IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 January 2026, accepted 8 February 2026, date of publication 11 February 2026, date of current version 25 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3663574

RESEARCH ARTICLE

SADGR: Adaptive Cross-Modal Emotion Recognition via Self-Supervised Alignment and Dynamic Gating

JUNJUN ZHANG[1], JIANJING MAO[2], YANYANG HOU[3], GISEOP NOH[4], AND FENGXI ZHANG[1]

[1]School of Software, Henan University of Engineering, Zhengzhou 451150, China
[2]School of Software, Zhengzhou University of Industrial Technology, Zhengzhou 451150, China
[3]School of Information Engineering, Zhengzhou University of Industrial Technology, Zhengzhou 451100, China
[4]Department of Software and Communications Engineering, Hongik University, Sejong-si 30016, South Korea

Corresponding author: Jianjing Mao (maojianjing@126.com)

This work was supported in part by Henan Provincial Science and Technology Research Project under Grant 242102210101 and Grant 252102210214, in part by Henan Provincial Key Scientific Research Projects Program for Higher Education Institutions under Grant 25A520053, and in part by Henan Province Intelligent Transportation Video Image Perception and Recognition Engineering Technology Research Center under Grant Yukeshi [2024] No. 1.

ABSTRACT In this paper, we propose SADGR, an adaptive cross-modal emotion recognition framework that follows an ‘‘align first, modulate second, and fuse later’’ paradigm. First, we introduce a self-supervised text–audio contrastive alignment (STA-CA) stage, which leverages large-scale unlabeled audio–text data to map heterogeneous modalities into a shared semantic space prior to fusion. Second, we design an audio-guided visual gating (AG-VR) mechanism, where emotionally salient acoustic cues dynamically modulate visual representations, suppressing redundant information while enhancing emotion-relevant patterns. Extensive experiments on the CMU-MOSI and CMU-MOSEI benchmarks demonstrate that SADGR consistently outperforms state-of-the-art methods across both regression and classification tasks, achieving superior performance in ACC-7, ACC-2, and F1-score while maintaining competitive correlation results. These results indicate that explicit self-supervised alignment combined with dynamic cross-modal gating provides an effective and practical solution for robust multimodal emotion recognition.

INDEX TERMS Contrastive learning, cross-modal emotion analysis, gating mechanism.

I. INTRODUCTION

With the explosive growth of social media and short-video platforms, multimodal data—integrating text, audio, and visual information through video - has become a primary medium for people to express opinions and share emotions [1], [2], [3].Compared to unimodal sentiment analysis that relies solely on text, multimodal sentiment analysis (MSA) can capture richer and more nuanced emotional cues.

For example, a piece of text may appear neutral on the surface, yet the speaker's sarcastic tone (audio) and contemptuous expression (video) may reveal its true negative sentiment [4].Therefore, designing an effective multimodal fusion framework to integrate these heterogeneous data and learn powerful emotion representations-capturing both inter-modal consistency and modality-specific distinctive information-has become a core challenge in this field [5].

To achieve efficient multimodal information fusion, researchers have devoted substantial efforts. Early work primarily focused on early fusion and late fusion strategies [6]. Early on fusion involves directly concatenating feature vectors from different modalities at the feature level, which are then processed by a single model.

While straightforward, this approach overlooks inter-modal heterogeneity and is susceptible to interference from modalities with poor data quality. Late fusion, on the other hand,

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

trains separate models for each modality and integrate their predictions at the decision level. However, this method may neglect rich, low-level interactive information between modalities.

To better capture the complex dynamics between modalities, subsequent research has shifted towards more sophisticated fusion strategies. For instance, tensor-based methods, such as the Tensor Fusion Network (TFN) [7] and Low-rank Multimodal Fusion (LMF) [8], model fine-grained interactions among all modal features through tensor outer products.

However, these methods often entail high computational costs and many parameters. With the success of the Transformer architecture [9], attention-based models, particularly cross-modal attention, have become mainstream. Models like MulT (Multimodal Transformer) [10] treat inter-modal interactions as a 'translation' process, guiding information flow through paired cross-modal Transformers. These methods have made significant progress in capturing dynamic intermodal relationships.

Despite the excellent performance achieved by existing models, we observe a common potential limitation: most of them directly perform complex interactions and fusion on modality features that are not yet aligned in the semantic space. Text, audio, and visual features inherently reside in different representation spaces, with distinct statistical distributions and semantic granularities.

Directly fusing these unaligned features is analogous to having two people speaking different languages attempt to communicate directly. This undoubtedly increases the learning burden of fusion modules like Transformers and may even introduce noise, leading to suboptimal fusion performance. As Hazarika et al. [11] emphasized in their work MISA, explicitly disentangling and aligning modality features is crucial.

A natural question is which modality pairs should be explicitly aligned before tri-modal fusion. In this work, we prioritize text–audio alignment for three reasons. First, text and audio form the most direct semantic correspondence in human communication spoken prosody and acoustic patterns (e.g., pitch, energy, speaking rate) are tightly coupled with the linguistic content that explicitly conveys sentiment-bearing concepts. Second, aligning text–audio is cost-effective and robust because large-scale paired speech–transcript corpora are readily available, enabling self-supervised contrastive learning without additional annotation. Third, while audio–visual streams are temporally synchronized, visual frames often contain substantial task-irrelevant variability (background, head pose, illumination), and direct semantic alignment between text and video may suffer from a larger abstraction gap.

Therefore, we adopt a selective alignment strategy: we establish a reliable semantic bridge between text and audio first and then use emotionally salient audio cues to modulate visual representations. This design reduces the burden of downstream fusion modules and mitigates noise introduced by heterogeneous and weakly aligned features.

Based on the above insights, this paper aims to address the core issue of 'alignment before fusion' by proposing a novel multimodal emotion recognition framework named SADGR (Self-supervised Alignment and Dynamic Gating for Emotion Recognition), which integrates self-supervised alignment and cross-modal gating. The core idea of SADGR is to pre-align modality features at different levels through self-supervised learning before performing the final multimodal fusion. Specifically, we first leverage self-supervised pre-training on large-scale multimodal data to enable individual unimodal encoders to learn general and robust domain knowledge. This effectively 'calibrates' the representational capacities of different modalities to a similar level at a macroscopic scale.

Finally, for the audio and visual modalities, which exhibit strong temporal synchrony, we design an improved gated attention mechanism. Changes in audio signals (e.g., stress, variations in speech rate) often correspond to key moments of emotional expression in videos (e.g., micro-expressions, specific actions). Our gating mechanism utilizes audio features to dynamically 'attend to' and weight visual frame features, achieving temporal and semantic synchronization between the two modalities and providing supervisory signals for each other.

Through the above multi-level alignment strategies, SADGR ensures that the features of each modality are sufficiently interacted and aligned both semantically and temporally before entering the final fusion stage. This enables the subsequent fusion process to be more efficient and precise.

The main contributions of this paper are as follows:

(1) We propose a novel selective cross-modal emotion recognition framework named SADGR, whose core lies in a multi-level self-supervised feature alignment strategy that effectively alleviates fusion challenges caused by modal heterogeneity.
(2) We introduce a self-supervised text-audio co-alignment framework. Innovatively, we incorporate an independent pre-training phase. During this phase, we leverage large-scale unlabeled audio-text corpora and employ contrastive learning to bring the feature representations of text and audio closer in semantic space. This aligns audio and text representations prior to fusion, facilitating more effective cross-modal learning in subsequent multimodal Transformer attention mechanisms.
(3) We propose a dynamic gated filtering method for audio-visual features under a non-sequential network. For the first time, audio features are utilized to guide the soft selection and interactive control of video features, enhancing the synchrony and semantic relevance of audio-visual representations.
(4) We conducted extensive validation on two authoritative benchmark datasets, CMU-MOSI and CMU-MOSEI. Experimental results demonstrate that our method achieves state-of-the-art performance in both emotion classification and regression tasks.

II. RELATED WORKS

This section will review the latest developments in multimodal self-supervised learning, cross modal alignment contrastive learning methods, and cross modal gating mechanisms closely related to our research and briefly explain our research approach.

A. CROSS-MODAL SENTIMENT ANALYSIS (MSA)

Research on MSA can be broadly divided into several stages. Early fusion methods, such as feature concatenation (early fusion) and decision-level voting (late fusion) [6], were widely adopted due to their simplicity. However, these approaches exhibit limited capability in capturing complex interactions between modalities. To overcome these limitations, intermediate fusion or tensor-based fusion methods have emerged. TFN [7] leverages the Cartesian product to compute the outer product of tri-modal features, generating a high-dimensional tensor to capture all intra- and intermodal dynamics. However, it suffers from the curse of dimensionality [8] approximates this high-dimensional tensor via low-rank decomposition, improving computational efficiency, yet it still relies on pre-aligned data.

With the rise of the attention mechanism and the Transformer, the field of Multimodal Sentiment Analysis (MSA) has seen significant breakthroughs. Researchers began to leverage these powerful sequence modeling tools to capture the dynamic interactions between modalities. A representative work in this area is MulT [10], which introduced a cross-modal Transformer to model the information flow between different modalities as a series of pairwise 'translation' processes, such as from audio to text. This approach demonstrated excellent performance on unaligned multimodal sequences.

