Deep Multimodal Learning for Emotion Recognition

Linoy Halifa[1], Bezalel Itzhaky[1], Ezra Ella[1]
[1]Afeka academic collage

Abstract

Accurate emotion recognition is a foundational component of effective human interaction and a critical enabler of social cohesion and conflict mitigation. Misinterpretation of emotional cues can lead to significant interpersonal misunderstandings, with implications for both individual well-being and societal stability. While most individuals can reliably interpret universal emotions such as happiness, anger, sadness, and fear, the ability to recognize more nuanced affective states—particularly those shaped by context, culture, and interpersonal dynamics—is less uniformly distributed and often associated with higher emotional sensitivity or intelligence. Notably, although core emotional expressions are largely consistent across cultures and even species, many sub-emotions are shaped by cultural and temporal contexts.

In the emerging era of emotionally aware artificial intelligence, the capacity of computational systems to perceive and interpret human emotions is becoming increasingly important. Applications span a wide range of domains, including human-computer interaction, education, digital mental health, security, and social media platforms. However, unimodal approaches have consistently fallen short in capturing the richness and variability of real-world emotional expression, which is inherently multimodal. Even state-of-the-art Multimodal Large Language Models (MLLMs) continue to struggle with the integration of audio inputs and the recognition of subtle micro-expressions.

To address these limitations, prior work introduced the MERR dataset, comprising over 28,000 coarse-grained and 4,400 fine-grained annotated emotional instances, enabling improved generalization across a variety of contexts. Building on this, the Emotion-LLaMA architecture was proposed to fuse audio, visual, and textual modalities using emotion-specific encoders. This model leverages a shared latent space and a fine-tuned LLaMA backbone, yielding significant improvements in both emotion recognition and affective reasoning.

In the present study, we explore the performance gains achieved by replacing the original LLaMA core in the Emotion-LLaMA architecture with a newer-generation LLaMA model. Furthermore, given that most prior work focused on Chinese participants, we investigate the model's cross-cultural robustness by evaluating its performance on emotionally annotated scenarios involving individuals of Middle Eastern descent. This evaluation provides insight into the model’s capacity for generalization across ethnocultural boundaries and its potential limitations in interpreting culturally situated emotional expressions Our analysis revealed that Emotion-LLaMA, as trained on the Chinese-dominated MERR dataset, exhibits limitations in cross-cultural generalization, particularly when applied to individuals from the MENA region. This suggests the need for incorporating cultural, contextual, and demographic awareness into future training pipelines. Additionally, we observed that minority outputs—often dismissed during prediction—carry valuable emotional and situational cues. To address this, we proposed a voting-based reasoning head that

aggregates multiple prompt-driven outputs. This architecture significantly improved both emotion recognition accuracy and contextual understanding by preserving and integrating diverse interpretations of the same scene.

1. Introduction

Emotion recognition has played a vital role in the evolutionary trajectory of social species, contributing to survival, communication, and the development of complex societies. Foundational emotions such as fear, happiness, anger, and sadness are known to be shared across cultures and even species, suggesting a deep evolutionary origin (Darwin, 1872; Ekman, 1992; de Waal, 2008).

In the computational domain, automatic emotion recognition remains a substantial challenge for two primary reasons. First, while current systems perform relatively well on well-defined or exaggerated emotional displays, they often struggle to interpret complex or ambiguous emotional cues (Poria et al., 2017). Second, emotional meaning is highly context-dependent—requiring a nuanced understanding of the interplay between facial expression, vocal tone, and semantic content (Kossaifi et al., 2019).

The emergence of transformer-based architectures and large language models (LLMs) has enabled significant progress in modeling multimodal data. These systems are capable of jointly processing audio, visual, and textual inputs, thereby capturing contextual dependencies that are essential for accurate emotion understanding. This multimodal integration mitigates information loss by aligning various input modalities into a unified representation space, offering a promising pathway for advancing affective computing (Tsai et al., 2019; Zadeh et al., 2018).

2. Related Work

This study builds upon the Emotion-LLaMA model introduced by Cheng et al. (2024), which aimed to improve multimodal emotion reasoning through the integration of audio, visual, and textual data (Cheng et al., 2024). Their work addressed a critical gap in existing multimodal large language models (MLLMs), which have shown considerable success in general vision-language tasks (e.g., VQA, video understanding) (Alayrac et al., 2022; Wang et al., 2023), but still fall short in emotional reasoning—particularly due to their limited capacity to process audio signals and detect subtle micro-expressions.

A central insight in Cheng et al.'s work is that the absence of a high-quality, instruction-tuned dataset for emotion recognition is a key limiting factor for MLLMs in affective domains. To address this, the authors introduced the MERR dataset, which includes 28,618 coarse-grained and 4,487 fine-grained annotated emotional samples. This dataset captures a wide range of emotion categories—including nuanced ones such as “doubt” and “contempt”—and was specifically designed to support instruction tuning and improve real-world generalization.

To fully leverage the dataset, the Emotion-LLaMA model was proposed, integrating HuBERT for audio encoding and multiple vision encoders (MAE, VideoMAE, EVA) to capture facial expression dynamics and scene context. These features were aligned within a LLaMA-based language model via instruction tuning, significantly enhancing emotional reasoning performance. Extensive evaluations demonstrated state-of-the-art results on multiple benchmarks, including EMER, MER2023-SEMI1, MER2024-NOISE2, and DFEW, with Emotion-LLaMA even outperforming ChatGPT-4V in zero-shot settings.

Key contributions of Cheng et al. include:
• Construction of the MERR dataset, offering a richly annotated and diverse corpus for multimodal emotion recognition.
• Development of a multimodal architecture that integrates audio, visual, and textual modalities with a LLaMA-2-based language model.
• Empirical validation across several benchmark datasets, establishing Emotion-LLaMA as a state-of-the-art model in public evaluations.

