

## 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
epochs	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