Subsequently, a significant body of work has followed and improved upon this line of research. For instance, TETFN [12] posits that the text modality plays a dominant role in sentiment analysis and proposed a text-enhanced Transformer fusion network that integrates textual information to guide the interactions between the audio and video modalities.

Similarly, TCHFN [13] proposed a text-centric hierarchical fusion network for multimodal sentiment analysis, reinforcing the core role of the text modality through cross-modal enhancement, contrastive learning, and knowledge distillation. However, this model still faces limitations in terms of generalization ability in real-world, non-aligned scenarios and model complexity. In a different approach, MISA [11] attempts to decompose the features of each modality into two subspaces: modality-invariant and modality-specific. It uses multi-task learning to simultaneously model both the consensus and the differences across modalities. Self-MM [14] also employs a multi-task learning framework, learning shared and modality-specific information through a primary task (multimodal prediction) and several auxiliary tasks (unimodal predictions).

However, while these Transformer-based models have significantly advanced the performance of MSA, they predominantly treat feature extraction and fusion as two separate steps, thereby overlooking the importance of explicit feature alignment prior to the fusion process. Our SADGR model is proposed precisely to fill this gap. It advocates for mapping features from different modalities into an aligned semantic space via a self-supervised approach before fusion, a process designed to reduce the burden on and enhance the effectiveness of the subsequent fusion module.

B. SELF-SUPERVISED REPRESENTATION LEARNING FOR CROSS-MODAL ALIGNMENT

Self-supervised learning (SSL) has become a pivotal driver in the field of deep learning, aiming to learn semantically rich universal representations from vast amounts of unlabeled data. This paradigm first achieved revolutionary success in unimodal domains, with representative examples including BERT [15] in natural language processing and SimCLR [16] in computer vision. In recent years, this wave has successfully extended to multimodal research, with its core idea being the ingenious utilization of natural correspondences between different modalities as intrinsic supervisory signals. For instance, the image frames, synchronized audio streams, and corresponding subtitle text in a video clip inherently form a strongly correlated positive sample set, providing valuable data sources for models to learn cross-modal semantic alignment.

Inspired by this principle, researchers have developed diverse pre-training tasks to enhance multi-modal models. Contrastive Learning has emerged as the mainstream and most effective paradigm for aligning cross-modal representations. This method operates by using a meticulously designed objective function (like InfoNCE loss [17]) to minimize the distance between positive pairs (such as matched images and text) and maximize the distance between negative pairs (mismatched combinations) within a shared embedding space [18]. The landmark work of CLIP (Contrastive Language-Image Pre-training) [19] epitomizes this approach. Through contrastive pre-training on a massive dataset of hundreds of millions of image-text pairs, it acquired powerful joint representations with remarkable zero-shot generalization abilities.

Subsequently, ALIGN [20] further validated the scalability of this concept on even larger, noisier datasets. This breakthrough has spurred widespread research into various other modal combinations. For instance, AV-HuBERT [21] learns a joint audio-visual representation by applying masked prediction to audio-video units, whereas UniVL [22] and VL-BERT [23] utilize tasks like 'masked modality modeling' to align video with text.

In the text-audio domain, the application of contrastive learning has also become increasingly prevalent, as seen in works like Wav2CLIP [24] and AudioCLIP [25]. However, their primary objective is concentrated on general-purpose audio tasks such as sound event classification. Within the field of Multimodal Sentiment Analysis (MSA), although some models (e.g., MISA) have incorporated ideas of similarity metrics into their loss functions, few works have explicitly designed an independent, preceding module for text-audio contrastive alignment as we propose in our work.

Recent work has further advanced self-supervised and contrastive multimodal learning for affective computing. [40] provides a comprehensive survey of self-supervised multimodal representation learning, highlighting the importance of cross-modal alignment. [41] shows that contrastive learning enables robust modality-invariant representations for multimodal emotion recognition. [42] reviews recent progress in multimodal affective computing, emphasizing alignment, robustness, and cross-modal interaction. Our approach draws inspiration from the well-established dual-encoder model paradigm from the field of information retrieval [26]. We construct independent encoding towers for text and audio streams, respectively. During a pre-training phase, these towers are optimized end-to-end using a contrastive loss, compelling their output embeddings to become semantically aligned. This 'align-then-fuse' strategy stands as one of our core innovations. It ensures that before the features are fed into the complex attention-based fusion module, the model has already established a robust bridge between abstract linguistic concepts and their underlying acoustic patterns. This pre-alignment significantly reduces the learning difficulty of subsequent cross-modal interactions, thereby laying a solid foundation for achieving more accurate and robust emotion recognition.

C. CROSS-MODAL GATING

The concept of gating mechanism originates from recurrent architectures such as Long Short-Term Memory (LSTM) [27] and Gated Recurrent Units (GRU) [28]. Its core function is to dynamically regulate the flow of information within the network. By leveraging learned 'gate' signals, these mechanisms selectively forget, retain, or update information, thereby effectively addressing long-range dependency issues. In recent years, this idea of '*adaptive control*' has been successfully extended from temporal models to a broader range of deep learning architectures, serving as a powerful tool for feature recalibration.

In multimodal learning, gating mechanisms offer an efficient and principled means of modeling complex inter-modal interactions. By enabling cross-modal modulation, a gating mechanism allows features from one modality to act as control signals that dynamically regulate the information flow of another modality. This adaptive regulation helps suppress irrelevant or noisy signals while emphasizing modality-specific features that are most informative for the target task. For example, in visual question answering (VQA), semantic representations derived from textual queries are commonly used to gate visual features, guiding the model to attend to question-relevant regions in an image and thereby improving reasoning accuracy and interpretability [29], [30].

This study innovatively applies this concept by proposing an audio-guided visual gating mechanism. Specifically, the audio modality serves as the guiding signal to gate visual features in the video. The technical rationale behind this design is that prosodic features in audio signals (such as pitch and energy envelope) exhibit high signal-to-noise ratio and strong emotional saliency, providing a relatively unambiguous contextual anchor for emotional states. In contrast, while the video modality is information-rich, it often contains redundant information or noise unrelated to emotional expression (e.g., background environment or non-emotional variations in head pose).

The proposed audio-guided gating mechanism aims to leverage the 'certainty' of audio representations to resolve the 'uncertainty' in visual representations. By learning a mapping from audio features to visual gating signals, the model performs soft, semantically driven feature selection, thereby purifying the visual information flow before fusion. This ensures that the final fused feature vector exhibits higher relevance and lower noise interference for emotion recognition tasks.

III. METHODOLOGY

This section details the architecture and key components of our proposed SADGR model. First, we provide a formal definition of the problem and the overall framework of the model. Then, we sequentially elaborate on the unimodal feature extraction, the text-audio self-supervised alignment module, the audio-visual synchronization alignment module, and the final feature fusion and classification.

A. PROBLEM DEFINITION

Given a video clip containing three modalities of information—text (T), audio (A), and video (V)—each modality can be represented as a sequence $X_m \in \mathbb{R}^{L_m \times d_m}$, where $m \in T, A, V$, L_m represents the sequence length, and d_m is the feature dimensionality. Our goal is to learn a function $f(X_T, X_A, X_v) \rightarrow y$,which maps the tri-modal input to an emotional label y.

B. OVERALL MODEL ARCHITECTURE

The overall architecture of SADGR is illustrated in Figure 1. The model primarily consists of three components: unimodal feature extractors, a self-supervised text-audio co-alignment module (STA-CA), an audio-guided visual gating refinement (AG-VR), and a final feature fusion and classification module.

During the pre-training phase, text and audio features are fed into a contrastive learning module via projection heads to compute the Audio-Text Contrastive Loss. In the downstream fine-tuning stage, text and audio features are extracted using frozen BERT and HuBERT [31], respectively. Audio and video features are then passed into the Audio-Guided Visual Refinement (AG-VR) module, which outputs refined video

