Multi-modal Fusion for Emotion and Intention Joint Understanding

Zijie Lin A0307218J
Jiayi Li A0279788W
Boyang Li A0286148U
Jiale Li A0296896X

Abstract

This report presents a multimodal fusion approach for emotion and intention recognition in video-based interactions, leveraging visual, audio, and textual modalities. We developed a comprehensive model that integrates features using a Transformer-based encoder, while employing advanced pre-trained models such as Qwen2-VL, Qwen2-AL, CLIP, HuBERT, and Wav2Vec for feature extraction. Furthermore, we designed a joint learning module to enhance the interdependency between emotion and intention recognition tasks. In the ICASSP 2025 MEIJU Grand Challenge, our model shows significant improvements over baseline methods on the validation set. In the test set, we rank 11th in Track 2. These results demonstrate our model's robustness in handling imbalanced data scenarios and its effectiveness in multimodal understanding tasks.

1 Introduction

Emotion recognition in video is a crucial task in the field of human-computer interaction, enabling systems to interpret and respond to human emotions effectively. Its applications range from mental health support to advanced human-machine dialogue systems. As highlighted by Gladys and Vetriselvi [2023], video emotion understanding contributes to enhancing emotional intelligence in automated systems, which is essential for creating empathetic and human-like interactions. The significance of this task lies in bridging the gap between computational intelligence and human emotional needs.

Over the years, the domain of video emotion recognition has evolved significantly. Early studies primarily focused on visual modalities, such as facial expressions and physical gestures Ebrahimi Kahou et al. [2015]. Subsequent advancements incorporated audio-visual approaches to leverage the complementary information Noroozi et al. [2017]. More recently, research has extended to combining visual, acoustic, and lexical modalities, which Gladys and Vetriselvi [2023] and Zhang et al. [2024] emphasize as key to achieving more robust and nuanced emotion recognition.

Building on this foundation, the MEIJU dataset was introduced as part of the ICASSP 2025 Grand Challenge Liu et al. [2024]. The dataset provides a comprehensive resource for advancing multimodal emotion and intent recognition, featuring annotations for seven emotion categories and nine intent categories. It comprises textual, acoustic, and visual modalities from seven TV series in English and Mandarin. The Grand Challenge consists of two tracks: Track 1 focuses on low-resource recognition with balanced datasets, while Track 2 targets imbalanced data scenarios for real-world applications.

The competition ran from August 26, 2024, to November 9, 2024, attracting global participation. Our multimodal fusion model surpassed the baseline on the validation set and ranked 11th in Track 2, placing in the top 37% of the 29 participating teams, demonstrating its effectiveness.

Specifically, our approach features:

- Multimodal fusion of visual, audio, and text modalities for comprehensive understanding.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

- Enriching text features with scene-specific descriptions generated by Vision Language Models (VLMs) and Audio Language Models (ALMs).
- Employing advanced feature extraction tools like CLIP, ResNet50, HuBERT, and Wav2Vec, combined via a Transformer Encoder for robust cross-modal integration.

2 Related Work

Emotion Recognition from Image and Video: CNN architectures, especially ResNet, have shown strong performance in detecting facial expressions Sarvakar et al. [2023]. Recent work has leveraged video-based features, such as the Clip model, for emotion recognition Bondielli et al. [2021].

Emotion Recognition from Speech: Pre-trained models like Wav2Vec have achieved accurate results Chen and Rudnicky [2023], and transformer models like HuBERT have shown potential in capturing speech emotions Wagner et al. [2023].

Emotion Recognition from Text: Large language models (LLMs) demonstrate strong emotion detection capabilities due to their extensive knowledge and reasoning abilities Xue et al. [2024].

Multi-modal Language model: VLMs and ALMs can reason about emotion and intention from visual or audio features. Recent work has used VLMs with fine-tuning techniques to achieve state-of-the-art results in the 2024 MER challenge Cheng et al. [2024a,b].

