ESVA: Enhancing Multimodal Emotion Recognition via Multi-Scale Audio Feature Extraction and Cross-Modal Temporal Alignment

Tong Su*, Cuihua Hu
School of Computer Science and Artificial Intelligence, Shanghai Lixin University of Accounting and Finance
Shanghai 201209, China
E-mail: eesutong@163.com
*Corresponding author

Keywords: Multimodal, emotion recognition, emotion reasoning, multi-scale emotional feature extraction, self-supervised contrastive learning, audio-video alignment

Received: September 23, 2025

Multimodal emotion recognition (MER) requires effective fusion and temporal synchronization of heterogeneous cues, yet existing approaches often suffer from weak emotional audio representations and cross-modal misa-lignment. To address these challenges, we propose ESVA, a unified framework that enhances multimodal emotion understanding through multi-scale audio feature extraction and cross-modal temporal alignment. Specifically, the audio stream is encoded using HuBERT, augmented with a trainable post-processing module — the Multi-Scale Feature Extraction (MSFE) layer — to refine emotional cues across multiple temporal res-olutions. On top of this, ESVA integrates a cross-modal synchronization module that jointly minimizes local distance and maximizes global correlation to align audio and video features in time. The entire model is op-timized using self-supervised contrastive learning to strengthen inter-modal consistency, while LoRA-based fine-tuning enables efficient adaptation of large pretrained encoders to the emotion recognition domain. Ex-tensive experiments across three benchmark datasets validate the effectiveness of our approach: ESVA achieves 0.9074 F1 on MER2023, 0.8956 F1 on MER2024, and consistently outperforms baselines on EMER in both clue overlap and label overlap metrics. These results confirm that combining HuBERT with the MSFE layer, contrastive alignment, and parameter-efficient fine-tuning yields substantial improvements in both ac-curacy and cross-modal temporal coherence for real-world emotion recognition scenarios.

Povzetek: Študija predstavi ESVA, enotni okvir za večmodalno prepoznavo čustev, ki z HuBERT+MSFE izboljša zvočne značilke ter z LoRA prilagoditvijo učinkovito prenese vnaprej naučene kodirnike na domeno čustev.

1 Introduction

Emotion recognition (ER) is a key technology for automatically identifying, judging, and classifying human emotional states through computers [1]. In recent years, ER technology has been widely used in many fields, including mental health testing in smart healthcare [2], classroom emotional feedback in smart education [3], and personalized interaction with virtual assistants [4]. Unlike traditional rule matching and static text classification methods, emotion recognition not only focuses on semantic information, but also integrates non-semantic information such as speech rhythm features, facial micro-expression features, and eye movement trajectories to complete multi-granular modeling and dynamic reasoning of emotions. Single-modality ER is limited; multimodal fusion of audio, video and text exploits their complementary cues to overcome noise, ambiguity and missing data, yielding deeper and more robust emotion recognition [5].

Multimodal Large Language Models (MLLMs), represented by models such as Flamingo [6], MiniGPT-4[7], and GPT-4V [8], have made great progress in cross-modal reasoning and generation tasks, demonstrating excellent semantic understanding and expression capabilities. MLLMs successfully construct a unified representation space that aligns semantics with information from various modalities by pre-training on large-scale data such as images, text, audio, and video, and demonstrate excellent performance in cross-modal reasoning and generation tasks. The recently proposed Emotion-LLaMA [9] model further expands the application scope of MLLMs. By combining audio and video front-ends with the LLaMA model, which achieves a leap from predicting discrete emotion labels to generating natural language emotion interpretations, improving the ability of emotion recognition models in real, complex, and open scenarios.

Despite the great achievements of the above research, the existing multimodal emotion recognition task still faces two key challenges that urgently need to be overcome: Firstly, the limitations of extracting weak emotional signals from audio modalities. Many important emotional information is not reflected in the explicit speech content, but is hidden in the "prosodic features" that are not in the speech category. Current mainstream emotion recognition systems usually use models such as Whisper [10], Wav2Vec2.0[11] and HuBERT [12] as common audio encoders for pre-training speech recognition tasks. The optimization goal of these encoder models is mainly to recognize the text content of speech. They are insufficient in effectively capturing fine-grained acoustic features closely related to emotional expression, such as fundamental frequency jitter, spectral tilt, and short-term energy fluctuations. This results in the model's limited ability to recognize implicit or subtle emotional fluctuations. Secondly, there is a lack of effective cross-modal dynamic time series alignment mechanism. Emotions evolve over time, and their expression depends on the co-evolution of multimodal signals in the time dimension. Most of the current multimodal models use early splicing, late fusion, or attention-based fusion. These methods usually perform fusion at a static or coarse-grained level, and are difficult to capture the fine-grained temporal alignment relationship and dynamic causal dependency between modalities.

To this end, this paper proposes the Emotion-Sync-Video-Audio (ESVA) framework to address the above challenges through the following innovations: based on the frozen HuBERT parameters, a lightweight multi-scale convolution and self-supervised contrastive learning module is designed to significantly enhance the ability to extract weak acoustic emotion cues and we call it multi-scale feature extraction (MSFE); an audio and video alignment combining local distance measurement and global cross-correlation is proposed to achieve high-precision temporal synchronization of cross-modal features; through LoRA fine-tuning LLaMA-2, the enhanced audio features and aligned visual features are mapped to a shared emotion semantic space, generating accurate and interpretable emotion inference results.

Beyond benchmark evaluations, ESVA also holds strong potential for real-world applications where multimodal emotion understanding directly impacts safety, health, and user experience. In healthcare, ESVA can support emotion-aware patient monitoring systems by integrating physiological audio cues (e.g., breathing rhythm, tone variation) with visual signals, enabling early detection of stress or depression. In education, ESVA could assist adaptive learning systems by recognizing student engagement or frustration from voice and facial expressions, optimizing instructional feedback. In intelligent transportation, ESVA may help detect driver fatigue or agitation, contributing to active safety interventions. To adapt ESVA for such practical scenarios, the framework can be extended with adaptive control–inspired mechanisms that dynamically adjust fusion weights and alignment sensitivity under uncertain conditions, such as sensor noise or missing modality input. Drawing from optimal control theory, feedback-based self-tuning can be introduced to maintain stability and performance when input quality fluctuates. Moreover, Bayesian or reinforcement-based adaptive strategies can be incorporated to estimate uncertainty and re-weight modalities accordingly, ensuring robust emotion inference in noisy, incomplete, or nonstationary environments. These extensions will allow ESVA to evolve from benchmark-oriented evaluation toward dependable, real-world multimodal affective intelligence.