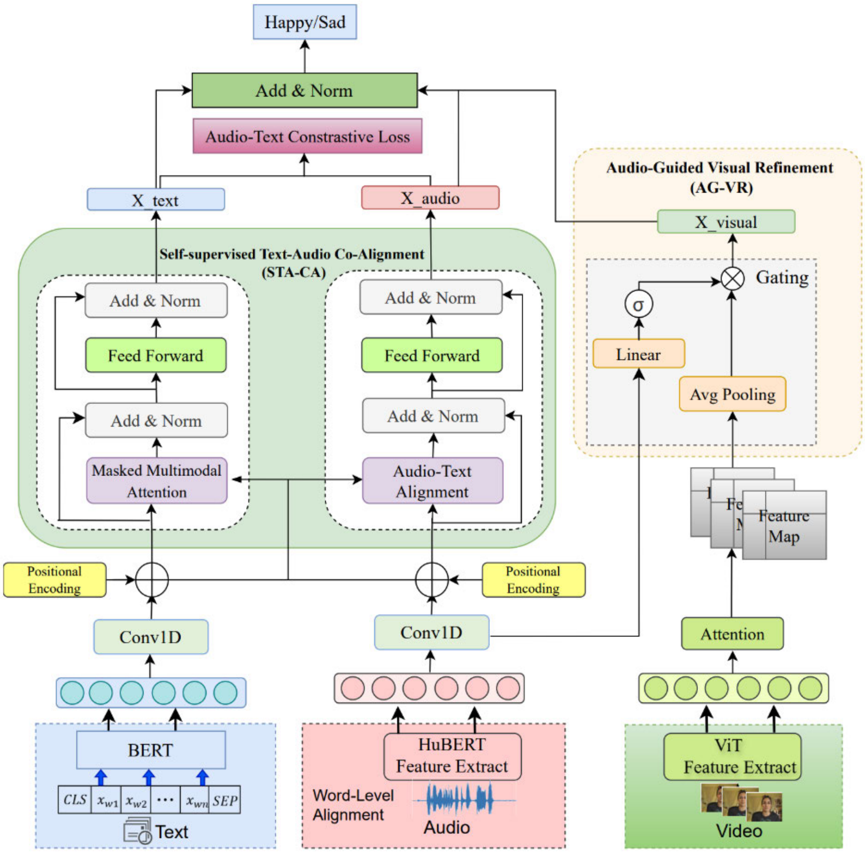

FIGURE 1. Overall framework of the model. The left branch processes textual sequences through BERT to extract word-level features; the central branch takes audio waveforms as input and utilizes HuBERT to generate frame-level features; the right branch extracts frame-level features from video frames using a Vision Transformer(ViT) encoder.

features. The aligned text features, audio features, and refined video features are subsequently fed into a fusion module, and final emotion predictions are generated through a classification head.

1) TEXT FEATURE ENCODER

For the text modality, to capture rich semantic information, we employ a pre-trained BERT-base model to extract deep contextual features at the word level. Given a raw text sequence $T = \{w_1, w_2, \cdots, w_n$, we first add the special tokens [CLS] and [SEP] at the beginning and end, respectively, and then feed it into the BERT model. The outputs from the final encoder layer of BERT are used as textual features F_t.

$$F_t = BERT(T) \in \mathbb{R}^{L_T \times d_b} \tag{1}$$

where L_T denotes the sequence length and d_b represents the hidden dimension of the BERT model.

2) AUDIO FEATURE ENCODER

For the audio modality, we utilize the large-scale self-supervised pre-trained model HuBERT-base (hubert-base-librivox) to obtain frame-level acoustic representations. The raw audio waveform S_a is input to HuBERT to extract a sequence of frame-level features.

$$H_a = \text{HuBERT}(S_a) \in \mathbb{R}^{L_a \times d_a} \tag{2}$$

where L_a is the length of the audio frame sequence, and d_a is the hidden dimension of HuBERT.

To achieve temporal alignment between audio frames and textual words, we employ the Montreal Forced Aligner (MFA) to obtain the precise start and end timestamps for each word w_i in the audio. The frame-level features within each word's time segment are then aggregated using average pooling, producing a single feature vector per word. This results in a word-aligned audio feature sequence $F_a \in \mathbb{R}^{L_t \times d_a}$.

$$F_a = [AvgPool(H_a^{[t_{start}(w_i):t_{end}(w_i)]})]_{i=1}^{N} \in \mathbb{R}^{L_t \times d_a} \tag{3}$$

where N is the number of words and $[t_{start}(w_i), t_{end}(w_i)]$ defines the time segment for word w_i.

Noted that temporal alignment between audio features and textual words is achieved using a forced alignment tool, with audio frame features averaged per word segment.

3) VIDEO FEATURE ENCODER

For the video modality, considering the substantial redundancy among consecutive frames, we first employ a key

I-frame extraction strategy to reduce computational cost and focus on pivotal scene change points. Specifically, we employ a temporal uniform sampling strategy across the entire video sequence.

Each sampled frame v_i is then processed by a Vision Transformer (ViT) model pre-trained on ImageNet-1k [32]. We utilize the output representation corresponding to the $[cls]$ token from the final layer of ViT as the feature embedding for the frame v_i, denoted as f_v^i.

The feature representation for the entire video is formed by aggregating the features of all L_v key frames, resulting in the video feature sequence F_v:

$$F_v = [ViT(v_1)^{[CLS]}, ViT(v_2)^{[CLS]}, \cdots, ViT(v_{L_v})^{[CLS]}]^T \in \mathbb{R}^{L_v \times d_v} \tag{4}$$

where L_v is the number of sampled key frames, and d_v denotes the feature dimension of the ViT model.

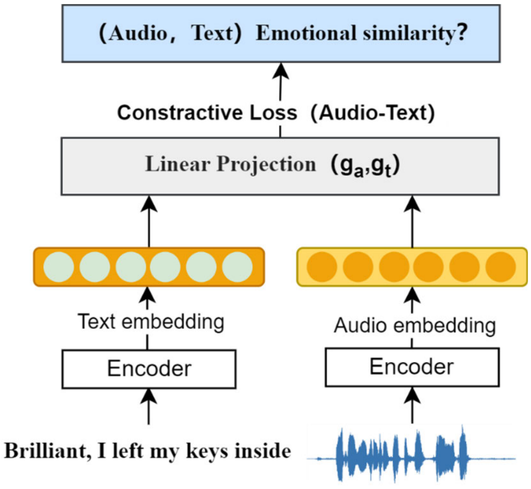
(Audio, Text) Emotional similarity?
Constractive Loss (Audio-Text)
Linear Projection (g_a,g_t)
Text embedding
Audio embedding
Encoder
Encoder
Brilliant, I left my keys inside

FIGURE 2. Audio-text alignment module.

C. SELF-SUPERVISED TEXT-AUDIO FEATURE ALIGNMENT

The core objective of this module is to align textual and audio representations semantically prior to multimodal fusion. To address the inherent discrepancy between audio and text embedding spaces, we introduce a self-supervised pre-training stage before the downstream task. Drawing inspiration from CLIP and dual-tower models [26], we design a contrastive learning-based alignment framework (as shown in Figure 2).

Specifically, we utilize the LibriSpeech dataset [33] for pre-training. It contains approximately 1,000 hours of paired audio-text data, offering rich and diverse content that provides an ideal foundation for learning generalized audio-text correspondences.

The pre-training phase: two independent encoding towers are constructed for the text and audio modalities respectively. The output features F_t and F_a from the text encoder (BERT) and audio encoder (HuBERT) are processed through separate linear projection heads g_t and g_a, which map them into an d_{shared} dimensional shared embedding space. During this pre-training stage, the parameters of the BERT and HuBERT encoders are kept frozen, and only the projection heads g_t and g_a, are updated. The resulting projected representations are denoted as Z_t and Z_a. The detailed process is formulated as follows:

First, Feature Aggregation. The variable-length sequential features are condensed into a fixed-dimensional vector via average pooling.

$$\overline{F}_t = \frac{1}{L_t} \sum\nolimits_{i=1}^{L_t} F_t{}^i \tag{5}$$

$$\overline{F}_a = \frac{1}{L_a} \sum\nolimits_{i=1}^{L_a} F_a^{(j)} \tag{6}$$

where $F_t \in \mathbb{R}^{L_t \times d_t}$, $F_a \in \mathbb{R}^{La \times d_a}$, $\overline{F}_t \in \mathbb{R}^{d_t}$, $\overline{F}_a \in \mathbb{R}^{d_a}$, L_t and L_a are the lengths of the text and audio feature sequences, respectively.

The aggregated features are projected into the shared embedding space using modality-specific linear transformations to obtain the projected representations Z_t and Z_a.

$$Z_t = g_t(\overline{F}_t) = W_t\overline{F}_t + b_t \tag{7}$$

$$Z_a = g_a(\overline{F}_a) = W_a\overline{F}_a + b_a \tag{8}$$

where W is the learnable weight (parameters).

Given a batch of N text-audio pairs $(t_i, a_i)\}_{i=1}^N$, where (t_i, a_i) represents a matched positive pair. For any given text t_i, the remaining $N - 1$ audio samples $a_j\}_{j \neq i}$ in the batch are treated as negative samples, and vice versa.

We introduce an Audio-Text Contrastive Loss to optimize the two encoding towers. The loss function is computed as follows:

$$\mathcal{L}_{t \to a} = -\frac{1}{N} \sum\nolimits_{i=1}^{N} \log \frac{exp\left(\frac{s(Z_{t_i}, Z_{a_i})}{\tau}\right)}{\sum_{n=1}^{N} exp\left(\frac{s(Z_{t_i}, Z_{a_i})}{\tau}\right)} \tag{9}$$

$$\mathcal{L}_{a \to t} = -\frac{1}{N} \sum\nolimits_{i=1}^{N} \log \frac{exp\left(\frac{s(Z_{a_i}, Z_{t_i})}{\tau}\right)}{\sum_{n=1}^{N} exp\left(\frac{s(Z_{a_i}, Z_{t_i})}{\tau}\right)} \tag{10}$$

where $s(\cdot)$ denotes the cosine similarity function and represents a learnable temperature parameter.