Joint Task Learning: Joint learning of emotion and intention may improve video understanding. Prior work has shown the effectiveness of multi-task learning [Ge et al., 2024], and our experiments on the MEIJU 2025 dataset further support the benefits of joint learning [Liu et al., 2024].

3 Method

3.1 Multi-modal Emotion & Intention Classifier

The multi-modal emotion and intention classifier integrates visual, audio, and text modalities to predict emotions and intentions in video inputs. Figure.1 provides an overview of the model architecture.

The pipeline starts by processing input videos through different modality-specific modules to extract relevant features. These modalities include visual, audio, and text, each providing unique information that contributes to understanding emotions and intentions. Then,the extracted features from each modality are then fused using an integration mechanism that captures inter-modal relationships and generates a unified representation. This integrated representation is used for making predictions about emotions and intentions, allowing the model to effectively interpret complex social interactions.

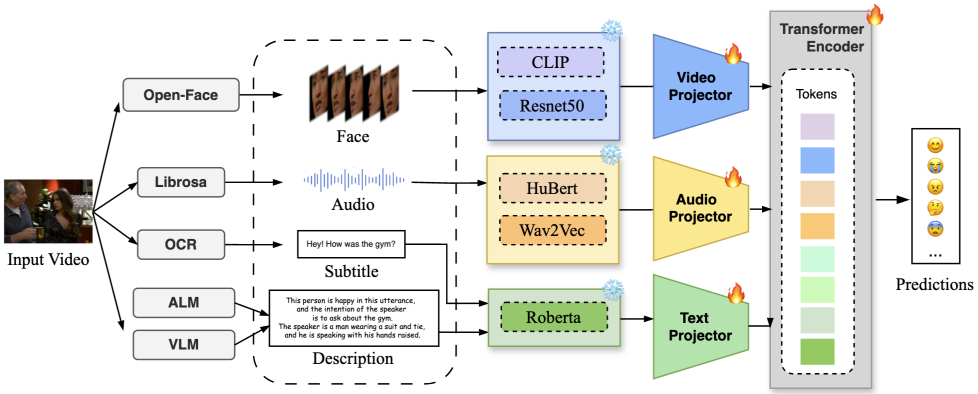

Figure 1: High-level overview of the multi-modal emotion & intention classifier pipeline. {Fire} represents modules with trainable parameters, and {Snowflake} represents modules with frozen parameters.

3.1.1 Vision Modality

The vision modality involves extracting and processing visual features from video frames to understand emotions and intentions.

Pre-process:

OpenFace :The process begins with OpenFace, a facial landmark detection and recognition library. OpenFace is used to extract rich facial features, including head pose, action units, and eye gaze. These features are highly relevant for recognizing emotions and intentions, as they provide essential information about facial expressions, body language, and attention focus. For example, head pose indicates the orientation of a person's face, which can suggest focus or attention, while action units represent muscle movements that help in identifying specific emotions. Eye gaze is also critical for understanding engagement and intention, as it reflects where a person is directing their attention.

Feature Extraction:

1.ResNet-50 :We use ResNet-50, a pre-trained deep convolutional neural network, to capture frame-level visual features Li and Lima [2021]. ResNet-50 is effective for this task because of its ability to learn complex visual patterns, which makes it well-suited for recognizing facial expressions and subtle visual cues in each video frame. These detailed frame-level features are crucial for understanding nuanced emotions, such as detecting slight changes in facial muscle movements that convey different emotional states.

2. CLIP :In addition to ResNet-50, we utilize the CLIP image encoder, which is trained on large-scale image-text data Radford et al. [2021]. CLIP is capable of learning high-level visual representations that complement the low-level features captured by ResNet-50. Unlike traditional vision models, CLIP leverages both visual and textual information, enabling it to understand the context of visual elements in a richer way. This multi-modal approach helps the model learn abstract features that go beyond individual facial expressions, thus enhancing the overall visual understanding.