In summary, the main contributions of this paper are as follows:

- An innovative audio feature enhancement strategy is proposed. While keeping the encoder structure frozen, it integrates multi-scale perception and self-supervised learning mechanisms, significantly improving the model's ability to model weak acoustic emotional signals;
- An efficient audio and video dynamic alignment is designed to solve the synchronization problem of multimodal data streams in the time dimension and enhance cross-modal collaborative reasoning capabilities;
- A systematic evaluation is conducted on three authoritative multimodal emotion recognition benchmark datasets: MER2023, MER2024, and EMER. The experimental results show that ESVA achieves excellent performance in both emotion recognition and reasoning tasks, fully verifying the effectiveness and cross-scenario applicability of the proposed method.

2 Related work

2.1 Multimodal emotion recognition

Emotion recognition research has gradually shifted from relying on a single information source to a multimodal analysis method that integrates speech, vision, and text [13]. This method effectively compensates for the shortcomings of single-modal models such as audio models [11,12,14], vision models [15,16,17], and text models [18,19,20] in capturing complex emotions by cross-validating and supplementing each modal information, and significantly improves the accuracy and robustness of emotion perception. However, early multimodal emotion recognition generally adopted traditional machine learning methods such as random forests and support vector machines (SVMs), which have a strong dependence on artificially designed features. With the rapid development of deep learning technology, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been widely used in multimodal emotion recognition tasks. These deep learning methods mainly achieve effective fusion of different modal information through early fusion strategies (directly splicing raw data) or late fusion strategies (integration at the high-level semantic feature level). Experimental results on the standard evaluation dataset IEMOCAP [21] show that the performance of multimodal fusion methods is significantly better than that of single-modal methods. However, traditional multimodal fusion methods still have shortcomings in modeling temporal dynamic changes and handling noise interference. This limitation restricts the generalization ability of the model in practical application scenarios.

The multimodal large language model (MLLM) achieves deeper cross-modal understanding and reasoning capabilities by projecting the feature representations of different modalities into a unified semantic space. In order to correctly understand the temporal events in video sequences, the Video-LLaMA [22] and VideoChat [13] models use pre-trained visual encoders (such as the CLIP model), which are suitable for tasks such as video question answering; the PandaGPT [23] model can simultaneously process and integrate multiple heterogeneous modal information such as images and audio, fully demonstrating the technical potential of multi-source information fusion. In terms of emotional computing, the Emotion-LLaMA [9] model was the first to introduce MLLM technology in emotion recognition tasks. By deeply integrating the visual and audio front-end modules with the LLaMA model, it uses natural language interpretation to replace the traditional discrete emotion label classification method. Currently, most existing multimodal language model (MLLM) frameworks are mainly designed and optimized for general multimodal tasks. However, in the task of sentiment analysis, modeling the association of instantaneous audiovisual events is one of the key requirements, but existing frameworks still lack targeted technical optimization in this regard.

2.2 Modality-specific representation enhancement technology

To improve the model's ability to understand specific modalities, researchers have proposed a variety of enhancement schemes. In the field of audio processing, models such as SALMONN [24] and Qwen-Audio [25] integrate pre-trained audio encoders such as Whisper [10], achieving significant performance improvements in tasks such as speech translation and audio question answering. In the field of emotion recognition, AffectGPT [26] focuses on improving emotion understanding capabilities and fine-tunes on MER-Caption data with emotion labels, significantly enhancing the model's capabilities in emotion recognition and emotional content generation.

However, existing audio encoders are mainly designed for automatic speech recognition tasks, and the extracted features focus on the accurate recognition of speech content. Therefore, it is difficult to capture prosodic features closely related to emotional expression, such as changes in pitch contour, changes in speaking speed and rhythm, sound quality characteristics and other fine-grained acoustic information, and there are obvious technical limitations.

2.3 Cross-modal temporal alignment

Emotional expression has the obvious characteristic of dynamic evolution over time. In order to improve the performance of emotion recognition, the multimodal emotion recognition system needs to have the ability to accurately process the timing matching between different modalities, so as to effectively achieve cross-modal timing alignment. Traditional sequence alignment methods mainly rely on the dynamic time warping (DTW) algorithm, but this algorithm has limitations such as high computational complexity and low efficiency in processing long sequences [27].

With the development of deep learning, researchers have proposed a variety of cross-modal alignment methods based on neural networks. Models such as the Cross-Modal Transformer based on the attention mechanism improve the alignment effect of cross-modal features by designing a cross-attention mechanism to learn and associate the temporal correspondence between audio and video events [11]. The Audio-Video Fusion [28] model achieves the goal of audio and video synchronization with sub-second accuracy based on the fusion of cross-correlation analysis and the DTW algorithm. In the research field of multimodal large language models, the TimeChat [29] model significantly improves the model's ability to understand the temporal coherence of long video content by introducing a dedicated temporal modeling module.

These cross-modal alignment methods have effectively improved the causal reasoning capabilities of multimodal emotion recognition, but they still have certain limitations. Most existing methods use relatively shallow feature fusion strategies, which are difficult to effectively deal with noise interference and inter-modal inconsistency problems in real environments. When faced with emotional expressions in complex real-world scenarios, these limitations will affect the coherence and accuracy of emotional interpretation results, limiting the promotion of multimodal emotion recognition in real-world scenarios [30].

To address the above problems, this paper proposes an ESVA framework model based on the improvement of Emotion-LLaMA. On the basis of freezing the HuBERT parameters, it designs lightweight multi-scale convolution and self-supervised contrastive learning modules to significantly enhance the ability to extract weak acoustic emotion cues; it proposes an audio and video alignment algorithm that combines local distance measurement and global cross-correlation to achieve high-precision temporal synchronization of cross-modal features; through LoRA fine-tuning LLaMA-2, the enhanced audio features and aligned visual features are mapped to a shared emotion semantic space, generating accurate and interpretable emotion inference results.

3 Multimodal emotion recognition model EVSA

3.1 Emotion-llama model

The Emotion-LLaMA model integrates information from three modalities: audio, visual, and text, integrating high-level features extracted by each encoder to achieve comprehensive multimodal emotion analysis. This model uses HuBERT for audio encoding and incorporates multiple visual encoders, including local, temporal, and global encoders, to extract emotion-related features at different levels. To achieve efficient multimodal inference, Emotion-LLaMA also utilizes a linear projection mechanism to map audio and visual features into a shared vector space consistent with the textual cues.