While most of the model architecture relies on LLaMA-2 as the core language model, the authors also utilized LLaMA-3 within their data annotation pipeline—specifically for generating refined emotional descriptions. This distinction is important: the fine-tuned emotion reasoning component is based on LLaMA-2, while LLaMA-3 was used upstream in the data generation process.

In our current work, we first sought to reproduce and extend the results of Cheng et al. Specifically, we considered to explore whether replacing the LLaMA-2 core with LLaMA-3 within the Emotion-LLaMA architecture could yield additional performance gains. This modification reflects the natural evolution of the model in parallel with ongoing improvements in foundation language models. However, during our experience with the mode, we discovered that there is a huge space where improvements could be made using our suggested voting head module and without replacing the LLaMA-2 core.

Additionally, we noted that the MERR dataset primarily features Chinese participants, raising important questions about the model's cross-cultural robustness. To evaluate this, we tested the model's generalizability on emotional scenarios involving individuals of Middle Eastern and North African (MENA) descent.

Lastly, based on our analysis of model outputs, we identified potential improvements through ensemble voting mechanisms, which we describe and implement in the architectural extensions presented in this paper.

3. Methodology

We investigated the Multimodal Emotion Recognition (MER) system originally designed by Cheng et al. (2024), incorporating video, audio, and text modalities. Our goal was to identify concise yet meaningful improvements, specifically by experimenting with the enhancement of the LLaMA-2 backbone with a voting head module, in order to assess whether this upgrade enhances classification accuracy and robustness.

We also probed the model's ability to generalize across cultures by evaluating performance using video clips of individuals from non-Chinese backgrounds, particularly those from the Middle Eastern and North African (MENA) region—addressing potential overfitting to the primarily Chinese demographic of the original MERR dataset. Based on anomaly patterns in output predictions, we further proposed an ensemble voting mechanism to boost confidence and mitigate misclassifications in ambiguous cases.

3.1 MER2025 Dataset

We utilized the MER2025 dataset, accessible via Hugging Face at the link: (https://huggingface.co/datasets/MERChallenge/MER2025) The official challenge is described in the baseline MER2025 paper (Lian et al. 2025)

The dataset is structured into four tracks, of which we concentrated on Track 1 (MER-SEMI)—basic categorical emotion recognition. According to the challenge specifications:

- Train+Validation set: 7,369 labeled samples
- Test set: 20,000 labeled samples selected from a larger 124,802 unlabeled pool (Lian et al. 2025), label is hidden for challenge participants.

Each sample includes aligned video clip, audio file, and a text transcript in CSV format—fully synchronized across modalities—providing a robust basis for analyzing the impact of architectural changes like upgrading the LLM backbone.

Our choice of Track 1 enables a controlled evaluation of the LLaMA core replacement (from 2 to 3), while also permitting scrutiny of cross-cultural performance given the dataset's standardized emotion categories.

3.2 Model Architecture

Similar to Cheng et all work, the architecture we used consists of three main components:

- Vision Encoder - Responsible for extracting visual features from video frames (based on ViT/VideoMAE).
- Audio Encoder - Extracts acoustic and speech-related features (based on HuBERT).
- Text Encoder – The RoBERTa encoder processes the textual input to generate contextual embeddings, which are then integrated with audio and visual features for multimodal reasoning.

The outputs from the modality-specific encoders are integrated via a fusion layer that projects all features into a shared embedding space. Subsequently, a cross-attention mechanism is employed to capture inter-modal correlations, enabling the model to produce either a final emotion classification or a descriptive reasoning output.

To enhance model performance, we introduce an additional layer that generates multiple output hypotheses based on varied prompt formulations. This layer facilitates a voting-based selection of the most reliable emotional interpretation, while also estimating a confidence score for each predicted outcome as described in section 3.4.

3.3 Training Procedure

- Stage 1 – unimodal baseline training - Each modality (video, audio, text) was first trained independently on Track 1 samples to establish baseline performance

- Stage 2 – Multimodal Fine- Tuning
  We then performed multimodal fine-tuning, combining all three modalities (video, audio, text) to train a unified model, using eight combinations: (AVT, AT, AV, VT) each with topn1 or topn2.
  This stage improves the model's ability to reason across modalities and detect subtle emotional cues.
- Comparison With Baseline - Our goal is to match or surpass the performance reported in the baseline paper by leveraging the voting head model ability to fuse multiple inputs and use data from the minority report.

3.4 Prompt-Ensemble Reasoning via MiniGPT

To enhance robustness, generalizability, and interpretability in multimodal emotion inference, we introduce a Prompt-Ensemble Reasoning mechanism at the final stage of the Emotion-LLaMA pipeline. Rather than selecting a single prediction or applying simple majority voting across outputs, we leverage the generative reasoning capabilities of a compact vision-language model—MiniGPT (Zhu et al., 2023) to synthesize multiple prompt-based outputs into a unified emotional judgment.

In this setup, we apply several distinct prompt formulations to the same multimodal input (video, audio, transcript), each generating a separate descriptive or categorical output via the Emotion-LLaMA core. These outputs are concatenated and passed as a structured prompt to MiniGPT, which performs a high-level reasoning task: identifying the most likely emotion, inferring a plausible situational context, and providing a self-assessed confidence level for its final prediction. We refer to the final emotion/situation prediction generated via prompt ensemble reasoning as a T^{vot} (Token Vote), reflecting a majority-informed token-based decision derived from multiple T^{ans} outputs (figure 1).

Example input

The following are descriptions of the same scene obtained from different prompt perspectives:

1. "The person looks uncertain and hesitant."