Once features are extracted from each frame, we use an LSTM layer to encode the sequence of frames. The LSTM network is employed to capture temporal dependencies and progression over time. By encoding the sequential frames, the LSTM compresses temporal information into a series of tokens that represent how visual features evolve throughout the video. This allows the model to capture not only static visual cues but also dynamic changes, which are essential for understanding how emotions and intentions unfold over time.

3.1.2 Audio Modality

The audio modality aims to extract and process acoustic features from video data to help understand emotions and intentions through vocal cues. The process begins with audio pre-processing, followed by feature extraction from the processed signals.

Pre-process:

Librosa : We use Librosa, an audio processing library, to handle the initial audio pre-processing. Librosa is capable of extracting audio from video files, performing tasks such as downsampling, noise reduction, and normalization. These pre-processing steps are essential for ensuring that the audio input is clean and consistent, allowing the downstream models to extract meaningful features without being affected by noise or inconsistencies. For example, downsampling reduces the sample rate to focus on the frequency range most relevant for speech analysis, while normalization ensures that the amplitude of audio signals remains consistent across different samples.

Feature Extraction:

1. Wav2Vec : We utilize Wav2Vec, a self-supervised learning model, to effectively capture speech representations Schneider et al. [2019] . Wav2Vec is pre-trained on large-scale unlabeled speech data, enabling it to learn a broad set of acoustic features without needing labeled data. Its ability to capture contextual speech information makes it particularly useful for emotion recognition, as it can understand both the content of speech and subtle cues like intonation, tone, and rhythm. Wav2Vec has demonstrated strong performance in various downstream speech tasks, including

emotion recognition.This model provides a detailed representation of the audio, which is crucial for identifying emotional nuances in speech.

2. HuBERT :In addition to Wav2Vec, we use HuBERT (Hidden-Unit BERT), a transformer-based model that builds upon masked language modeling techniques for audio Hsu et al. [2021]. HuBERT provides more robust speech representations by predicting hidden units in a self-supervised manner, which makes it well-suited for capturing both prosodic and linguistic features. This model has shown state-of-the-art performance in speech understanding tasks, particularly in scenarios involving emotions, as it can handle complex audio dynamics like intonation and stress patterns.These features are key to understanding the emotional state of the speaker.

Once the features are extracted, we employ an LSTM layer to encode the temporal progression of the audio. The LSTM network helps capture the temporal dependencies present in the audio signal, such as changes in pitch or tone over time. By encoding these features into a sequence of tokens, the model can effectively represent the dynamics of speech, which are critical for distinguishing between different emotional and intentional states. For example, a gradual increase in pitch may indicate excitement or urgency, while a monotone voice might indicate disinterest or calmness.

3.1.3 Text Modality

Our text modality features are derived from two primary sources: subtitles extracted through OCR and descriptions generated by multi-modal language models.

OCR Optical Character Recognition technology enables the extraction of textual information from video frames, allowing us to capture any displayed text or subtitles that may provide additional context for emotion and intention recognition.

Multi-modal Language Models We leverage large multi-modal language models for their strong reasoning capabilities and contextual understanding. Specifically, we employ Qwen-2-VL Wang et al. [2024] for visual-language understanding and Qwen-2-AL Chu et al. [2024] for audio-language understanding. These models are prompted with specific queries to extract relevant information:

```
Prompts for Qwen-2-VL:
"Who is the speaker, and describe the speaker's facial expression and body movement."

Prompts for Qwen-2-AL:
1. "Describe the audio"
2. "Based on the utterance and voice, what is the intention of the speaker?"

Generated Examples:
Visual description: The speaker is a woman with long brown hair, wearing a black turtleneck.
She is standing in front of a man who is facing away from the camera. The woman is speaking
to the man, and her facial expression appears to be one of concern or worry. She is
slightly leaning forward, indicating that she is engaged in the conversation.

Audio transcription: The audio is of a woman speaking, in English, saying, "Would you
promise you won't tell anyone?"

Audio tone and purpose: Based on the voice, it sounds like this person is fearful or
anxious, possibly about keeping a secret or about someone finding out something.
```

Text Feature Extraction We utilize RoBERTa Liu [2019] to process both the OCR-extracted subtitles and language model outputs. For each text source:

- Extract the CLS token representation for sentence-level features
- Process word embeddings through TextCNN Zhang and Wallace [2015] for local feature extraction
- Project features through fully connected layers for dimension alignment

This comprehensive text processing pipeline effectively captures both explicit information from transcribed speech and implicit contextual information from visual and audio descriptions, providing rich features for emotion and intention recognition.