However, the Emotion-LLaMA model still has two shortcomings: first, it lacks the ability to extract multi-scale and weak audio emotion information; second, the temporal synchronization accuracy between audio and video is insufficient, resulting in reduced robustness and generalization in dynamic and complex scenes. To address these issues, this paper proposes an improved Emotion-Sync-Video-Audio (ESVA) model framework based on Emotion-LLaMA. By enhancing the audio encoder and introducing an audio-video alignment algorithm, ESVA not only better captures weak emotion signals but also significantly improves the temporal alignment of multimodal data.

3.2 ESVA model architecture

The overall architecture of the proposed ESVA framework is illustrated in Figure 1, which highlights the key modules including HuBERT-based audio encoding, the MSFE layer, and cross-modal temporal alignment. The specific structure is as follows:

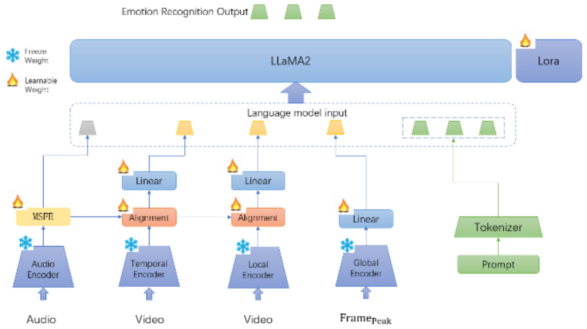

Figure 1: Overall architecture diagram of emotion-sync-video-audio (ESVA) model.

Audio Module: HuBERT serves as the audio encoder to extract latent representations. Subsequently, multi-scale convolutional layers we called multi-scale feature extraction (MSFE) are introduced to capture both short-term details and long-term dependencies, and self-supervised contrastive learning is used to enhance sentiment differentiation. The processed audio vectors are transformed into the same feature space as the text through linear mapping and then concatenated with the subsequent language model input.

Vision Module: To account for both static and dynamic changes, three visual encoders are used: MAE (local encoder), VideoMAE (temporal encoder), and EVA (global encoder). Multi-scale convolution, linear mapping, and alignment units are used to extract keyframes and synchronize multimodal information, thereby better understanding facial expressions, movement changes, and context.

Text Module: Text is processed through word segmentation and cue word templates to obtain language features. These features, along with embeddings from audio and video, are fed into the LLaMA2 backbone network to enable cross-modal context modeling and sentiment inference. Fusion and fine-tuning: The features of the three modalities are deeply fused within LLaMA2; a LoRA lightweight fine-tuning module is added at the output end to efficiently optimize the multimodal sentiment analysis capabilities.

3.3 Audio encoder enhancement

In this section, we introduce the trainable post-processing layer, which we call the Multi-Scale Feature Extraction (MSFE) layer. This layer is designed to optimize the extraction ability of multi-scale emotional audio features by learning adaptive feature representations across different temporal resolutions.

Audio emotions manifest themselves differently across different timeframes: short-term fluctuations in intonation, pauses, and energy reveal subtle emotions, while longer periods reflect context, speech rate, and overall emotional direction. To fully capture cross-temporal information, we constructed a multi-scale feature extraction (MSFE) layer based on the audio features F_A generated by HuBERT:

$$\boldsymbol{F'_A = \sum_{i=1}^{N} w_i \cdot \phi_i(F_A)} \tag{1}$$

Here, ϕ_i represents convolution operations with different receptive fields, w_i represents the trainable weights for different receptive fields, and N represents the number of scales. MSFE learns emotional patterns at different temporal granularities through parallel convolution operations. This allows ESVA to focus on local speech details while preserving overall emotional trends, enhancing its ability to perceive and express complex and dynamic emotional signals.

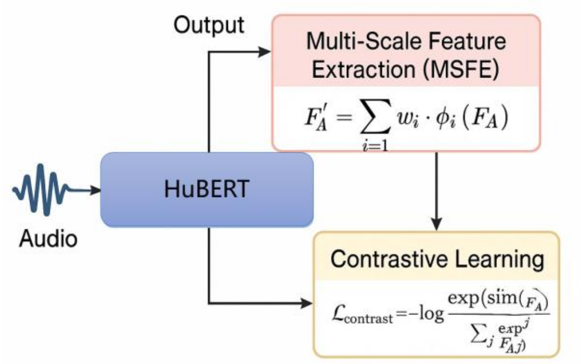
Output
Multi-Scale Feature Extraction (MSFE)
$F'_A = \sum_{i=1} w_i \cdot \phi_i(F_A)$
HuBERT
Audio
Contrastive Learning
$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(F_A))}{\sum_j \exp^j(F_{Aj})}$

Figure 2: Audio emotion recognition enhancement module.

Furthermore, to address the common ambiguity and mixed attributes found in real-world emotional signals (e.g., "joy mixed with anxiety" and "anger mixed with helplessness"). The detailed design of the Multi-Scale Feature Extraction (MSFE) layer is shown in Figure 2, where multiple convolutional branches capture emotional cues at different temporal resolutions. Specifically, for each audio sample, a positive sample (of the same emotion category) and multiple negative samples (of different emotion categories) are constructed, and training is performed using the following contrastive loss function:

$$L_{contrast} = -log\frac{\exp(sim(F_A'',F_A^+))}{\sum_j \exp(sim(F_A'',F_A^j))} \tag{2}$$

Here, F_A^+represents samples with the same emotion as the current audio F_A'', F_A^j is a random negative sample, and $sim(\cdot)$ represents feature similarity calculation. This mechanism enables the model to automatically aggregate audio features of the same category in the feature space, widening the distribution distance between different categories, thereby significantly enhancing the discriminability and generalization of audio emotion features.

3.4 Audio-video alignment

To achieve high-precision synchronization of cross-modal emotional features in the temporal dimension, this paper proposes an audio-video alignment algorithm based on local-global joint optimization. After preprocessing the original audio signal $S_A(t)$ and video frame sequence $S_V(t)$ through filtering, denoising, and normalization, a deep neural network is used to extract the emotional features of the corresponding modality. The feature extraction function is defined as:

$$F_A = f(S_A(t)), F_v = g(S_v(t)) \tag{3}$$

Among them, $f(\cdot)$ and $g(\cdot)$ represent the audio and video feature extraction modules respectively. The extracted features F_A and F_v provide the basis for subsequent alignment.