2. "They seem confused, possibly not understanding the conversation."

3. "There is a lack of eye contact and signs of mental disengagement."

Based on this information:

- What is the most likely emotional state of the individual?

- What situation could explain this behavior?

- On a scale from 1 to 10, how confident are you in this interpretation?

MiniGPT has shown strong capabilities in open-ended reasoning tasks and contextual understanding, making it particularly well-suited for cases where subtle emotional cues and conflicting interpretations arise (Zhu et al., 2023; Dai et al., 2023). This approach aligns with broader work in self-consistency prompting and multi-step inference in LLMs (Wang et al., 2022), where generating and aggregating multiple outputs improves reliability.

Key Advantages:

- **Zero-shot generalization**: No task-specific fine-tuning is needed.
- **Interpretability**: The output includes both emotion classification and explanatory reasoning, along with a confidence estimate.
- **Multimodal extension**: MiniGPT can optionally incorporate raw visual frames or scene descriptions for richer contextual grounding.

This reasoning head effectively acts as a final **aggregation and validation layer**, improving resilience to prompt variation and enhancing model transparency—both essential in real-world emotion recognition tasks where ambiguity and cultural variability are common (Zadeh et al., 2018; Lian et al., 2025).

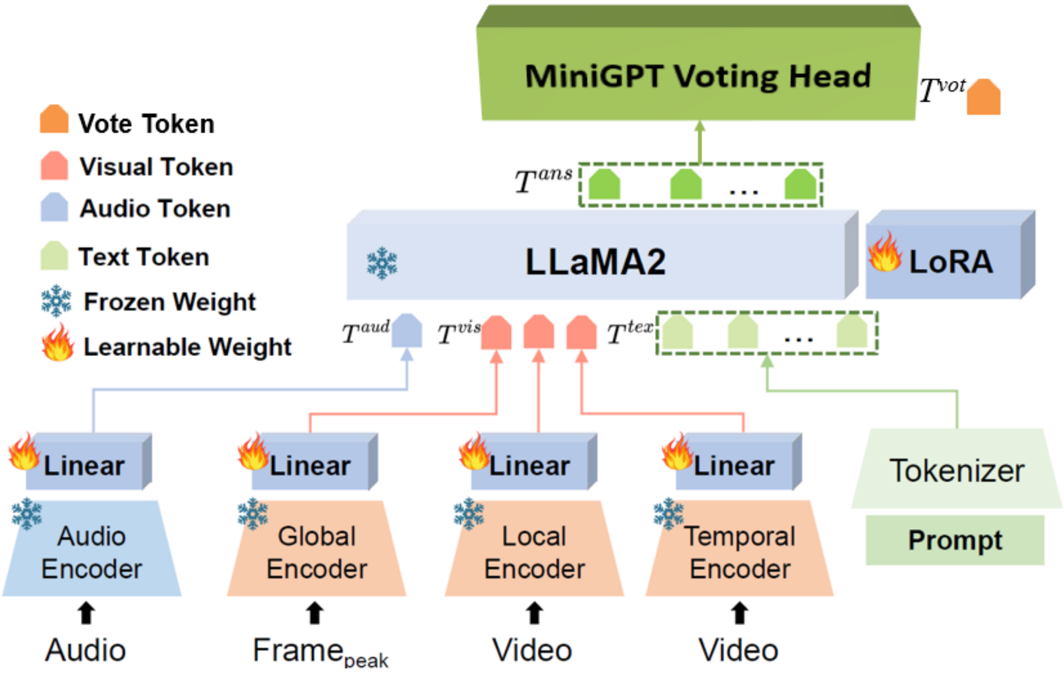


Figure 1: *Architecture of the improved Emotion-LLaMA model.*
The base architecture integrates audio, visual, and textual inputs using modality-specific encoders for multimodal emotion recognition and reasoning. In our extension, we introduce a MiniGPT-based voting head that aggregates multiple prompt-based outputs from the Emotion-LLaMA core and generates a unified final prediction—referred to as a T^{vot} (Token Vote) through contextual reasoning and confidence estimation.

3.5 Weighted Late-Fusion Post-Analysis of Multimodal LLaMA Outputs

In our post-hoc evaluation of the Emotion Recognition Multimodal LLaMA outputs, the raw unimodal prediction results, exported in CSV format, were subjected to a weighted late-fusion analysis. Each of the six unimodal classifiers—two per modality (audio, video, and text)—was assigned a modality-specific weight proportional to its confidence score, as produced by the respective model. The weighting scheme serves as a calibration mechanism, ensuring that modalities with inherently higher discriminative power exert greater influence on the final decision boundary. The weighted outputs were then aggregated via a summation-based voting framework, wherein the class with the highest cumulative weighted score was selected as the final predicted emotion. This method provides a principled integration of heterogeneous input streams, capturing the asymmetric contribution of each modality and enhancing both the robustness and interpretability of the multimodal inference process. Post-hoc evaluation Excel files are available in the project's public GitHub repository.

3.6 Code and Files Availability

All source code, experimental configurations, and processed datasets used in this study are openly accessible in the project's public GitHub repository: https://github.com/LinoyHalifa/Multimodal-Learning. The repository includes scripts for data preprocessing, model inference, and post-hoc analysis, enabling full reproducibility of the results reported herein.

4 Results

4.1 Emotion-LlaMA Zero-Shot Evaluation on MENA Individuals Revealed Accuracy Limitations

A notable limitation of the MERR dataset is that it consists exclusively of video and audio samples from Chinese participants. This raised an important question regarding the generalizability of the model—specifically, its ability to accurately detect facial expressions and interpret vocal tones in individuals from other ethnic backgrounds, such as those from the MENA (Middle East and North Africa) region, even when evaluating basic emotions in Track 1.