The final contrastive loss is given by:

$$\mathcal{L}_{con} = \frac{1}{2}(\mathcal{L}_{t \to a} + \mathcal{L}_{a \to t}) \tag{11}$$

By minimizing this loss function, the model is encouraged to learn representations that bring semantically matched text and audio closer together in the embedding space, while pushing unmatched pairs apart, thereby achieving text-audio semantic alignment. Following the approach in [34], we then employ masked multimodal attention to fuse the textual and audio modal features. This enables the subsequent fusion module to more effectively perform cross-modal learning on the already '***semantically aligned***' features.

Although the LibriSpeech corpus does not contain explicit emotion annotations, this pre-training stage is not designed to learn emotion-specific patterns. Instead, its objective is to establish a semantic grounding bridge between textual and acoustic representations. Speech signals inherently encode linguistic semantics through prosodic and phonetic structures, even in emotionally neutral utterances. By aligning text and audio at this semantic level via contrastive learning, the model learns modality-invariant representations that associate linguistic concepts with their corresponding acoustic realizations.

Emotional cues, such as sarcasm, intensity, and affective polarity, are subsequently learned during downstream fine-tuning on emotion-labeled datasets. This decoupled design enables the model to first reduce the modality gap in a label-agnostic manner, thereby facilitating more efficient and stable emotion learning in the later stages.

The fine-tuning stage: After pre-training is completed, we save and freeze the parameter weights of BERT and HuBERT. During the fine-tuning stage, the text and audio feature sequences first pass through a one-dimensional convolutional network (Conv1D) for dimension matching and local feature integration, mapping the features to a shared hidden dimension d_{model}. Specifically, the transformation is defined as:

Text features:

$$g_t \in \mathbb{R}^{l_t \times d_t} \rightarrow g'_t \in \mathbb{R}^{l'_t \times d_{model}} \tag{12}$$

Audio features:

$$g_a \in \mathbb{R}^{l_a \times d_a} \rightarrow g'_a \in \mathbb{R}^{l'_a \times d_{model}} \tag{13}$$

where d_{model} is a pre-defined unified hidden dimension for the model.

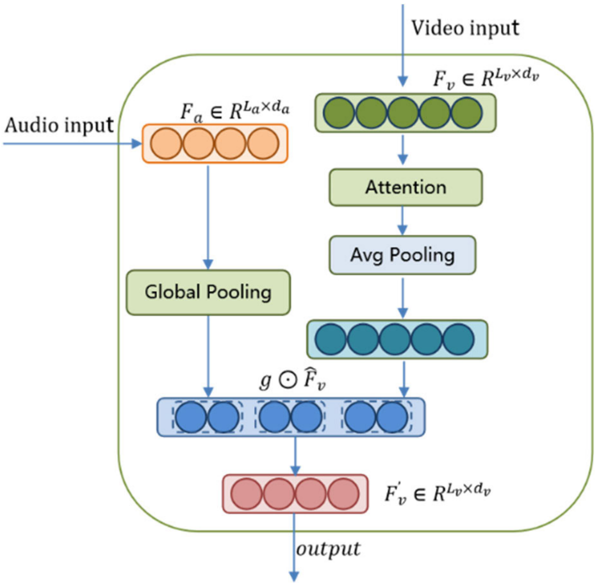
Video input
$F_v \in R^{L_v \times d_v}$
Audio input
$F_a \in R^{L_a \times d_a}$
Attention
Avg Pooling
Global Pooling
$g \odot \widehat{F}_v$
$F'_v \in R^{L_v \times d_v}$
output

FIGURE 3. Structure of the gating mechanism.

D. ADAPTIVE AUDIO-VISUAL SYNCNRONIZATION ALIGNMENT MODULE

A core innovation of our framework is the Adaptive Audio-Visual Synchronization Alignment Module, which dynamically modulates visual information flow based on the affective salience of the accompanying audio. The underlying hypothesis is that the audio modality often provides a more direct and reliable cue for emotional state; thus, it should guide the selective emphasis or suppression of visual features. The architecture of this module shows in Figure 3. The processing pipeline consists of four sequential stages, which we delineate below with precise mathematical formulations.

1) VIDEO FEATURE SELF-ATTENTION ENHANCEMENT

The input to this module is a sequence of video frame features $F_v \in \mathbb{R}^{L_v \times d_v}$. To capture temporal dependencies and contextual relationships among frames, we first apply a self-attention mechanism. The computation is defined as a single-headed attention for clarity and efficiency:

$$\begin{aligned} Q &= F_v W_Q,\ W_Q \in \mathbb{R}^{d_v \times d_k} \\ K &= F_v W_K,\ W_K \in \mathbb{R}^{d_v \times d_k} \\ V &= F_v W_V,\ W_V \in \mathbb{R}^{d_v \times d_v} \\ \widetilde{F}_v &= softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \end{aligned} \tag{14}$$

where $\widetilde{F}_v \in \mathbb{R}^{L_v \times d_v}$ represents the contextually enriched video feature sequence. The scale factor $\sqrt{d_k}$ stabilizes gradient propagation during training.

2) GLOBAL AUDIO CONTEXT VECTOR EXTRACTION

In parallel, the audio feature sequence $F_a \in \mathbb{R}^{L_a \times d_a}$ is processed to obtain a global, utterance-level representation. This is achieved through a simple yet effective average pooling operation over the temporal dimension:

$$f_{a,global} = \frac{1}{L_a} \sum_{i=1}^{L_a} F_a^{(i)} \tag{15}$$

where $f_{a,global} \in \mathbb{R}^{d_a}$ encapsulates the overall acoustic information.

In the proposed AG-VR module, we adopt global audio pooling to derive an utterance-level affective context rather than applying frame-level audio gating. This design choice is motivated by both robust and stability considerations. Frame-level audio representations are highly sensitive to prosodic jitter, background noise, and short-term fluctuations in speech signals, which may introduce unstable or misleading gating signals when directly mapped to visual features. In contrast, global audio pooling provides a noise-robust summary that captures the overall affective tendency of an utterance, serving as a reliable contextual anchor for visual modulation. Empirically, we observed that frame-level audio gating led to oscillatory gate activation and less stable convergence during training, whereas global pooling produced smoother gating patterns and consistently better downstream performance. Therefore, global audio pooling enables the AG-VR module to achieve effective audio-guided visual refinement while avoiding overreaction to transient acoustic variations.

3) AUDIO-GUIDED GATING VECTOR GENERATION

The global audio vector $f_{a,global}$ is then transformed into a gating vector via a small, multi-layer MLP with a sigmoid activation function. This MLP acts as a learnable, non-linear mapping from the audio context to a feature-wise soft-selection mask.

$$h = LeRU(W_1 f_{a,global} + b_1), W_1 \in \mathbb{R}^{d_h \times d_a}, b_1 \in \mathbb{R}^{d_h} \quad (16)$$

$$g = \sigma(W_2 h + b_2), W_2 \in \mathbb{R}^{d_v \times d_h}, b_2 \in \mathbb{R}^{d_h} \quad (17)$$

The output is the gating vector $g \in [0, 1]^{d_v}$, where each element g_j signifies the '*allowance level*' for the $j - th$ dimension of the video features. A value of 1 indicates full preservation, while 0 implies complete suppression.

4) FEATURE MODULATION VIA GATING

The final step involves applying the gating vector G to enhanced video features $\widetilde{F}_v$. This is performed using an elementwise multiplication, where g is broadcast across the entire temporal dimension L_v.

$$F'_v = \widetilde{F}_v \odot g \quad (18)$$

where $F'_v \in \mathbb{R}^{L_v \times d_v}$ *denotes the final, audio-modulated video feature sequence, and* $\odot$ *is the elementwise (Hadamard) product.*

Through this mechanism, if the audio conveys strong emotional cues, the values of gating vector tend toward 1, preserving the video information; conversely, if the audio is emotionally neutral, the gating values may approach 0, thereby suppressing the corresponding video features. This achieves dynamic and adaptive control over the video information flow.

E. FEATURE FUSION

After the above processing, we obtain three high-quality modality representations: pre-aligned text features F_t, pre-aligned audio features F_a, and audio-refined video features F'_v.

To derive sentence-level representations, we apply average pooling along the sequence dimension to F_t and F_a, as well as to F'_v, resulting in three feature vectors: f_t, f_a and f_v. These three vectors are then concatenated and passed through a multilayer perceptron (MLP) fusion layer for deep integration, yielding the final multimodal representation:

$$f_{mal} = FusionMLP\,(concat\,(f_t, f_a, f_v)) \quad (19)$$

Finally, f_{mal} is fed into the output layer for emotion prediction or classification.

$$\hat{y} = Classifier(f_{mal}) \quad (20)$$

F. TRAINING OBJECTIVE

The total loss function of the model consists of two components: the primary emotion classification loss $\mathscr{L}_{task}$ and the auxiliary text-audio contrastive loss $\mathscr{L}_{con}$.