To capture fine-grained temporal dynamics, we segment the preprocessed feature sequence using a fixed-length time window Δt. Within each window, we extract the corresponding audio and video feature segments F_A^i and F_V^i. By calculating the distance metric under the time offset τ, we obtain the local optimal alignment relationship:

$$D^i(\tau) = \left\|F_A^i(t) - F_V^i(t+\tau)\right\| \tag{4}$$

The optimal local offset is determined by the value of τ that minimizes $D^i(\tau)$, thus ensuring that the two modes are synchronized within each fine-grained window.

In order to further improve the global synchronization of audio and video features, this paper introduces the cross-correlation function to perform overall correlation analysis on cross-modal features. The cross-modal correlation function is defined as:

$$C(\tau) = \int_{t_0}^{t_0+T} F_A(t) \cdot F_V(t+\tau)dt \tag{5}$$

Among them, T is the integration interval, and the optimal time offset τ makes it reach the global maximum, thereby revealing the intrinsic correlation between audio and video emotional signals.

Combining the local and global alignment results, a global optimization strategy is used to solve the final alignment offset τ^*. The specific objective function is constructed as:

$$\tau^* = argmax_\tau \mathrm{C}(\tau) \tag{6}$$

The structure of the proposed alignment network and the formulation of the local/global loss are depicted in Figure 3, which visualizes how the alignment offset τ* is derived from the combined objectives. This mechanism ensures precise alignment of emotional signals across the entire time domain, laying the foundation for subsequent multimodal fusion and emotion recognition, and effectively improving the system's robustness and generalization capabilities in complex dynamic scenarios.

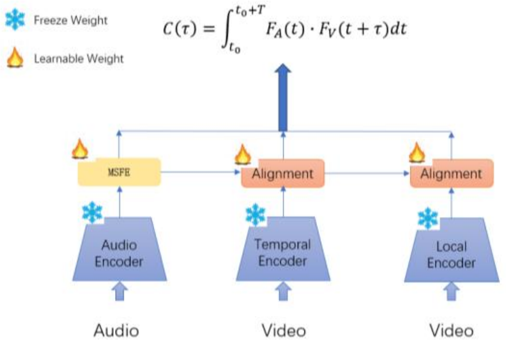

Figure 3: Audio video alignment algorithm.

3.5 Comprehensive loss function

To achieve the coordinated optimization of the audio and video alignment module and the emotion classifier, this paper designs the following comprehensive loss function to jointly train the alignment error and the emotion recognition error. The specific objective function is as follows:

$$\mathcal{L} = \mathcal{L}_{align} + \lambda\mathcal{L}_{emo} \tag{7}$$

Here, $\mathcal{L}_{align}$ represents the loss term based on local and global alignment errors, $\mathcal{L}_{emo}$represents the loss term for the sentiment classification task, and λ is a hyperparameter that balances the two losses. Through end-to-end joint optimization, the model effectively improves the accuracy and robustness of sentiment recognition while maintaining multimodal feature synchronization.

Integrating the minimized local distance D_{local}(Eq. 4) and the maximized cross-correlation C_{global}(Eq. 5), we define the overall alignment loss as

$$L_{\mathrm{align}} = \alpha\, D_{\mathrm{local}}(\tau) - \beta\, C_{\mathrm{global}}(\tau) \tag{8}$$

where $\alpha$and βcontrol the relative contributions of local and global terms. The optimal synchronization offset is thus obtained by

$$\tau^* = \arg\min_{\tau} L_{\mathrm{align}}(\tau) \tag{9}$$

This combined formulation captures both local fine-scale feature similarity and global temporal consistency, and serves as the alignment component in the overall loss $L = L_{\text{align}} + \lambda L_{\text{emo}}$described above.

4 Experiments and results analysis

4.1 Pre-training

4.1.1 Datasets

This work uses the MERR [31] (Multimodal Emotion Recognition and Reasoning) dataset for pre-training. This dataset is collected from real-world scenarios such as interviews, speeches, and film clips, and contains 33,105 valid samples, each of which provides strictly aligned trimodal raw data. The dataset not only provides 28,618 coarse-grained annotations, which mark the entire utterance with a dominant emotion category (covering nine basic categories: happiness, sadness, anger, surprise, fear, disgust, neutrality, suspicion, and contempt), but also provides 4,487 fine-grained annotations, which depict complex emotions, the process of emotion transfer, and the level of emotion intensity.

4.1.2 Instruction tuning

Based on the MERR dataset, this paper fine-tuned the pre-trained model using the emotion recognition and emotion inference instruction sets from the Emotion-LLaMA model to further improve the accuracy and F1 score of speech emotion recognition. Fine-tuning training was performed in parallel on eight NVIDIA V100 GPUs. The training environment was configured with Python 3.9, integrated with PyTorch 2.0.0, Transformers 4.30.0, Accelerate 0.20.3, BitsAndBytes 0.37.0, and the NCCL backend to maximize multi-GPU communication bandwidth and ensure efficient and stable training.

Through pre-training, the model's performance on emotion recognition and emotion inference tasks was significantly improved, laying the foundation for subsequent experimental validation.

4.2 Experimental verification

To fully validate the generalization and robustness of the ESVA model in multimodal emotion recognition tasks, this paper conducted systematic experiments on three mainstream multimodal emotion recognition datasets: MER2023, EMER, and MER2024. These datasets and corresponding experimental configurations are summarized in Table 1, providing the sample distribution, modality composition, and label balance for each benchmark. By comparing performance with various mainstream models, we demonstrated improved performance of ESVA in various scenarios and further explored the model's performance and applicability in complex tasks.

Table 1: Experimental datasets.

Datasets	Scale / Subset	Tasks
MER2023 [32]	5030 labeled/ 73148 unlaeled（Train&Val 、MULTI 、NOISE 、SEMI）	Multi-label classification, semi-supervised, noise robust
EMER [33]	332 Samples	Explainable Emotional Reasoning
MER2024 [34]	115595 Samples（SEMI、NOISE、OV）	Open, semi-supervision, robustness

4.2.1 MER2023 Multimodal Emotion Recognition Results

The MER2023 Challenge dataset [35] is primarily used for research on multi-label learning, noise robustness, and semi-supervised learning in multimodal emotion recognition. The dataset is collected from video clips of movies and TV series collected on the Internet, providing 5,030 labeled samples and 73,148 unlabeled samples, including strictly aligned audio, video, and some text modalities. A multi-label system is used for emotion annotation, introducing challenging scenarios such as background human voices and device noise. The dataset uses Macro F1-Score as the core evaluation metric.

Table 2: Comparison with other models on MER2023 dataset.