In our analysis, the model demonstrated some ability to recognize facial and vocal emotional cues in MENA individuals. However, it also exhibited systematic **misclassifications**, particularly in scenarios involving non-Asian facial morphology and dialect-influenced vocal prosody.

For example:

- In one case, the model classified a combination of **frustration** and a slight **smile** as "angry," despite the mixed effect being more complex and ambiguous (Figure 2a).
- In another case, the same "angry" label was assigned based on vocal tone, although the acoustic profile more closely matched **fear or anxiety** (Figure 2b).
- In a further instance, the model incorrectly described a **clean-shaven** Mediterranean individual as a "man with a beard and mustache," suggesting a bias or limitation in visual feature recognition (Figure 2c).

These examples underscore the importance of incorporating **culturally and ethnically diverse datasets** during pretraining and tuning phases to ensure reliable cross-demographic emotion recognition.

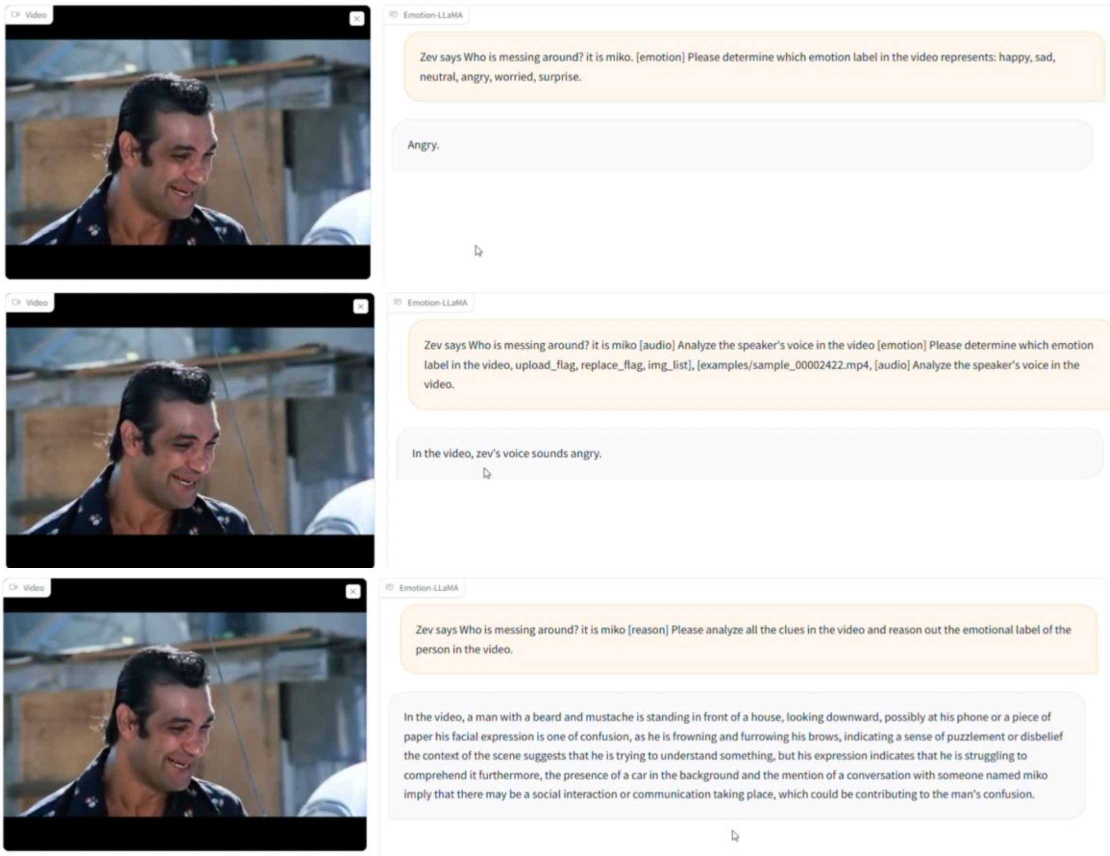


Figure 2: a. Misclassification of a frustrated but smiling expression as “angry.” B. Misinterpretation of a fearful vocal tone as “angry.” C. Visual recognition error: Clean-shaven MENA subject identified as having a beard and mustache.