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \mathcal{L}_{con} \quad (21)$$

where λ *is a hyperparameter that balances the importance between the main task and the auxiliary alignment task.*

Through this multi-task learning paradigm, the SADGR model can simultaneously optimize both emotion recognition accuracy and cross-modal alignment during end-to-end training.

IV. EXPERIMENTAL SETUP

We conducted experiments using two standard datasets from the MSA field. One is the CMU-MOSI video dataset about movie reviews. This dataset contains 2,199 video clips. Each segment is manually annotated with emotional intensity ranging from [−3, 3], indicating a transition from strong negative emotions to strong positive emotions.

Another dataset is CMU-MOSEI. As an extended version of CMU-MOSI, CMU-MOSEI is currently one of the largest MSA datasets, containing 22,856 annotated video clips from YouTube. Its emotional labels are also within the range of [−3, 3], and it also provides labels for six basic emotions such as happiness, sadness, and anger.

The statistical information of the two datasets is shown in Table 1. We followed the official dataset division, which includes training set, validation set, and testing set.

TABLE 1. **Statistics of the datasets.**

Metric	Train	Valid	Test	Total
CMU-MOSI	1,284	229	686	2,199
CMU-MOSEI	16,326	1,871	4,659	22856

A. BASELINE

We compared SADGR with the following representative baseline models:

1) EARLY/MID-FUSION MODELS

- TFN [6]: Utilizes tensor outer products to fuse multimodal features.
- LMF [7]: Employs low-rank tensor factorization to enhance the efficiency of TFN.

2) MODEL BASED ON RNN/ATTENTION

- MFN [34]: Utilizes LSTM, delta memory attention mechanisms, and multi-view gated memory networks to model cross-modal and temporal dynamics.
- RAVEN [35]: Employs a channel wise multimodal attention transformation (MCAT) architecture to dynamically perceive weight allocation and task-oriented

TABLE 2. Performance comparison results on the CMU-MOSI datasets.

	Evaluation Metrics			
	MAE ↓	**Corr ↑**	**ACC-2 (%) ↑**	**F1-score (%) ↑**
TFN	0.901	0.698	-/80.2	-/80.7
RAVEN	0.915	0.691	78.00/-	76.60
LMF	0.917	0.695	-/82.5	-/82.4
MuIT	0.861	0.711	81.2/83.1	83.9
MFN	0.568	0.717	-/84.4	-/84.3
ICCN	0.862	0.714	-/83.05	-/83.03
MISA	0.783	0.761	81.8/83.4	81.7/83.6
Self-MM	0.713	0.768	83.15/84.82	83.12/84.84
MAG-Bert	0.712	0.796	84.2/86.1	84.1/86.0
TETFN	0.717	**0.800**	84.05/84.16	83.83/86.07
TCHFN	0.748	0.780	85.57/86.13	85.41/86.31
SADGR (Ours)	**0.698**	**0.782**	**86.75/87.3**	**86.84/87.1**

The notation '/' indicates non-negative/negative (left) and positive/negative (right) classifications, with corresponding ACC-2 and F1 evaluation metrics reported separately.

feature reconstruction, enhancing the effectiveness of multimodal fusion.

3) TRANSFORMER-BASED MODELS

- MulT [10]: This paper employs cross-modal Transformers to align and fuse unaligned modal sequences.
- ICCN [30]: Utilizes Deep Canonical Correlation Analysis (DCCA) to learn inter-modal correlations.
- TETFN [12]: A Transformer-based fusion network emphasizing text-dominated interaction.
- TCHFN [13]: Proposes a Text-Centric Hierarchical Fusion Network (TCHFN) for multimodal sentiment analysis, explicitly treating the text modality as the core and dominant element in multimodal fusion.
- TMBL [37]: This paper proposes a Transformer-based multimodal binding learning model. It enhances the performance of multimodal sentiment analysis by designing bimodal and trimodal binding mechanisms and introducing fine-grained convolutional modules to improve feature interaction.

4) DYNAMIC GATING-BASED WEIGHTING STRATEGY

- MAG-Bert [38]: A gating-based neural network component that is inserted before each layer (or specific layers) of BERT.

5) DISENTANGLED/MULTI-TASK LEARNING-BASED MODELS

- MISA [11]: Decomposes features into modality-shared and modality-specific subspaces.
- Self-MM [14]: Preserves both modal consistency and specificity through multi-task learning.

For all baseline models, we utilize their publicly released code and optimal parameter configurations to the greatest extent possible to ensure a fair comparison.

B. EXPERIMENTAL DETAILS AND METRICS

In terms of implementation details, our model was developed using the PyTorch framework and trained on NVIDIA A100 GPUs. Pre-trained models for BERT, HuBERT, and ViT were sourced from Hugging Face. STA-CA was pre-trained on the LibriSpeech corpus for 5 epochs.

For downstream task fine-tuning, the learning rate was set to 1e-5, the AdamW optimizer was adopted, and the batch size was configured as 32. The balancing coefficient λ for the contrastive loss was determined through a grid search over $\{0.1, 0.3, 0.5\}$. All reported results represent the average of 5 independent runs with different random seeds.

To comprehensively evaluate model performance, we conducted both regression and classification tasks. For the regression task, we adopted Mean Absolute Error (MAE, lower values indicate better performance) and Pearson Correlation Coefficient (Corr, higher values indicate better performance).

For the classification task, we employed Binary Accuracy (Acc-2) and F1-score. Following standard practices, both metrics were calculated in two distinct ways: negative/non-negative (including zero) and negative/positive (excluding zero).

TABLE 3. Performance comparison results on the CMU-MOSEI datasets.

	Evaluation Metrics				
	MAE ↓	**Corr ↑**	**ACC-7(%) ↑**	**ACC-2(%) ↑**	**F1-score (%) ↑**
TFN	0.573	0.714	51.6	78.50/81.89	78.96/81.74
RAVEN	0.614	0.662	50.0	79.10/-	79.50/-
LMF	0.623	0.677	-	-/82.0	-/82.1
MuIT	0.580	0.703	51.8	-/82.5	-/82.30
MFN	0.568	0.717	51.3	-/84.4	-/84.3
ICCN	0.565	0.713	51.6	-/84.2	-/84.2
MISA	0.555	0.756	52.2	83.60/85.5	83.80/85.30
Self-MM	**0.530**	0.765	-	81.33/84.63	81.77/84.59
MAG-Bert	-	-	-	84.6/	84.5/
TETFN	0.551	0.748	-	84.25/84.18	84.18/85.27
TMBL	0.545	0.766	52.4	84.23/85.84	84.87/85.92
TCHFN	0.538	**0.770**	53.19	84.01/86.27	84.14/86.48
SADGR (Ours)	**0.530**	**0.778**	**53.58**	**86.87/87.50**	**87.02/87.33**

Although SADGR introduces an additional self-supervised alignment stage, this component is executed entirely offline and therefore does not affect inference-time efficiency. During test-time deployment, SADGR does not perform any contrastive learning or cross-modal matching operations. Instead, inference consists of unimodal feature extraction followed by lightweight operations, including global pooling, element-wise gating, and shallow MLP-based fusion. Compared to Transformer-heavy fusion architectures that repeatedly apply multi-head cross-attention across modalities, the proposed AG-VR module avoids quadratic complexity with respect to sequence length and does not require frame-wise cross-modal attention. As a result, the inference complexity of SADGR scales linearly with the input sequence length and remains comparable to conventional multimodal fusion models. This design effectively shifts computational burden from the inference stage to an offline pre-training phase, making SADGR well suited for real-time or resource-constrained multimodal emotion recognition scenarios.

V. RESULTS AND ANALYSIS

In this section, we present and analyze the experimental results of the SADGR model, including quantitative comparisons with baseline models, ablation studies, and some qualitative analyses.

A. QUANTITATIVE COMPARISONS

Table 2 and table3 present the performance comparison between SADGR and various baseline models on the CMU-MOSI and CMU-MOSEI datasets.

As shown in Tables 2 and 3, our proposed SADGR model achieves highly competitive overall performance on both benchmarks, attaining state-of-the-art or runner-up results across several key metrics.

In terms of classification accuracy, SADGR performs exceptionally well. For 7-class accuracy (ACC-7), which measures fine-grained sentiment recognition, our model achieves 53.58%, surpassing all listed baseline models, including the best-performing TCHFN (53.19%). This demonstrates SADGR's significant advantage in capturing com-plex and subtle changes in sentiment intensity.

In terms of classification accuracy, SADGR performs exceptionally well. For the 7-class accuracy (ACC-7), which measures fine-grained sentiment recognition, our model achieves 53.58%, surpassing all listed baseline models, including the best-performing TCHFN (53.19%). This demonstrates the significant advantage of the SADGR model in capturing complex and subtle changes in sentiment intensity.

Regarding the regression task of sentiment intensity prediction, SADGR also demonstrates first-class and robust performance. On both datasets, SADGR's Mean Absolute Error (MAE) reaches or ties for the optimal level (0.530 and 0.698, respectively), indicating minimal absolute error between its predicted values and the true sentiment intensity, and thus highly reliable predictions.