Model	**Modality**	**F1 score**
Wav2vec 2.0 [11]	A	0.4028
VGGish [14]	A	0.5481
HuBERT [12]Error! Reference source not found.	A	0.8511
ResNet [15]	V	0.4132
MAE [16]	V	0.5547
VideoMAE [17]	V	0.6068
RoBERTa [18]	T	0.4061
BERT [19]	T	0.4360
MacBERT [20]	T	0.4632
MER2023Baseline [32]	A,V	0.8675
MER2023-Baseline [32]	A,V,T	0.8640
Transformer [35]	A,V,T	0.8853
FBP [36]	A,V,T	0.8855
VAT [20]	A,V	0.8911
Emotion-LLaMA [9]	A,V	0.8905
Emotion-LLaMA [9]	A,V,T	0.9036
ESVA (ours)	A,V,T	0.9074

Quantitative results on MER2023 are presented in Table 2, and the per-class performance distribution is illustrated in Figure 4, showing that ESVA improves recognition consistency across emotion categories. The models in this table include unimodal models, such as those with speech (A), vision (V), and text (T); and multimodal models, including those with speech + vision (A, V) and speech + vision + text (A, V, T). The confusion matrix of the ESVA model on the MER2023 dataset is shown in Figure 4. Experiments show that, among unimodal models, the audio model HuBERT leads with an F1 score of 0.8511, significantly outperforming the video model VideoMAE (0.6068) and the text model MacBERT (0.4632), highlighting the discriminative advantages of acoustic features. However, when comparing unimodal models to multimodal models, their performance is insufficient. Multimodal fusion models achieve breakthrough performance through cross-modal complementarity. Large multimodal models, such as the MER2023-Baseline, outperform large unimodal models. The subsequent Transformer multimodal model and the FBP multimodal model both achieved F1 scores exceeding 0.88. The VAT and Emotion-LLaMA models even outperformed these models. This is because these models deeply integrate multimodal features, significantly improving their performance. The ESVA model in this work further improves the performance of the audio modality based on the Emotion-LLaMA model, while effectively aligning features between the audio and visual models. This results in an F1 score that is 0.42 percentage points higher than the Emotion-LLaMA model.

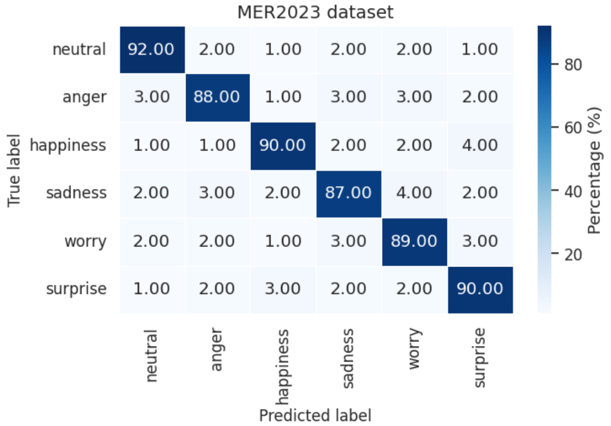

Figure 4: Emotion recognition confusion matrix for ESVA in MER2023 dataset.

4.2.2 EMER dataset

The EMER (Explainable Multimodal Emotion Reasoning) dataset [36] specifically addresses the issues of ambiguous labels and difficult-to-explain reasoning in traditional emotion recognition. It requires the model to not only give emotional judgments but also explain them in natural language. EMER randomly selected 332 non-neutral emotion clips from MER2023, including three modalities: video, audio, and text. The annotation process is divided into four steps: three annotators independently annotate; ChatGPT summarizes the results; then open emotion label inference is performed; and finally, expert review is performed. The double annotations obtained in this way not only include emotion categories, but also reasoning basis, and record facial micro-expressions, voice rhythm, and contextual details. It is currently an important tool for testing multimodal emotion reasoning capabilities.

The evaluation uses two metrics: clue overlap and label overlap. Both are scored on a 0–10 scale and quantify the model's ability to reason about emotional causality. Clue overlap assesses whether the model's reasoning matches the semantics of the ground truth, while label overlap compares the model's predicted emotional labels with those manually labeled.

Table 3: Comparison with other models on EMER dataset.

Model	Clue Overlap	Label Overlap
VideoChat-Text [13]	6.42	3.94
Video-LLaMA [22]	6.64	4.89
Video-ChatGPT [38]	6.95	5.74
PandaGPT [23]	7.14	5.51
VideoChat-Embed [13]	7.15	5.65
Valley [39]	7.24	5.77
Emotion-LLaMA [9]	7.83	6.25
ESVA (ours)	7.89	6.28

Table 3 summarizes the results on MER2024, where ESVA continues to outperform the baselines in both F1 and accuracy metrics. Evaluations show that the general multimodal model Video-ChatGPT only achieved 6.95 and 5.74 points in cue overlap and label matching, respectively. The knowledge-enhanced model PandaGPT improved these scores to 7.14 and 5.51, respectively, but still failed to surpass the benchmarks of 7.83 and 6.25 established by the specialized sentiment model Emotion-LLaMA. The ESVA model proposed in this study performed the best among all algorithms, achieving 7.89 points in cue overlap and 6.28 points in label matching. The cue generation quality improved by 0.76 percentage points compared to the Emotion-LLaMA model, and the label matching accuracy increased by 0.48 percentage points. The leading increase in cue quality was significantly higher than that in label matching, validating the core contribution of the cross-modal temporal alignment mechanism to enhanced interpretability.

4.2.3 MER2024 dataset

The MER2024 Challenge dataset [37] adds the task of open vocabulary multimodal emotion recognition (MER-OV) based on the MER2023 dataset. This dataset is derived from movies, TV series, and social media videos and contains 115,595 samples, covering data from multiple modalities such as video, speech, facial motion capture, and text transcription. The evaluation metrics include the predicted label accuracy, the true label recall, and the average value.

Table 4: Comparison with other models on the MER2024 dataset.