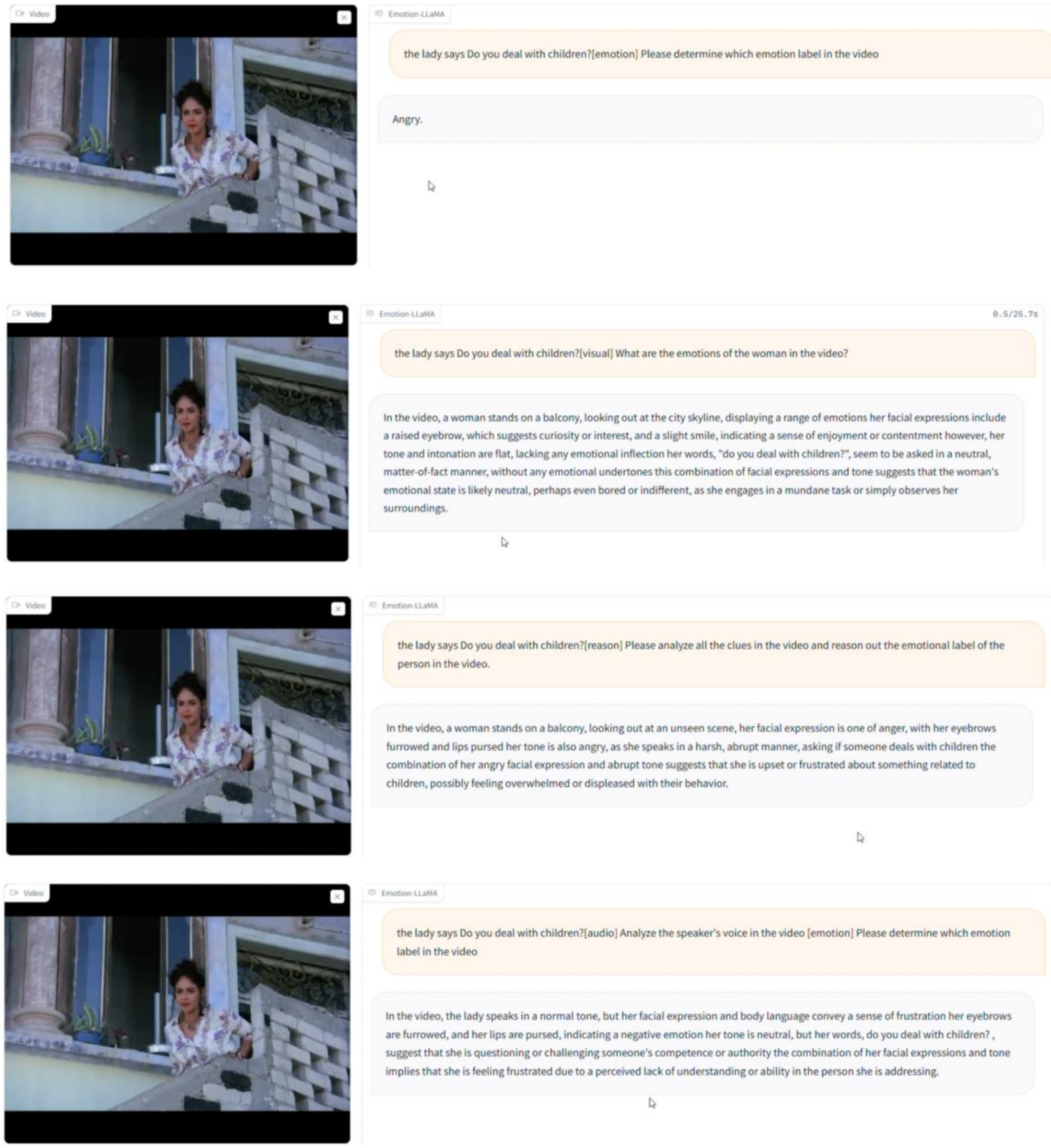
4.2 Using a MiniGPT-Based Voting Head Improves Model Accuracy and Reasoning

To evaluate our hypothesis that adding a reasoning and voting layer could improve Emotion-LLaMA's accuracy, we tested a prompt ensemble approach using MiniGPT as a reasoning head. We generated multiple outputs for the same video scene using different prompt formulations and provided these outputs to the MiniGPT model for aggregation and inference.

Initial results suggest that this approach can effectively prevent critical misclassifications. In the example shown in Figure 3, three out of four prompts correctly identified the woman’s emotional state as angry. However, one prompt (Figure 3b) generated clearly inaccurate interpretation—labeling the emotion as curiosity, enjoyment, and neutrality.

By aggregating the four interpretations, the MiniGPT-based head correctly predicted the dominant emotion as "Angry", with a high confidence score of 9/10. Additionally, the model produced a nuanced explanation of the scene, demonstrating improved understanding of situational context and emotional tone. The generated output (Figure 3e) not only affirmed the correct emotion but also included fine-grained reasoning that surpassed the descriptive clarity of any single prompt or the standalone LLaMA model (full text in supplementary Data 1).

This result illustrates the potential of combining multiple perspectives with a reasoning engine to enhance multimodal emotional inference and reduce false positives or misleading interpretations.



Tvot (Token Vote) Output:

- **Emotion**: Angry
- **Situation**: The woman appears to be addressing a stressful situation involving children, potentially in a caregiving or educational context. Her stern facial expression, sharp tone, and abrupt questioning suggest she is experiencing frustration or irritation—possibly due to challenging behavior or lack of support. The setting (balcony/staircase) implies she might be confronting someone from a physical or emotional distance, adding to the tension.
- **Confidence**: 9/10

Figure 3: a. Prompt-based output labeling the woman as “Angry” (true lable). b. Incorrect prompt-based output suggesting “Curious, Neutral, Enjoying”. 3c & 3d: Additional prompt outputs reinforcing the “Angry” label. Figure 3e: GPT demo aggregated output with final decision: "Angry", confidence: 9/10, and extended situational reasoning.

4.3 Audio Features Contributed Most Significantly to Model Accuracy

To evaluate the relative contribution of each modality to model performance, we conducted a controlled ablation analysis by inputting each modality—audio, text, and video—independently into the Emotion-LLaMA model. Performance was assessed using F1 score and accuracy metrics across a fixed evaluation set (*Supplementary* Data 2).

The results revealed that audio features consistently outperformed both text and video in terms of emotional classification accuracy. Notably, there was no significant performance gap between models using baseline audio features and those using enhanced (large) audio representations.

The finding that audio surpassed text is unsurprising, as intonation, stress, and prosody often convey emotional nuance beyond the semantic content of speech—emphasizing the notion that *“emotion lies more in how something is said than in what is said.”*

However, the performance gap between audio and video was less intuitive. One might expect facial expressions and gestures to provide equally rich emotional cues. This discrepancy and its possible implications are further addressed in the Discussion section.

modality	Audio features		Text features		Video features	
Feature extraction Model	chinese-hubert-base-UTT	chinese-hubert-large-UTT	chinese-roberta-wwm-ext-UTT	chinese-roberta-wwm-ext-large-UTT	clip-vit-base-patch32-UTT	clip-vit-large-patch14-UTT
Feature Dimension	(1, 768)	(1, 1024)	(1, 768)	(1, 1024)	(1, 512)	(1, 768)
f1_score	0.74	0.73	0.50	0.53	0.60	0.63
Acc. score	0.59	0.58	0.33	0.36	0.42	0.46

Table 1: F1 and accuracy scores for single-modality inputs evaluated using the Emotion-LLaMA model. For each modality (audio, video, and text), both the base and large variants of the model were tested.