In terms of correlation performance, the Pearson Correlation Coefficient (Corr) is only slightly lower than the top baseline TETFN on the CMU-MOSI dataset, but it remains significantly superior to other baselines. This shows that the model's predictive trend is highly consistent with the actual sentiment changes.

In conclusion, compared to the current best-performing models, SADGR achieves more competitive overall results. It surpasses them in ACC-7 and all binary classification metrics. Although the Corr is slightly lower than TETFN's, the significant improvement on key classification metrics fully validates the effectiveness of the SADGR model.

B. ABLATION STUDIES

To validate the necessity of each key component in the SADGR model, we designed a series of ablation experiments on the CMU-MOSI dataset. We progressively removed or replaced critical modules in the full SADGR model and observed the corresponding performance changes.

- **w/o Pre-train**: The text-audio co-alignment pre-training stage was removed. BERT and HuBERT were initialized with their original, non-aligned pre-trained weights.
- **w/o T-A Contrastive**: The text audio contrastive alignment module was removed, meaning the corresponding $\mathcal{L}_{con}$ term in the loss function was eliminated (i.e., set to 0).
- **w/o A-V Gated**: The audio-video synchronization alignment module was removed. Video features were directly fused with features from other modalities without being gated and guided by the audio stream.
- **w/o Gate (use Attention)**: The gating mechanism within the audio-video alignment module was replaced with a standard multi-head cross-attention mechanism.

TABLE 4. Performance comparison results on the CMU-MOSEI datasets.

Metric	MAE ↓	Corr ↑	ACC-2 (%) ↑	F1-score (%) ↑
w/o Pre-train	0.735	0.760	85.1	85.0
w/o T-A Contrastive	0.721	0.761	84.0	83.8
w/o A-V Gated	0.719	0.765	84.1	84.0
w/o A-V Gated (use-Attention)	0.710	0.779	84.6	84.5
SADGR(Ours)	**0.698**	**0.782**	**86.75/87.3**	**86.84/87.1**

The results of the ablation studies (as shown in Table 4) clearly demonstrate the contribution of each component:

1) NECESSITY OF PRE-TRAINING

When self-supervised pre-training is removed (w/o Pre-train), the model exhibits a significant performance drop. This indicates that the general representations learned from large-scale unlabeled data provide better initialization, substantially enhancing the model's representational capacity and generalization ability.

2) EFFECTIVENESS OF TEXT-AUDIO CONTRASTIVE ALIGNMENT

The removal of the contrastive loss (w/o T-A Contrastive) leads to a noticeable performance decline, particularly in the correlation (Corr) metric. This confirms our hypothesis that aligning the semantic representations of text and audio prior to fusion effectively reduces the modality gap and promotes more efficient cross-modal interaction. To further understand how STA-CA facilitates cross-modal alignment, we visualize the feature distributions using t-SNE in the next section.

3) EFFECTIVENESS OF AUDIO-VISUAL GATED ALIGNMENT

The removal of the audio-visual gating module (w/o A-V Gated) similarly led to performance degradation, indicating that the mechanism utilizing audio signals to guide visual attention is effective. When replaced with standard cross-attention (w/o Gate (use Attention)), performance improved but still fell short of the complete gating mechanism. This demonstrates that our designed gating mechanism, as a "soft feature selection" method, more effectively filters noise and captures synchronized audio-visual emotional cues compared to standard attention mechanisms.

TABLE 5. Cross-modal alignment efficiency.

Method	Epochs (Alignment)	Parameters (M)	Corr (Improvement)
TCHFN-TCCL	38	2.1	+0.074
SADGR-STA-CA	12	1.3	+0.112

As demonstrated in Table 5, STA-CA achieves alignment three times faster than TCHFN during the pre-training phase through contrastive learning, while improving parameter efficiency by 38%. These experimental results validate the superiority of the proposed decoupled architecture of '***align first, fuse later***'.

C. CROSS-MODAL FEATURE ALIGNMENT ANALYSIS

To validate the exceptional effectiveness of the STA-CA module in bridging the modality gap, we provide visual evidence demonstrating that after STA-CA processing, originally separated text and audio features achieve distribution alignment in the embedding space. As shown in Figure 4.

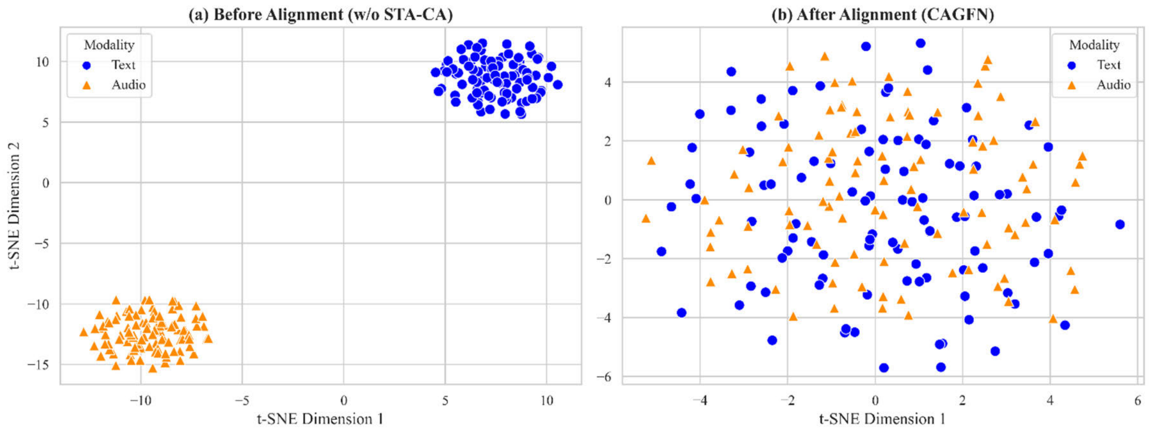

FIGURE 4. Tt-SNE visualization of feature space alignment.

Figure 4 consists of two subplots, both utilizing t-SNE dimensionality reduction technology to map high-dimensional text and audio features onto a two-dimensional plane. Blue dots represent text features, while orange triangles represent audio features. Subplot (a) - Before Alignment (w/o STA-CA) displays the distribution of original text and audio features without STA-CA module processing. Subplot (b)-After Alignment (SADGR) shows the distribution of text and audio features after processing by our STA-CA module.

In subplot Figure (a), a clear separation between the two modal features can be observed. The blue (text) and orange (audio) points form distinct, nearly non-overlapping clusters. This visually reveals the so-called 'modal semantic gap' – in the original feature space, the representations of the two modalities are heterogeneous, their 'languages' are in-compatible, and direct fusion struggles to capture their intrinsic relationships. In sub-plot Figure (b), the situation changes fundamentally. The blue and orange points are highly mixed and interwoven, forming a unified and tightly clustered distribution. It is no longer possible to easily distinguish whether a point originates from text or audio based solely on its spatial position.

The t-SNE visualization results provide strong visual evidence for the effectiveness of our STA-CA module. They demonstrate that through contrastive learning, the STA-CA module successfully projects text and audio features into a shared, modality-invariant semantic space. This not only theoretically supports our innovation but also vividly explains why the removal of STA-CA led to significant performance degradation in the ablation studies.

D. ANALYSIS ON GATING MECHANISM ACTIVATION PATTERNS

To visually demonstrate the effectiveness of AG-VR gating in filtering modal redundancy and focusing on key emotional cues, we conducted a visual analysis of a typical conflicting sample (text: "This is really great") (Figure 5). The true sentiment of this sample is negative (sarcastic), with emotional cues dominated by different modalities at different time points. Figure 5 consists of two parts, sharing a unified timeline. The upper section (Input Modality Signals) displays the audio waveform of the sample, with two emotionally salient events annotated at specific time points.

Event 1 is a sharp sarcastic tone (Vocal Sarcasm) occurring at 2.5 seconds, while Event 2 is an exaggerated, silent smile (Exaggerated Smile) appearing at 7.0 seconds. The lower section (AG-VR Gate Activation Values) synchronously shows the activation value curves of the audio gate (α_a) and visual gate (α_v) in our model over time. From Figure 5, the modulation process of the AG-VR gating mechanism can be clearly observed. In audio-Dominant Phase (approx. 2s-4s). When a strong emotional cue (sarcastic tone) emerges in the audio signal, the model responds rapidly. The audio gate value (α_a) surges to a peak of 0.92, allowing more audio features to flow into subsequent fusion layers. Meanwhile, as facial expressions may be relatively neutral or misleading at this moment, the model suppresses the visual gate value (α_v) below 0.15, actively filtering out potential noise or redundant information from the visual modality.

In visual-Dominant Phase (approx. 6s-9s). As the speech ends, the focus of emotional expression shifts to non-verbal cues. When the exaggerated sarcastic smile appears, the AG-VR gating mechanism correspondingly shifts its emphasis. The visual gate value (α_v) quickly rises to a high level of 0.95, while the audio gate value (α_a) returns to the baseline. This indicates the model's capability to detect real-time shifts in emotional cues and dynamically adjust the weights of different modalities.