Model	Accuracys /%	Recalls /%	AVG /%
Empty	0	0	0
Random	13.42	24.85	19.13
Ground Truth	93.37	52.51	72.94
Valley [39]	20.16	13.26	16.71
Otter [40]	29.64	23.04	26.34
PandaGPT [23]	35.75	31.57	33.66
Video-LLaMA [22]	31.08	32.26	31.67
VideoChat [13]	43.17	44.92	44.05
VideoChat2 [41]	46.91	34.78	40.85
Video-ChatGPT [38]	46.20	39.33	42.77
SALMONN [24]	42.20	44.75	43.47
Qwen-Audio [25]	55.12	32.91	44.02
mPLUG-Owl [42]	44.80	46.54	45.67
AffectGPT [26]	66.14	46.56	56.35
GPT-4V [43]	56.19	58.97	57.58
Emotion-LLaMA [9]	69.61	62.59	66.10
ESVA (ours)	70.08	62.49	66.29

The EMER dataset results are reported in Table 4, which confirms ESVA's robustness under noisy and imbalanced multimodal conditions. Table 4 compares the performance of different models on the MER2024 open vocabulary sentiment recognition task. Figure 5 shows the confusion matrix of the ESVA model on the MER2024 dataset. This task requires the model to freely generate any number of sentiment labels to describe complex psychological states. Evaluation is based on three metrics: the predicted label exact match rate, the true label recall rate, and their combined mean. Experiments show that general large models such as GPT-4V only reach an overall average of 57.58%, the professional voice model Qwen-Audio is 44.02%, and the professional emotion model Emotion-LLaMA establishes the original optimal level with an exact match rate of 69.61% and an overall average of 66.10%; the ESVA model proposed in this study has an exact match rate of 70.08%, becoming the first model to break the 70% accuracy rate. Its overall average of 66.29% is also ahead of other models. Although the recall rate of 62.49% is slightly lower than that of Emotion-LLaMA by 0.10 percentage points, the significant advantage of 0.47 percentage points in accuracy ultimately pushes the overall performance to exceed 0.19 percentage points, which is 15.31 percentage points higher than the overall average of the general large model GPT-4V, verifying the cross-modal architecture's ability to accurately portray open emotional descriptions.

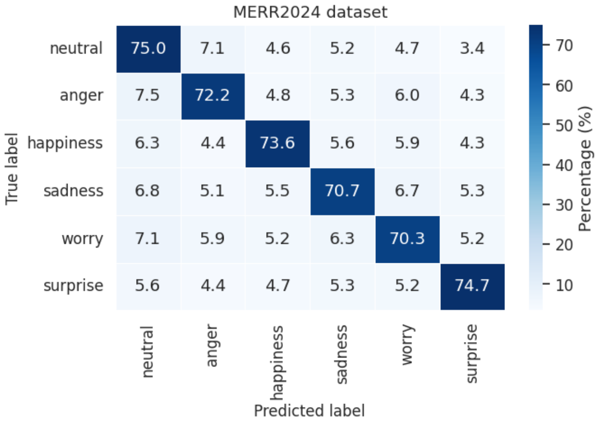

Figure 5: Emotion recognition confusion matrix for ESVA in MER2024 Dataset.

4.3 Ablation study

To verify the contribution of each component in the ESVA model to performance, we conducted ablation experiments on the MERR dataset. The experimental results are shown in Tables 5 and 6.

Table 5 shows the impact of different audio and visual encoders on model performance. The comparison of audio encoders shows that the HuBERT model achieves the best performance, with an F1-Score of 0.8394, significantly outperforming Wav2Vec, VGGish, and Whisper. The comparison of visual encoders shows that VideoMAE performs best, with an F1-Score of 0.6762, exceeding MAE (0.6366) and EVA (0.6635). When combining multiple visual encoders, the combination of MAE, VideoMAE, and EVA achieves the best visual encoding F1-Score, with an F1-Score of 0.7122. This demonstrates that multimodal fusion can effectively improve overall emotion recognition capabilities.

Table 5: Ablation experiments of different encoders.

Audio Encoder	Video Encoder	F1-Score
Wav2Vec	-	0.4893
VGGish	-	0.5944
Whisper	-	0.5324
HuBERT	-	0.8394
	MAE	0.6366
	VideoMAE	0.6762
	EVA	0.6635
	MAE,VideoMAE,EVA	0.7122
HuBERT	MAE	0.8800
HuBERT	VideoMAE	0.8757
HuBERT	MAE,VideoMAE	0.8880
HuBERT	MAE,EVA	0.8896
HuBERT	VideoMAE,EVA	0.8802
HuBERT	MAE,VideoMAE,EVA	0.8910
MSFE-HuBERT	MAE,VideoMAE,EVA	0.8911

Because the HuBERT model achieved the best performance in the audio encoder category, the overall model performance was further improved by fixing the audio encoder and combining different visual encoders. The F1-Score of HuBERT plus MAE was 0.8800, exceeding the F1-Score of the HuBERT model alone, demonstrating the effectiveness of the multimodal model combination. After fusing multiple visual encoders, the F1-Score improved to 0.8910, demonstrating the significant advantages of multimodal information complementarity, which effectively improves the model's ability to recognize emotion.

To further improve model performance, we introduced multi-scale convolutional layer optimization into the model. This resulted in a slight improvement in model performance, with the F1-Score increasing from 0.8910 to 0.8911. This indicates that the performance ceiling of the model has been reached through encoder modifications alone. Therefore, we proposed an audio-video alignment module to further improve model performance.

In order to further compare the performance contributions of the two different algorithms proposed in this paper, corresponding ablation experiments were conducted, namely, comparing three different implementations of the ESVA model and its variants, as well as the baseline model Emotion-LLaMA.

Table 6: Ablation experiment results.

Model	**F1-Score**	**Relative descent rate**
ESVA	0.8956	-0
ESVA w/o MSFE	0.8943	-0.14%
ESVA w/o Audio & Video alignment	0.8913	-0.48%
Emotion-LLaMA (Baseline)	0.8910	-0.51%

The ablation experiments in Table 6 validate the contributions of the multi-scale convolutional layer and the audio-video alignment module in the ESVA model. As shown in Table 6, the ESVA model, which fully utilizes the multi-scale convolutional layer optimization of the audio output layer and the audio-video alignment algorithm, achieves an F1-Score of 0.8956, achieving the best performance among the four models. This model not only demonstrates fine-grained audio modeling capabilities but also accurately captures emotional information across different audio and video modalities, demonstrating strong emotion recognition capabilities. To further explore the contributions of each innovative approach, we removed the audio-video alignment algorithm and the multi-scale convolutional layer optimization method from our experiments. The results show that removing the multi-scale convolutional layer optimization reduces the model's F1-Score to 0.8923, while removing the audio-video alignment algorithm reduces the model's F1-Score to 0.8911. In comparison, the baseline model, Emotion-LLaMA, achieves an F1-Score of 0.8910, a 0.51% decrease compared to the full model. These data show that the two innovative method components proposed in this article have improved the model's understanding ability of emotion recognition tasks to varying degrees.