4.3 The Combination of Multimodalities into the LLaMA Model Had a Relatively Small Effect on Scores

In the next stage of our evaluation, we examined the effect of combining multiple modalities—audio, video, and text—on the performance of the Emotion-LLaMA model (Table 2). We tested several input configurations across these modalities and measured their performance using F1 score and accuracy.

Surprisingly, the results showed that adding more modalities resulted in only modest performance gains, with all configurations remaining below 80% in both metrics. This finding suggests that the additive value of multimodal integration is limited in this setup.

Two possible explanations emerge:

1. There is a high degree of redundancy across the modalities, meaning that much of the emotional information is already present in any single modality.
2. The model may lack sufficient capacity or design to extract and align the complementary features across modalities needed to significantly improve predictions.

These observations raise important questions about how modality-specific features are processed and fused in the current model, and they motivate future architectural improvements.

Score/Model	multi model Fusion 1	Multi model Fusion 2	Multi model Fusion 3	Multi model Fusion 4	multi model Fusion 5	multi model Fusion 6	multi model Fusion 7	multi model Fusion 8
Fusion	AVT	AVT	AV	AV	AT	AT	VT	VT
topn	1	2	1	2	1	2	1	2
f1_score	79.27%	79.96%	73.39%	74.71%	79.25%	77.77%	69.78%	70.88%
Acc. score	65.66%	66.61%	57.97%	59.62%	65.63%	63.63%	53.59%	54.89%

Table 2: fusion combination: A = Audio input. V = Video input. T = Text input. AVT = Audio + Video + Text (all modalities). Fusion topn = Number of top-scoring predictions from each modality considered in fusion (e.g., Top-1 or Top-2 predictions retained during fusion)

4.4 In the Case of Six Different Models, Only 14% of Multimodal Predictions Agreed on the Same Label

In earlier analyses, we observed that model performance improved slightly when multiple modalities were fused, rather than relying on a single dominant modality. To further explore the potential for improvement using a voting mechanism, we analyzed the level of agreement between predictions made by six different multimodal configurations (as described in Table 2).

We evaluated 20,000 samples and measured how often these models converged on the same predicted label (Table 3). Agreement levels ranged from full consensus (all six models predicting the same label) to complete disagreement (each model outputting a unique label). A score of 1 indicates that only one model aligned with the majority label, while a score of 6 reflects full agreement across all models.

Results indicate a high level of divergence, with only 14% of cases showing complete agreement among all six models. Conversely, in 86% of samples, at least one model deviated from the majority prediction. These "minority reports" represent valuable alternative perspectives that are currently discarded—but could provide important complementary insights if incorporated into a voting-based head module. This highlights the potential of ensemble methods to capture a broader emotional context and improve final decision-making.

Number of models in agreed majority	Count of samples	Percentage of total samples
1/6	4	0.02%
2/6	2520	12.60%
3/6	5106	25.53%
4/6	6609	33.05%
5/6	2950	14.75%
6/6	2811	14.06%

Table 3: Agreement Level: Number of models (out of 6) that agreed with the majority label for a given sample. If all six models gave different outputs, the **majority label** was selected based on the highest confidence or internal model score, in this table there are four cases where each of the models provided a different result. Percentages reflect how frequently each level of agreement occurred across 20,000 tested samples.

4.5 Hyperparameter Adjustment and Unification Improved Performance

The original Emotion-LLaMA model uses randomly initialized hyperparameters for each run, potentially introducing performance variability between trials. To improve consistency and performance, we conducted a series of trial-and-error experiments to optimize and stabilize the initial hyperparameter configurations (see *Supplementary Data 3*). Once the optimal hyperparameters were identified, we applied these fixed settings uniformly across all model evaluations.

This unification led to two key improvements. First, the overall F1 scores and accuracy rates increased across tasks and configurations (Table 2, *Supplementary* Data 5), providing results similar to those obtained by Chent et al. (*Supplementary* Data 4). Second, and perhaps more importantly, model agreement levels improved—that is, different model variants (based on modality combinations and fusion topn settings) more frequently produced consistent outputs (*Table 4*). This consistency suggests that stabilized hyperparameters reduce randomness in internal representations and enhance alignment between multimodal reasoning paths.

Number of models in agreed majority	Incidence in all emotions	Percentage of total samples
1/8	0	0%
2/8	22	0.03%
3/8	394	1%
4/8	1481	4%
5/8	1983	7%
6/8	3932	17%
7/8	2251	12%
8/8	9937	58%
Grand Total	20000	100%

Table 4: Agreement Level: Agreement analysis across models using different modality combinations and fusion_topn values, after applying unified hyperparameters. The table shows the number of models (out of eight) that agreed with the majority decision on predicted emotion. “Incidence in all emotions” refers to the number of samples for which a specific agreement level occurred.“Percentage of total samples” denotes the relative frequency of that agreement level across the entire test set.