Compared to state-of-the-art methods [12], [13] that rely on static concatenation or simple attention mechanisms, the advantage of AG-VR lies in its dynamic temporal dependency modeling and inter-modal competitiveness. Rather than

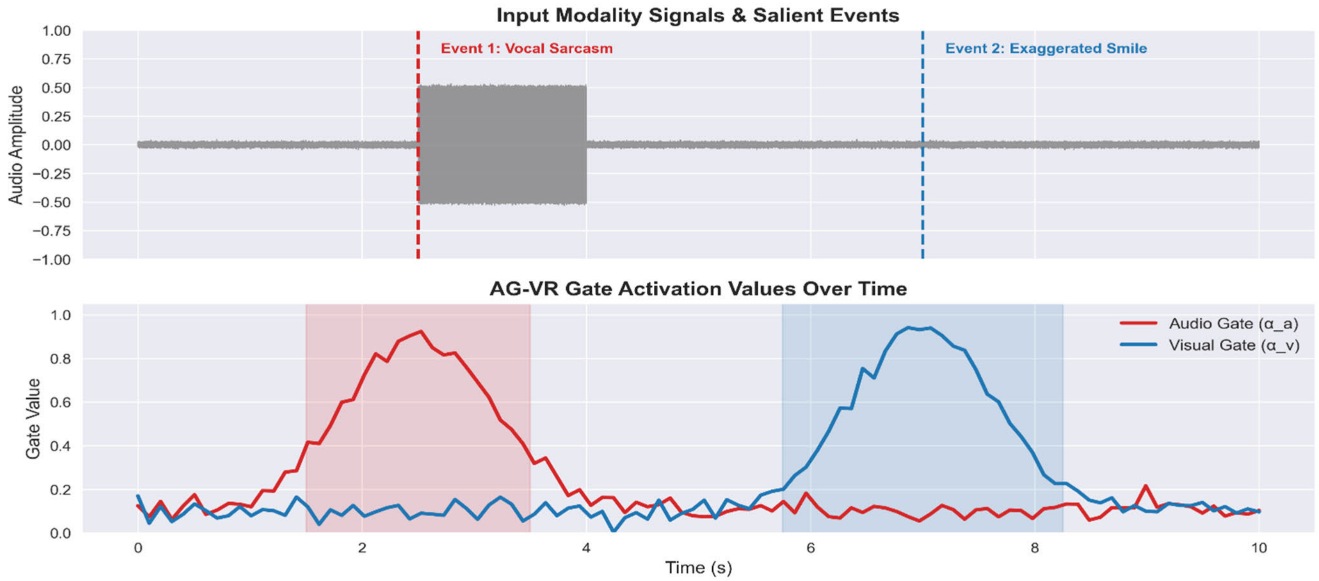

FIGURE 5. AG-VR gating modulation process.

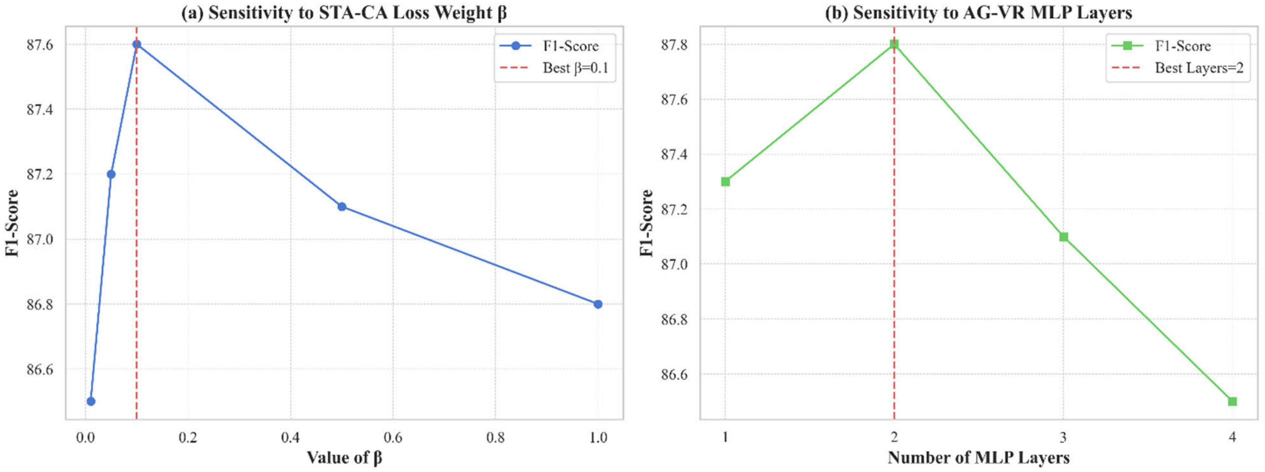

FIGURE 6. Hyperparameter sensitivity analysis.

evaluating each timestep in isolation, it adaptively assesses the importance of each modality within context.

This demonstrates the ability of the SADGR model to dynamically adjust the contribution weights of different modalities, enabling efficient complementary integration of cross-modal information. It effectively mitigates multimodal redundancy or conflict, thereby improving the accuracy and robustness of emotion recognition.

E. HYPERPARAMETER SENSITIVITY

Beyond architectural design, we validate the model's stability to hyperparameter choices, a critical factor for real-world deployment.

This experiment aims to evaluate the sensitivity and robustness of our model to two key hyperparameters. The first is the weight of the contrastive loss in the STA-CA module, and the second is the number of MLP layers in the AG-VR module. This analysis helps demonstrate the rationality of the selected parameter values and confirms that the model operates stably within a certain parameter range.

Figure 6 consists of two line charts. Subplot (a) displays the curve of F1-Score as a function of the β value (ranging from 0.01 to 1.0). Subplot (b) shows the F1-Score curve relative to the number of MLP layers (from 1 to 4) in the AG-VR module. The optimal parameter values are indicated by red dashed lines.

Both curves exhibit a typical inverted U-shape. When β is too small (e.g., 0.01), the constraint of cross-modal alignment is insufficient to effectively bridge the modality gap, resulting in lower performance. As it increases, performance improves steadily and peaks at 0.1. This suggests an optimal balance between the main task of emotion classification and the auxiliary task of cross-modal alignment at this point.

When β exceeds 0.1, performance begins to decline. This may be due to excessive emphasis on the alignment task, causing the model to ''over-focus'' on inter-modal matching while overlooking more nuanced emotional discriminative cues, thereby interfering with the learning of the main task.

The MLP in the AG-VR module (Subplot b) learns to generate audio-guided gating signals for visual features. Its depth determines modeling capacity. Subplot (b) in Figure 6 shows that even a 1-layer MLP achieves solid performance, proving the baseline effectiveness of the gating mechanism. Performance peaks at 2 layers, where increased nonlinearity captures complex audio-visual dependencies for finer noise filtering. Beyond 2 layers, performance declines, likely due to overfitting on limited data, reducing generalization.

Hyperparameter sensitivity analysis confirms the model's robustness across reasonable parameter ranges.

VI. CONCLUSION

This paper proposes an innovative SADGR model to address two core challenges in multimodal sentiment analysis: modality misalignment and information redundancy.

Firstly, to tackle the semantic gap between modalities, particularly between text and audio, which often hinders effective information fusion, the model introduces a self-supervised text-audio cross-alignment (STA-CA) pre-training phase. This phase leverages self-supervised learning to capture latent emotional correlations between textual and acoustic representations, thereby mitigating inter-modal heterogeneity and improving multimodal fusion accuracy.

Secondly, to resolve the issue of emotional recognition noise caused by redundant video information, SADGR designs an audio-guided video gated refinement (AG-VR) mechanism. Specifically, the model uses emotionally salient audio signals as guidance to dynamically adjust the weights of video frames, adaptively filtering out key emotional frames via a gating mechanism while suppressing irrelevant or interfering information, thereby enhancing the precision and robustness of sentiment recognition. Systematic experiments on mainstream benchmark datasets for multi-modal sentiment analysis demonstrate that SADGR outperforms existing methods in key metrics such as recognition accuracy, highlighting its advanced capabilities and practical value in complex multimodal tasks.

From a system perspective, SADGR demonstrates that explicitly decoupling cross-modal alignment from downstream fusion enables a more efficient and scalable multimodal learning pipeline. By shifting the computational burden of alignment to an offline self-supervised stage and relying on lightweight gating and pooling operations at inference time, the proposed framework achieves a favorable balance between representation quality and deployment efficiency. This design paradigm provides a practical blueprint for building robust multimodal systems that can be readily adapted to real-world, latency-sensitive applications.

While the results presented in this paper are encouraging, the study has certain limitations. Although SADGR avoids heavy cross-modal attention during inference, the use of attention mechanisms in upstream feature extractors may still introduce non-negligible computational overhead in large-scale or highly resource-constrained environments. Therefore, further optimization is required to fully realize real-time deployment under strict latency constraints.

Furthermore, like many existing models, its performance may be sensitive to the quality and nature of the dataset, and its generalization ability in real-world scenarios with greater noise and variability requires further investigation.