3.2 Multi-modal Fusion module

The multi-modal fusion module integrates features extracted from visual, audio, and text modalities to generate a unified representation for emotion and intention recognition. We employ a two-layer Transformer encoder that effectively handles complex inter-modal interactions using its self-attention mechanism. This allows the model to weigh the importance of features from different modalities, such as the connection between visual cues and audio signals that convey the same emotional state.

The first layer of the Transformer encoder captures basic relationships within and between modalities, processing features like facial expressions, speech tones, and textual content independently while identifying interactions. The second layer further refines these relationships, focusing on high-level interdependencies that contribute to an integrated understanding of emotional and intentional context. By stacking two layers, the model captures both low-level and high-level interactions for a cohesive representation.

The self-attention mechanism provides several advantages for multi-modal fusion, including the ability to model inter-modal relationships by focusing on relevant features across different modalities, dynamic weighting of features based on context, which is crucial for interpreting complex scenarios, and efficient integration of long-range dependencies to capture subtle cues, such as delayed emotional responses. These advantages enable the two-layer Transformer encoder to effectively integrate features from visual, audio, and text modalities, capturing the richness of human emotions and intentions, ultimately leading to more accurate and empathetic human-computer interaction.

3.3 Emotion-intention Joint Learning

The concept of emotion-intention joint learning was initially proposed in this dataset challenge Liu et al. [2024]. We build on this approach with an improved module, called the Emotion-Intention Joint Learning Module, which moves beyond basic cross-attention.

Our method begins with pre-training two separate classifiers—one for emotion and one for intention classification. After pre-training, we freeze the parameters of these classifiers and connect them to the joint learning module, as shown in Figure.2. The experiment section will discuss the impact of this module on performance in detail.

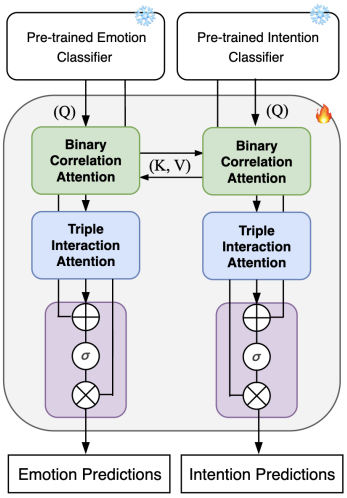

Figure 2: Overview of the emotion-intention joint learning module. {Fire} represents modules with trainable parameters, and {Snowflake} represents modules with frozen parameters.

4 Experiments

4.1 Experimental Setup

Due to strict computation resources, we only use a 3090 from AutoDL platform by sself-funding. We develop our model on foundation of MEIJU2025 baseline codebase.

4.2 Metrics

To evaluate emotion and intent recognition, we use metrics tailored for fair and effective assessment across tasks. The selection of these metrics aligns with the objectives of this task, as emotion and intent recognition often deal with imbalanced datasets and multiple labels. The weighted F1 and micro F1 scores provide robust assessments under these conditions, while JRBM emphasizes the joint recognition performance, reflecting the interdependency of emotion and intent in real-world applications.

4.2.1 F1 Score

The F1 score, balancing precision and recall, is defined as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The F1 score is well-suited for imbalanced data, ensuring effective prediction for both majority and minority classes.

4.2.2 Micro F1 Score

The micro F1 aggregates true positives, false positives, and false negatives across all classes:

$$\text{Micro F1} = \frac{2 \cdot \text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}.$$

It evaluates overall performance, making it ideal for imbalanced datasets.