5. Conclusions and Future Directions

5.1 Emotion-LLaMA Requires Cultural and Racial Generalization Capabilities

While the Emotion-LLaMA model benefits from training on a large-scale, high-quality dataset and demonstrates impressive multimodal reasoning, our evaluation reveals its limited ability to generalize to non-Chinese populations in zero-shot settings. These findings suggest that the model is, to some extent, overfitted to culturally and demographically homogeneous data. Future iterations of emotion recognition models should integrate mechanisms for cultural, racial, and temporal adaptation, analogous to how language models can detect and respond to linguistic context. Notably, the model was still capable of correctly identifying basic emotions across populations, which suggests that targeted retraining (e.g., via LoRA fine-tuning on the final layers) could significantly enhance generalization without requiring full retraining.

5.2 Visual Features Are Underutilized in Emotion Perception

Our analyses show that audio features contributed most significantly to the model's performance, reflecting the strength of the HuBERT encoder in extracting emotional cues from voice. However, this also underscores a critical limitation in the video modality processing. Despite visual input often being richer in emotional cues, the model relies heavily on facial expressions, likely underutilizing other body language signals that are important for nuanced emotion interpretation. Additionally, the selection of a single "best" facial frame for emotion classification may not represent the broader emotional context of the entire scene. Incorporating full-body cues and dynamic context into the video analysis pipeline could yield substantial performance improvements.

5.3 Model Output Variability Suggests High Potential for Improvement

Our evaluation demonstrated that only 14% of multimodal model configurations achieved complete agreement on the predicted label across all six tested combinations. This result highlights the high degree of prediction variability and the untapped potential for improvement in the remaining 86% of cases. We hypothesize that this variability arises from the internal focus of each configuration on different modalities, leading to biased decision-making. Integrating a reasoning and voting module that conserves and leverages these "minority opinions" could allow the system to achieve greater consistency and reliability, particularly in ambiguous or complex scenarios.

5.4 Improving Emotion-LLaMA Without Changing the Core Language Model

Although we initially hypothesized that replacing the LLaMA-2 core with LLaMA-3 would yield performance improvements, our findings suggest that architectural enhancements may contribute more significantly than core upgrades alone. Specifically, we propose the addition of a MiniGPT-based voting and reasoning head, which processes multiple outputs derived from different prompt formulations of the same scene. This layer synthesizes the inputs to generate a final, contextually accurate emotional prediction and confidence score. Our experiments indicate that this approach substantially improves reasoning depth and interpretability, potentially matching human-level emotional comprehension in certain tasks.

We also noticed that in our hands we could improve the model's performance by adjusting and freezing specific hyperparameters. We believe that this improvement primarily reduced the training time required to reach optimal performance, since in most cases, even with random initializations, sufficient training would eventually lead to convergence. However, by fixing the hyperparameters early, we not only increased the model's average performance metrics, but also observed greater consistency and agreement between different modality combinations and fusion strategies.

6. Discussion

This work sought to address critical gaps in multimodal emotion recognition by enhancing the Emotion-LLaMA architecture with a modular voting-based reasoning layer. One of the most surprising findings was that improved emotional understanding did not require retraining the base model or modifying its architecture. Instead, a significant improvement in accuracy and reasoning was achieved simply by aggregating multiple outputs generated from varied prompt perspectives into a unified decision-making process. This insight suggests that the limitation lies not only in model capacity, but in the strategy used to interpret and reconcile outputs.

Among the enhancements evaluated—including cultural robustness, modality sensitivity, and architectural extensions, the addition of the MiniGPT-based voting head proved most impactful. While cultural and demographic generalization remains an important direction for long-term progress, the voting mechanism specifically addresses data loss occurring when minority outputs are discarded by default majority voting schemes. By integrating these "discarded" perspectives into the reasoning pipeline, the model achieved a richer, more nuanced understanding of emotional context—closer to human cognitive flexibility.

Despite these advancements, the model still lacks dedicated processing for full-body gestures and environmental cues, both of which play a critical role in natural emotional interpretation. Incorporating these into a future multimodal pipeline, possibly through an additional encoder specialized in body language and scene context, could enable near-human-level emotional recognition.

Looking forward, the Emotion-LLaMA framework could benefit from cultural awareness modules capable of adapting output to societal norms, much like how multilingual LLMs adapt to different languages. Real-time emotion recognition systems—especially those used in education, mental health, or human-computer interaction—could leverage this architecture for more context-aware, empathetic, and socially intelligent responses.

These findings position the voting head as a simple but powerful module to bridge the gap between deterministic prediction and true emotional reasoning. By giving space to diverse outputs, the system not only improves accuracy but more faithfully mirrors the ambiguity and complexity of human emotion.

Author's Note

Some portions of the text, including editing, structuring, and refinement of the manuscript, were assisted by OpenAI's ChatGPT. The authors are solely responsible for the content, interpretations, and conclusions presented in this work.

Acknowledgments

We would like to extend our sincere gratitude to Dr. Zheng Lian for his generous technical support and guidance throughout this project. His insights into the MER 2025 Challenge framework and the implementation of the Emotion-LLaMA architecture were invaluable in helping us navigate the technical intricacies of the study

References

1. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan K. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*. https://arxiv.org/abs/2204.14198

2. Cheng, Z., Li, S., Xu, X., Wang, D., Xu, Y., & Yang, H. (2024). Emotion-LLaMA: Emotion reasoning with multimodal large language models. *arXiv preprint arXiv:2406.11161*. https://arxiv.org/abs/2406.11161

3. Darwin, C. (1872). *The expression of the emotions in man and animals*. London: John Murray. https://www.gutenberg.org/ebooks/1227

4. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D., (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint* arXiv:2203.11171. https://arxiv.org/abs/2203.11171

5. de Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology, 59*, 279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625

6. Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3–4), 169–200:2008. https://doi.org/10.1080/02699939208411068

7. Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing, 65*, 23–36. https://doi.org/10.1016/j.imavis.2017.02.001