For future work, we have identified several promising directions. First, we will explore model compression techniques to develop a more lightweight version of SADGR. Second, we plan to extend our framework to incorporate other modalities, such as physiological signals (e.g., EEG, ECG), to build a more comprehensive emotion recognition system. In conclusion, the principles of dynamic gating and explicit feature alignment proposed in this paper provide a solid and extensible foundation for future research and system design in multimodal machine learning.

ACKNOWLEDGMENT

The authors acknowledge that this study benefited from interdisciplinary discussions enabled by the funded research projects, which facilitated effective integration of algorithmic design, system-level considerations, and empirical evaluation. This work was conducted as a collaborative research effort supported by multiple research funding programs. The conceptual design, methodological development, and model architecture were jointly developed through close collaboration among the core researchers. Software implementation, verification, and experimental validation were carried out with coordinated contributions across institutions. Data preprocessing, result analysis, visualization, and manuscript refinement were performed through distributed yet complementary efforts, ensuring methodological rigor and reproducibility.

REFERENCES

[1] Q. Shi, J. Fan, Z. Wang, and Z. Zhang, ''Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain,'' *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108837, doi: 10.1016/j.patcog.2022.108837.

[2] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, ''Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research,'' *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132, Jan. 2023, doi: 10.1109/TAFFC.2020.3038167.

[3] J. Zhang, Z. Cui, H. J. Park, and G. Noh, "BHGAttN: A feature-enhanced hierarchical graph attention network for sentiment analysis," *Entropy*, vol. 24, no. 11, p. 1691, Nov. 2022, doi: 10.3390/e24111691.

[4] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, New Orleans, LA, USA, Feb. 2018, pp. 5642–5649, doi: 10.1609/aaai.v32i1.12024.

[5] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Feb. 2017, doi: 10.1016/j.inffus.2016.12.001.

[6] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2017, pp. 1103–1114.

[7] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Brussels, Belgium, Jul. 2018, pp. 2247–2256, doi: 10.18653/v1/p18-1209.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[9] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019, pp. 6558–6569.

[10] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -Specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, San Jose, CA, USA, Oct. 2020, pp. 1122–1131, doi: 10.1145/3394171.3413678.

[11] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109259, doi: 10.1016/j.patcog.2022.109259.

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

[13] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, pp. 10790–10797, doi: 10.1609/aaai.v35i12.17289.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2019, pp. 4171–4186, doi: 10.18653/v1/P19-1656.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, Jul. 2020.

[16] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[17] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[19] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[20] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," 2022, *arXiv:2201.02184*.

[21] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, E. Choi, and M. Zhou, "UniVL: A unified video and language pre-training model for multimodal understanding and generation," 2020, *arXiv:2002.06353*.

[22] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," 2019, *arXiv:1908.08530*.

[23] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning robust audio representations from clip," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4563–4567, doi: 10.1109/ICASSP43922.2022.9747669.

[24] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 976–980, doi: 10.1109/ICASSP43922.2022.9747631.

[25] J.-T. Huang, A. Sharma, S. Sun, L. Xia, D. Zhang, P. Pronin, J. Padmanabhan, G. Ottaviano, and L. Yang, "Embedding-based retrieval in Facebook search," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2553–2561, doi: 10.1145/3394486.3403305.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[27] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[28] Y. Ding, J. Yu, B. Liu, Y. Hu, M. Cui, and Q. Wu, "MuKEA: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5079–5088.

[29] Y. Wang, M. Yasunaga, H. Ren, S. Wada, and J. Leskovec, "VQA-GNN: Reasoning with multimodal knowledge via graph neural networks for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21525–21535.

[30] W.-N. Hsu, B. Bolte, Y.-H.-H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] K. Yang, H. Xu, and K. Gao, "CM-BERT: Cross-modal BERT for text-audio sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 521–528, doi: 10.1145/3394171.3413690.

[33] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 5634–5640.

[34] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7216–7223.

[35] Z. Sun, P. K. Sarma, W. A. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 8992–8999.

[36] J. Huang, J. Zhou, Z. Tang, J. Lin, and C. Y.-C. Chen, "TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis," *Knowl.-Based Syst.*, vol. 285, Feb. 2024, Art. no. 111346.

[37] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[38] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.

[39] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[40] Y. Zong, O. M. Aodha, and T. M. Hospedales, "Self-supervised multimodal learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 5299–5318, Jul. 2025, doi: 10.1109/TPAMI.2024.3429301.

[41] R. Liu, H. Zuo, Z. Lian, B. W. Schuller, and H. Li, "Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities," *IEEE Trans. Affect. Comput.*, vol. 15, no. 4, pp. 1856–1873, Oct. 2024, doi: 10.1109/TAFFC.2024.3378570.

[42] P. Liang, Z. Lian, B. Liu, and B. Schuller, "Recent advances in multimodal affective computing: A review," 2024, *arXiv:2409.07388*.

JUNJUN ZHANG received the B.S. degree in engineering from Nanyang Institute of Technology, China, in 2008, the M.S. degree in computer technology from Zhengzhou University, China, in 2017, and the Ph.D. degree in computer engineering from Cheongju University, Republic of Korea, in 2023. From 2009 to 2016, she was a full-time Lecturer in information engineering with Zhengzhou Vocational College of Technology, China. From 2017 to 2019, she was a Lecturer with the School of Software, Zhengzhou University of Industrial Technology, China. Since 2024, she has been an Associate Professor with the School of Software, Henan University of Engineering, China, and a Key Member of Henan Engineering Technology Research Center of Intelligent Transportation Video Image Perception and Recognition. She has published over 15 research papers, holds two national invention patents, and has led or participated in several government- and industry-funded projects in artificial intelligence and software engineering. Her research interests include deep learning, pattern recognition, multimodal data fusion, and intelligent transportation.

JIANJING MAO received the B.S. degree in computer and application from Henan Vocational-Technical Teachers College, China, the M.S. degree in software engineering from Wuhan University, China, in 2010, and the Ph.D. degree in curriculum and instruction (computer) from Central Philippine University, Philippines, in 2023.

From 2004 to 2007, she was a Faculty Member of the Department of Information Engineering, Henan Economics & Management School, China. Since 2007, she has been a Faculty Member with the School of Information Engineering and subsequently with the School of Artificial Intelligence and Big Data, Zhengzhou University of Industrial Technology, China, where she has also been an Associate Professor, since 2019, holding leadership roles with the School of Information Engineering and the School of Artificial Intelligence and Big Data. She is a principal leader of several key provincial initiatives, including Henan Engineering Technology Research Center of Intelligent Transportation Video Image Perception and Recognition, Henan Provincial Intelligent Transportation Big Data Industry-Integration Innovation Center, and Henan Provincial Characteristic Development Program (Data Science and Big Data Technology). She has published over 20 academic papers, holds four patents, and has led or participated in numerous funded projects in the fields of big data and artificial intelligence. Her research interests include deep learning and the application of modern educational technology.

Dr. Mao is an Executive Council Member of Henan Electronics Society, a Council Member of Henan Provincial Higher Education Computer Education Research Association, and a Senior Member of China Computer Federation (CCF).

YANYANG HOU received the B.Eng. degree from Tianjin Ren'ai College, China, in 2010, the M.Eng. degree from Qiqihar University, China, in 2014, and the Ph.D. degree in Management from Central Philippine University, Iloilo City, Philippines, in 2023. From 2017 to 2025, he was a Faculty Member of the School of Information Engineering, Zhengzhou University of Industrial Technology, China. During this period, he has authored more than ten peer-reviewed research papers and led multiple government-funded projects in artificial intelligence. His primary research interests include deep learning, artificial intelligence, and digital image processing.

GISEOP NOH received the B.S. degree in industrial engineering from Korea Air Force Academy, South Korea, the M.S. degree in computer science from the University of Colorado at Denver, USA, in 2009, and the Ph.D. degree in computer science from Seoul National University, South Korea, in 2014.

From 1994 to 2005, he was an Officer with Republic of Korea Air Force, where he was responsible for the operation and maintenance of advanced weapon computer systems. Later, he worked at the Defense Acquisition Program Administration on aviation electronics & computer projects. From 2017 to 2018, he was a Faculty Member with the Department of Computer Science, Korea Air Force Academy. From 2018 to 2024, he was a Professor with the Department of Artificial Intelligence Software, Cheongju University, South Korea. Since 2025, he has been a Professor with the Department of Software Convergence, Hongik University, South Korea, where he established and currently directs the DeepShark Laboratory. His research interests include machine learning, computer vision, retrieval-augmented generation (RAG) systems, graph neural networks, and AI applications in semiconductor inspection and smart manufacturing. He has published over 60 research papers, holds more than 12 registered patents, and has led numerous government- and industry-funded projects in artificial intelligence and software engineering. He is a member of Korean Institute of Information Scientists and Engineers (KIISE) and Korea Information Processing Society (KIPS).

FENGXI ZHANG is currently pursuing the B.Eng. degree in data science and big data technology with Henan University of Engineering, China. He joined Dr. Junjun Zhang's research team, in 2024, where he supports foundational experimental work, including literature review and data processing. He has participated in projects related to multimodal learning and natural language processing (NLP). As the first author, he has applied for and been granted two patents, holds one software copyright, and has contributed to one research paper.