4.2.3 Joint Recognition Balance Metric (JRBM)

To measure joint emotion and intent recognition, we use the Joint Recognition Balance Metric:

$$\text{JRBM} = \frac{2 \cdot M_{\text{emotion}} \cdot M_{\text{intent}}}{M_{\text{emotion}} + M_{\text{intent}}}$$

Here, M_{emotion} and M_{intent} are the chosen metrics (Weighted F1 or Micro F1). This ensures balanced evaluation of both tasks.

4.3 Results and Analysis

Table 1 shows the performance of our model compared to the baseline. Our model achieved significant improvements across all metrics, particularly in Intent Micro F1 and Joint F1 scores.

Table 1: Comparison with Baseline Model

Model	Emo Micro F1	Int Micro F1	JRBM	Parameters(M)
Baseline	0.5342	0.5412	0.5377	3.082
Our Model	**0.5574**	**0.6026**	**0.5791**	4.128

Our team ranked 11th out of 29 participating teams, as shown in Figure 3. The competition was highly competitive, showcasing the robustness of our approach.

Results

#	User	Entries	Date of Last Entry	score ▲	Detailed Results
1	yihongzhu	7	11/08/24	0.5882 (1)	View
2	xuyangyan	18	11/08/24	0.5882 (1)	View
3	xuezhiyuan	1	11/08/24	0.5877 (2)	View
4	Aizen_Sousuke	35	11/08/24	0.5847 (3)	View
5	Soceremite	4	11/07/24	0.5795 (4)	View
6	SEU_AIPLab	8	11/08/24	0.5792 (5)	View
7	Haorr	5	11/05/24	0.5780 (6)	View
8	fengkuiQian	44	11/06/24	0.5774 (7)	View
9	Djctionary	21	11/08/24	0.5772 (8)	View
10	Zq5437	10	11/05/24	0.5768 (9)	View
11	zj-lin	3	11/08/24	0.5736 (10)	View
12	Zhouyang_Chi	5	11/02/24	0.5721 (11)	View
13	zzzfly	19	11/08/24	0.5703 (12)	View
14	XiaolinXu	4	11/05/24	0.5703 (12)	View
15	murakami	3	11/03/24	0.5703 (12)	View
16	honghongWang	8	10/25/24	0.5625 (13)	View
17	yangbin	6	11/09/24	0.5582 (14)	View
18	chexinyi	1	11/01/24	0.5530 (15)	View
19	MSXF_Audio	1	11/07/24	0.5332 (16)	View
20	KTTRCDL	3	11/05/24	0.5298 (17)	View
21	liyihao	5	11/05/24	0.5280 (18)	View
22	Jasonerrr	13	11/04/24	0.5180 (19)	View
23	GMLAB-SZU	1	10/31/24	0.4713 (20)	View
24	JialongMai	3	10/30/24	0.4628 (21)	View
25	Jasonxiei	12	11/03/24	0.4404 (22)	View
26	Xiao_MG	28	11/09/24	0.0887 (23)	View
27	DxBOW	18	11/08/24	0.0887 (23)	View
28	lyy2002	19	11/09/24	0.0785 (24)	View
29	msxf	9	11/05/24	0.0750 (25)	View

Figure 3: Competition Results

Table 2: Ablation Study on Modalities

Vision	Audio	Text	Emo Micro F1	Int Micro F1	JRBM	Parameters (M)
✓	X	X	0.5331	0.4101	0.4636	1.416
X	✓	X	0.4913	0.4560	0.4730	1.416
X	X	✓	0.4841	0.5534	0.5165	1.618
✓	✓	X	0.5438	0.5206	0.5319	2.896
✓	X	✓	0.5002	0.5954	0.5437	3.098
X	✓	✓	0.4658	0.5963	0.5230	3.098
✓	✓	✓	**0.5574**	**0.6026**	**0.5791**	4.128

The ablation study (Table 2) evaluates the contribution of each modality—vision, audio, and text—towards the overall performance.