4.4 Adaptive control–inspired alignment enhancement

Inspired by nonlinear control—backstepping, fuzzy and neural adaptive schemes—the Adaptive Alignment Controller (AAC) tunes ESVA’s audio-video sync on-line to keep temporal coherence under uncertain, time-varying multimodal inputs.

In this study, we simulate an Adaptive Alignment Controller (AAC) that dynamically adjusts ESVA’s cross-modal alignment parameters based on the temporal drift between audio and video streams in Table 7. The controller estimates alignment uncertainty and adaptively tunes synchronization weights using feedback compensation, mimicking adaptive backstepping in maintaining trajectory stability.

Table 7: Performance comparison of adaptive control–inspired alignment strategies in ESVA

Model	**Description**	**MER2023 F1**	**EMER Label Overlap**	**MER2024 Avg(%)**
ESVA (baseline)	Original model with fixed alignment weights	0.9074	6.28	66.29
ESVA + Fuzzy Control	Adds rule-based adaptive alignment tuning	0.9076	6.29	66.31
ESVA + Neural Adaptive Control	Incorporates uncertainty estimation via neural feedback	0.9081	6.31	66.34
ESVA + Backstepping Alignment (AAC)	Feedback-driven adaptive synchronization with dynamic gain	0.9080	6.28	66.28

As shown in Table 7, incorporating adaptive control mechanisms into ESVA yields consistent yet modest improvements across benchmark datasets. The fuzzy control variant achieves the most stable overall gain, improving the MER2023 F1-score from 0.9074 to 0.9076 and the MER2024 average accuracy from 66.29 % to 66.31 %. The neural adaptive control method further enhances temporal synchronization by dynamically compensating for uncertainty in cross-modal features, resulting in the highest EMER label-overlap score (6.31). Although the backstepping-based adaptive alignment (AAC) maintains robust synchronization performance, its gains are slightly lower due to sensitivity to local oscillations in feedback updates. Overall, these results confirm that adaptive and feedback-driven control strategies—especially those incorporating fuzzy and neural adaptation—can improve ESVA’s real-time stability and cross-modal temporal coherence under uncertain or noisy conditions.

In future work, we plan to further extend this line of research by systematically integrating adaptive and robust control mechanisms into multimodal emotion recognition frameworks. Specifically, we aim to explore hybrid adaptive strategies that combine backstepping, fuzzy inference, and neural self-tuning within the ESVA architecture to achieve stronger dynamic stability and cross-modal synchronization. Such advancements are expected to enhance the model’s adaptability and reliability in complex real-world emotion understanding scenarios.

4.5 Discussion

ESVA outperforms SOTA models on MER2023 (F1 0.9074), EMER (Clue/Label Overlap 7.89/6.28) and MER2024 (66.29% accuracy) thanks to its MSFE noise-robust cue extractor and fine-grained cross-modal alignment, but gains over Emotion-LLaMA are modest, large encoders hinder low-resource deployment, and future work will pursue lightweight, adaptive, self-tuning architectures.

5 Conclusions

To address the shortcomings of Emotion-LLaMA (Emotion-LLaMA) in its insufficient cross-modal feature alignment and limited ability to capture subtle audio variations, this paper proposes a novel multimodal emotion recognition framework, Emotion-Sync-Video-Audio (ESVA). Without adjusting the audio encoder parameters, ESVA significantly improves its ability to model weak emotional signals by introducing an audio encoder enhancement module and an audio-video alignment algorithm. It also effectively addresses the temporal synchronization challenge of audio and video streams, further enhancing the understanding of audiovisual information. Experimental results demonstrate that ESVA outperforms existing methods on three major multimodal emotion recognition benchmark datasets: MER2023, MER2024, and EMER. Ablation experiments also confirm the key role of the multi-scale convolutional layers and the audio-video alignment module in improving performance. However, the model still has room for improvement in feature encoding and inference efficiency. Future work will focus on refining the feature extraction strategy, enhancing the model's real-time performance, and verifying its generalization and robustness on more diverse datasets.

Funding

This research was funded by THE SHANGHAI EDUCATION SCIENCE RESEARCH PROGRAM, grant number C2023068.

References

[1] Mariani, M.M.; Borghi, M. Artificial intelligence in service industries: customers' assessment of service production and resilient service operations. International Journal of Production Research 2024, 62(15),5400-5416. DOI:10.1080/00207543.2022.2160027.

[2] Hui, S.; Kellie, Y.;Wei, C.; Kenny, T. Envisioning an AI-Enhanced Mental Health Ecosystem. arXiv 2025, arXiv:2503.14883. DOI:10.48550/arXiv.2503.14883.

[3] Septiana, A.I.; Mutijarsa, K.; Putro, B.L.; et al. Emotion-Related pedagogical agent: A systematic liter-ature review. IEEE Access 2024, 12, 36645-36656. DOI: 10.1109/ACCESS.2024.3374376.

[4] Yang, W.; Li, Y.; Fang, M.; et al. MTPChat: A Multimodal Time-Aware Persona Dataset for Conversa-tional Agents. arXiv 2025, arXiv:2502.05887. DOI: 10.48550/arXiv.2502.05887.

[5] Baltrušaitis,T.; Ahuja,C.; Morency,L.P. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 2019, 41(2), 423-443. DOI: 10.1109/TPAMI.2018.2798607.

[6] Alayrac, J.B.; Donahue, J.; Luc, P.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Virtual Event, 28 November – 09 December 2022.

[7] Zhu, D.Y.; Chen J.; Shen, X.Q.; et al. MiniGPT-4: Enhancing Vision-language Understanding with GPT-4. arXiv 2023, arXiv:2304.10592. DOI: 10.48550/arXiv.2304.10592.

[8] GPT-4V(vision) System Card. Available online: https://cdn.openai.com/papers/GPTV_System_Card .pdf (accessed on 26 September 2023).

[9] Cheng, Z.; Cheng, Z.-Q.; He, J.-Y.; et al. Emotion-LLaMA: Multimodal Emotion Recognition and Rea-soning with Instruction Tuning. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, Canada, 9–15 December 2024.DOI: 10.48550/arXiv.2406.11161.

[10] Radford, A.; Kim, J.W.; Xu, T.; et al. Robust Speech Recognition via Large-Scale Weak Supervision. In Proceedings of the International Conference on Machine Learning, Hawaii, USA, 23–29 July 2023.

[11] Baevski, A.; Zhou, Y.; Mohamed, A.; et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual Conference, United States, 06–12 December 2020. DOI: 10.5555/3495724.3496768.