8. Lian, Z., Liu, R., Xu, K., Liu, X., Zhang, Y., Liu, P., Cheng, Z., Zuo, H., Ma, Z., Peng, X., Chen, X., Li, Y., Cambria, E., Zhao, G., Schuller, B.W., Tao, J. (2025). MER 2025: When affective computing meets large language models. *arXiv preprint arXiv:2504.19423*. https://arxiv.org/abs/2504.19423

9. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

10. Tsai, Y-H. H., Bai, S., Liang, P. P., Kotler, Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. https://doi.org/10.18653/v1/P19-1656

11. Zhang, H., Li, X., Bing, L. (2023). Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*. https://arxiv.org/abs/2306.02858

12. Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L.-P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2236–2246. https://doi.org/10.18653/v1/P18-1208

13. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M. (2023). MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint* arXiv:2304.10592. https://arxiv.org/abs/2304.10592

Supplementary data

1. Input prompt to the GPT model and instructions to act as voting model:

Based on descriptions of the same scene obtained from different prompt perspectives:

1. Angry.

2. In the video, a woman stands on a staircase, looking downward with a furrowed brow and a stern expression, which suggests she is feeling upset or frustrated her tone and body language convey a sense of anger, as she speaks in a sharp, abrupt manner, asking if someone deals with children her facial expressions and tone imply that she is not happy about the situation, possibly related to her work or personal life, and her direct questioning suggests a sense of urgency or frustration.

3.In the video, a woman stands on a balcony, looking down at something with a stern expression her tone and body language convey a sense of anger or frustration, as she speaks in a sharp, abrupt manner, asking if the person deals with children the combination of her stern facial expression, sharp tone, and direct question suggests that she is upset or annoyed about something related to children, possibly feeling overwhelmed or frustrated by their behavior.

4.The lady's tone in the video is angry.

Based on this information:

- What is the most likely emotional state of the individual?
- What situation could explain this behavior?
- On a scale from 1 to 10, how confident are you in this interpretation?

Model’s output: Tvot (Token Vote) Output:

Emotion: Angry

Situation: The woman appears to be addressing a stressful situation involving children, potentially in a caregiving or educational context. Her stern facial expression, sharp tone, and abrupt questioning suggest she is experiencing frustration or irritation—possibly due to challenging behavior or lack of support. The setting (balcony/staircase) implies she might be confronting someone from a physical or emotional distance, adding to the tension.

Confidence: 9/10

2. Example for output data in unimodal assay. The results for multimodal were obtained in a similar format.

			Audio features		Text features		Video features	
model_name	Most Common	Final score	attention	attention	attention	attention	attention	attention
feature_type	-		chinese-hubert-base-	chinese-hubert-large-	chinese-roberta-wwm-ext-	chinese-roberta-wwm-	clip-vit-base-	clip-vit-large-
f1_score	0.74		0.74	0.73	0.50	0.53	0.60	0.63
acc_score	0.59		0.59	0.58	0.33	0.36	0.42	0.46
Sample Name	**Majority Vote**							
samplenew3_00043973	**sad**	**67%**	sad	sad	sad	sad	happy	happy
samplenew3_00010321	**happy**	83%	happy	happy	happy	happy	neutral	happy
samplenew3_00086154	**sad**	50%	neutral	angry	sad	neutral	sad	sad
samplenew3_00003258	**angry**	50%	angry	happy	neutral	neutral	angry	angry
samplenew3_00010237	**sad**	50%	sad	sad	angry	angry	happy	sad
samplenew3_00006734	**sad**	33%	sad	sad	neutral	neutral	happy	happy

3. Hyperparameters before and after adjustment

All parameters were set to have the same starting point for all model tests, for better comparison between model's performance.

hyperparameter	Base value	Run value
epocs	10	20
Learning rate	0.001	0.0001
Drop out	0.2	0.5
Hidden dimension	64	64
Hidden dimension	64	128

4. Results Obtained in the Original Emotion-LLaMA Article

In the original *Emotion-LLaMA* study, Cheng et al. (2024) reported the model's performance across multiple datasets, including MER2023-SEMI1, MER2024-NOISE2, EMER, and DFEW, using several evaluation metrics such as F1 score, WAR (Weighted Average Recall), and UAR (Unweighted Average Recall). The results were broken down per emotion label, enabling a granular view of the model's classification accuracy across categories such as angry, sad, happy, worried, surprise, neutral, and disgust.

For a direct and meaningful comparison with our evaluation on Track 1 of the MER2025 Challenge, we observed that Track 1 includes only six emotion labels, explicitly excluding the "disgust" category. Therefore, when comparing our model's performance to that of Cheng et al., we calculated an adjusted average performance metric based on the six relevant categories (*excluding "disgust"*). This allowed for a consistent and fair comparison between the models under the same emotion label constraints.

Emotion	Happy	Sad	Neutral	Angry	Surprised	Disgust	Fear	Average score	Avg. without Disgust
Score	93.05	79.42	72.47	84.14	72.79	3.45	44.2	64.22	74.35

5. results obtained in our model with default hyperparameter configurations.

Model performance across different multimodal fusion configurations under unified hyperparameter settings. The table presents the F1 and accuracy scores for five different multimodal input combinations (Fusion1–Fusion5), each using a specific fusion strategy. fusion_topn: Number of top frames selected per video input. Input: Combination of modalities used — A (Audio), V (Video), T (Text). f1_score: F1 macro score across all emotional labels. acc_score: Overall classification accuracy.

Hyperparameter	Fusion1	Fusion2	Fusion3	Fusion4	Fusion5
fusion_topn	1	1	1	2	2
Input	AT	AVT	AV	AVT	AV
f1_score	0.74	0.79	0.77	0.79	0.76
acc_score	0.59	0.65	0.63	0.66	0.62