The full tri-modal combination significantly improves performance across all metrics, achieving the highest Joint F1 (0.5791). This confirms that integrating all three modalities provides the most comprehensive understanding of both emotions and intents, leveraging their complementary features.

Table 3 demonstrate the significant impact of the joint learning module on performance. Without joint learning, individual modalities or their pairwise combinations fail to achieve the optimal performance observed with full tri-modal integration. Specifically:

- The tri-modal configuration with the joint learning module achieves **the highest JRBM** score of 0.5791, outperforming all pairwise combinations and single-modality setups.
- Both Emotion Micro F1 and Intention Micro F1 metrics improve with joint learning, showing effective utilization of cross-modal interactions and task interdependencies.

These results confirm that the joint learning module is essential for achieving superior performance by effectively integrating and balancing contributions from all modalities.

Table 3: Effect of Joint Learning Module

Joint Module	Emo Micro F1	Int Micro F1	JRBM	Parameters (M)
Without	0.5572	0.5905	0.5733	3.872
With	**0.5574**	**0.6026**	**0.5791**	4.128

5 Conclusion

In this study, we developed a multimodal fusion model for joint emotion and intention recognition, leveraging visual, audio, and text modalities. Our approach demonstrated significant improvements over the baseline across all evaluation metrics, including Emotion Micro F1, Intent Micro F1, and Joint scores (JRBM). Through rigorous experimentation, we highlighted the critical role of multimodal integration and the joint learning module in achieving comprehensive and robust recognition performance.

The ablation study underscored the complementary strengths of individual modalities and revealed the superior effectiveness of combining all three modalities. Furthermore, the joint learning module proved to be indispensable, optimizing cross-modal interactions and enhancing task interdependencies. These results validate the importance of incorporating multimodal data and joint learning strategies in advancing emotion and intention recognition systems.

Future work could explore the integration of additional modalities, such as physiological signals, and refine the joint learning module to further enhance performance. Additionally, expanding the dataset diversity and adopting more advanced pre-trained models could unlock new possibilities for emotion and intention recognition tasks.

In summary, this research contributes to the growing field of human-computer interaction by providing a comprehensive, multimodal approach to emotion and intention understanding, setting a strong foundation for future advancements in the domain.

References

A. Bondielli, L. C. Passaro, et al. Leveraging clip for image emotion recognition. In *Ceur workshop proceedings*, volume 3015. CEUR-WS, 2021.

L.-W. Chen and A. Rudnicky. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *arXiv preprint arXiv:2406.11161*, 2024a.

Z. Cheng, S. Tu, D. Huang, M. Li, X. Peng, Z.-Q. Cheng, and A. G. Hauptmann. Sztu-cmu at mer2024: Improving emotion-llama with conv-attention for multimodal emotion recognition. *arXiv preprint arXiv:2408.10500*, 2024b.

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 467–474, 2015.

M. Ge, M. Li, D. Tang, P. Li, K. Liu, S. Deng, S. Pu, L. Liu, Y. Song, and T. Zhang. Early joint learning of emotion information makes multimodal model understand you better, 2024. URL `https://doi.org/10.1145/3689092.3689415`.

A. A. Gladys and V. Vetriselvi. Survey on multimodal approaches to emotion recognition. *Neurocomputing*, page 126693, 2023.

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 transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

B. Li and D. Lima. Facial expression recognition via resnet-50. *International Journal of Cognitive Computing in Engineering*, 2:57–64, 2021.

R. Liu, H. Zuo, Z. Lian, X. Xing, B. W. Schuller, and H. Li. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*, 2024.

Y. Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, 10(1):60–75, 2017.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. S. Kumar, R. Manjunath, and S. Jaiswal. Facial emotion recognition using convolutional neural networks. *Materials Today: Proceedings*, 80: 3560–3564, 2023.

S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759, 2023.

P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

J. Xue, M.-P. Nguyen, B. Matheny, and L.-M. Nguyen. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pages 277–292. Springer, 2024.

S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, 237:121692, 2024.

Y. Zhang and B. Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.