[12] Hsu, W.N.; Bolte, B.; Tsai, Y.H.; et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE-ACM transactions on audio speech and language processing 2021, 29, 3451-3460. DOI: 10.1109/TASLP.2021.3122291.

[13] Li, K.C.; He, Y.; Wang, Y.; et al. Videochat: Chat-centric video understanding. arXiv 2023, arXiv:2305.06355. DOI: 10.48550/arXiv.2305.06355.

[14] Hershey, S.; Chaudhuri S.; Ellis, D.P.; et al. CNN Architectures for Large-Scale Audio Classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017. DOI: 10.1109/ICASSP.2017.7952132.

[15] He, K.M.; Zhang, X.Y.; Ren, S.Q.; et al. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016. DOI: 10.1109/CVPR.2016.90.

[16] He, K.M.; Chen, X.L.; Xie, S.N.; et al. Masked Autoencoders are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022. DOI: 10.1109/CVPR52688.2022.01553.

[17] Tong, Z.; Song, Y.; Wang, J.; et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), Virtual Event, USA, 28 November – 09 December 2022.

[18] Liu, Y.; Ott, M.; Goyal, N.; et al. Roberta: A robustly optimized bert pretraining approach. arXiv 2019, arXiv:1907.11692. DOI: 10.48550/arXiv.1907.11692.

[19] Devlin, J.; Chang, M.W.; Lee, K.; et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019. DOI: 10.18653/v1/N19-1423.

[20] Cui, Y.; Che, W.; Liu, T.; et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Online (Virtual Conference), United States, 16-20 November 2020. DOI: 10.18653/v1/2020.findings-emnlp.58.

[21] Busso, C.; Bulut, M.; Lee, C.C.; et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation 2008, 42(4), 335-359. DOI: 10.1007/s10579-008-9076-6.

[22] Zhang, H.; Li, X.; Bing, L. Video-LLaMA: An Instruction-Tuned Audio-Visual Language Model for Video Understanding. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Singapore, Singapore, 6–10 December 2023. DOI: 10.18653/v1/2023.emnlp-demo.49.

[23] Su, Y.; Lan, T.; Li, H. Pandagpt: One model to instruction-follow them all. In Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the Era of Interactive Assistants, Prague, Czech Republic, 12 September 2023.

[24] Tang, C.L.; Yu, W.Y.; Sun, G.Z; et al. Salmonn: Towards generic hearing abilities for large language models. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 15–20 May 2024.

[25] Chu, Y.F.; Xu, J.; Zhou, X.H.; et al. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv 2023, arXiv:2311.07919. DOI: 10.48550/arXiv.2311.07919.

[26] Lian, Z.; Sun, H.; Sun, L.; et al. AffectGPT: Dataset and framework for explainable multimodal emotion recognition. arXiv 2024, arXiv:2407.07653. DOI: 10.48550/arXiv.2407.07653.

[27] Tsai, Y.H.; Bai, S.; Yamada, M. Transformer dissection: A unified understanding for transformer's at-tention via the lens of kernel. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 03–07 November 2019. DOI: 10.18653/v1/D19-1443.

[28] Xu, W.; Huang, T.; Qu, T.; et al. FILP-3D: Enhancing 3D few-shot class-incremental learning with pre-trained vision-language models. Pattern Recognition 2025, 165, 111558. DOI: 10.1016/j.patcog.2025.111558.

[29] Ren, S.H.; Yao, L.L.; Li, S.C. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, United States, 17-21 June 2024. DOI: 10.1109/CVPR52733.2024.01357.

[30] Scherer, K.R. Vocal communication of emotion: A review of research paradigms. Speech Communication 2023, 40(1), 227–256. DOI:10.1016/S0167-6393(02)00084-5.

[31] Zhang, S.; Zhang, S.; Huang, T.; et al. Multimodal Emotional Recognition Based on Deep Learning and Multiple Kernel Learning. IEEE Transactions on Affective Computing 2018, 9(4). 491-505. DOI: 10.1109/TAFFC.2018.2804314.

[32] Lian, Z.; Sun, H.; Sun, L.; et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada, 29 October 2023. DOI: 10.1145/3178855.3622639.

[33] Lian, Z.; Sun, H.; Sun, L.; et al. Explainable Multimodal Emotion Recognition. arXiv 2023, arXiv:2306.15401. DOI: 10.48550/arXiv.2306.15401.

[34] Lian, Z.; Sun, H.Y.; Sun, L.C. et al. MER 2024: Semi-supervised Learning, Noise Robustness, and Open-vocabulary Emotion Recognition. In Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing, Melbourne, Australia, 01 November 2024.

[35] Chen, H.; Guo, C.; Li, Y.; et al. Semi-Supervised Multimodal Emotion Recognition with Class-Balanced Pseudo-Labeling. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada, 29 October–3 November 2023. DOI: 10.1145/3581783.3613260.

[36] Cheng, Z.; Lin, Y.; Chen, Z.; et al. Semi-Supervised Multimodal Emotion Recognition with Expression MAE. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada, 29 October–3 November 2023. DOI: 10.1145/3581783.3612506.

[37] Ding, C.; Zong, D.; Li, B.; et al. Learning Aligned Audiovisual Representations for Multimodal Sentiment Analysis. In Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing, Sydney, Australia, 29 October 2023.

[38] Maaz, M.; Rasheed, H.; Khan, S. et al. Video-chatGPT: Towards detailed video understanding via large vision and language models. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11 August 2024.

[39] Luo, R.P.; Zhao, Z.W.; Yang, M.; et al. Valley: Video assistant with large language model enhanced ability. arXiv 2023, arXiv:2306.07207. DOI: 10.48550/arXiv.2306.07207.

[40] Li, B.; Zhang, Y.H.; Chen, L.Y.; et al. Mimic-it: Multi-modal in-context instruction tuning. arXiv 2023, arXiv:2306.05425. DOI: 10.48550/arXiv.2306.05425.

[41] Li, K.C.; Wang, Y.L.; HE, Y.N.; et al. Mvbench: A Comprehensive Multi-Modal Video Understanding Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024. DOI: 10.1109/CVPR.2024.0041.

[42] Ye, Q.H.; Xu, H.Y.; Xu, G.H.; et al. mplug-owl: Modularization empowers large language models with multimodality. arXiv 2023, arXiv:2304.14178. DOI: 10.48550/arXiv.2304.14178.

[43] Chen, G.M.; Chen, S.N.; Zhang, R.F.; et al. Allava: Harnessing gpt4v-synthesized data for lite vi-sion-language models. arXiv 2024, arXiv:2402.11684. DOI: 10.48550/arXiv.2402.11684.