Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning

Zebang Cheng*, **Zhi-Qi Cheng***†, Jun-Yan He, Jingdong Sun, Kai Wang,
Yuxiang Lin, Zheng Lian, **Xiaojiang Peng†**, Alexander Hauptmann

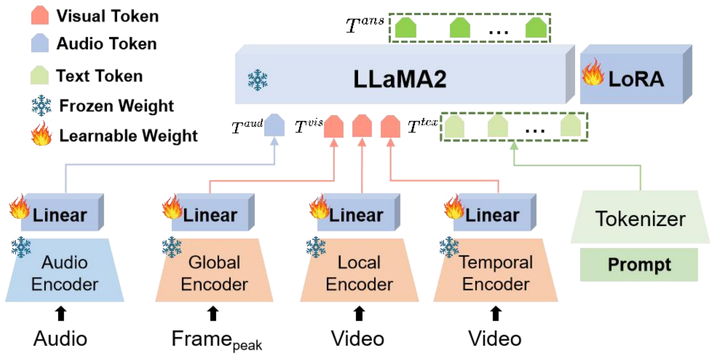
NEURAL INFORMATION PROCESSING SYSTEMS

Main Contributions

- We constructed the **MERR** dataset, consisting of 28,618 coarse-grained and 4,487 fine-grained annotated samples across diverse emotional contexts, enabling models to generalize to real-world applications and advance large-scale multimodal emotion model training and evaluation.
- We developed the **Emotion-LLaMA** model, which integrates HuBERT for audio processing and multiview visual encoders (MAE, VideoMAE, EVA) to capture facial details, dynamics, and context, enhancing emotional recognition and reasoning by aligning these features to a textual semantic space.
- Extensive experiments show that Emotion-LLaMA outperforms other MLLMs across multiple datasets, establishing it as the state-of-the-art model in public competitions. It achieved top scores on the EMER dataset (Clue Overlap: **7.83**, Label Overlap: **6.25**), F1 scores of **0.9036** on MER2023-SEMI and **0.8452** on MER2024-NOISE, and surpassed ChatGPT-4V in zero-shot evaluations, including DFEW (+**4.37%**) and MER2024-OV (+**8.52%**).

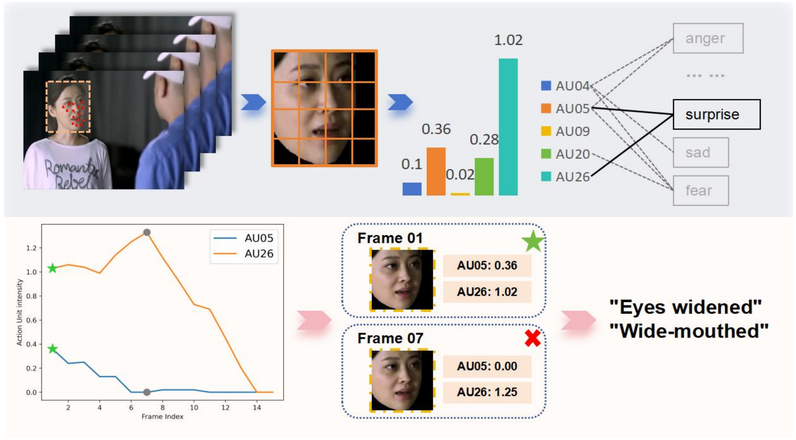
Emotion-LLaMA

To capture emotional cues in audio and visual modalities, we leverage the HuBERT model as our audio encoder and a multiview visual encoder (EVA, MAE, VideoMAE). **Emotion-LLaMA** is trained in a **coarse-to-fine** manner, consisting of the Pre-training and Multimodal Instruction Tuning.
By employing this linear projection and multimodal token representation, **Emotion-LLaMA** processes and integrates information from various modalities, leveraging the strengths of the underlying LLaMA model while incorporating essential emotional cues from audio and visual sources.



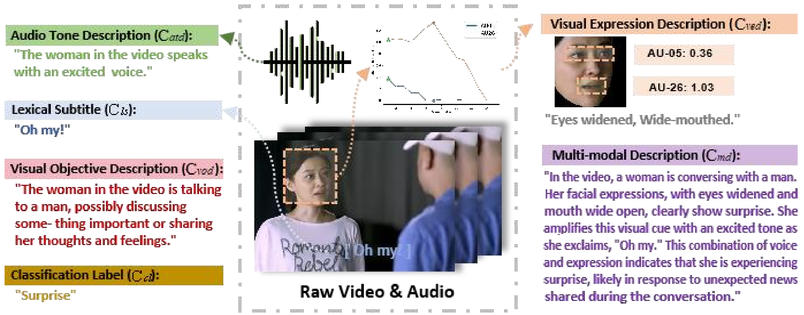
Architecture of Emotion-LLaMA, which integrates multimodal inputs for multimodal emotional recognition and reasoning.

MERR Dataset Construction



Overview of the video expression annotation process using Action Units (AUs).

The MERR dataset includes audio tone description, lexical subtitle, visual objective description, visual expression description, classification label, and multimodal description. It extends the range of emotional categories and annotations beyond those found in existing datasets.



Example of the MERR dataset.

	Sufficient Quantity	Audio Description	Visual Objective Description	Visual Expression Description	Classification Label	**Multimodal Description**
EmoSet [83]	✔	✗	✔	✓	✓	✗
EmoVIT [82]	✔	✗	✔	✓	✓	✓
DFEW [39]	✔	✗	✗	✗	✔	✗
MER2023 [51]	✔	✗	✗	✗	✔	✗
EMER [54]	✗	✔	✔	✔	✔	✔
MERR (ours)	✔	✔	✔	✔	✔	✔

A comparative analysis of several key emotional datasets, including DFEW, MER2023, EMER, and MERR.

Experiments

Models	Clue Overlap	Label Overlap
VideoChat-Text [47]	6.42	3.94
Video-LLaMA [86]	6.64	4.89
Video-ChatGPT [59]	6.95	5.74
PandaGPT [72]	7.14	5.51
VideoChat-Embed [47]	7.15	5.65
Valley [58]	7.24	5.77
Emotion-LLaMA (ours)	**7.83**	**6.25**

Method	Modality	F1 Score
MER2023-Baseline	A, V	0.8675
MER2023-Baseline	A, V, T	0.8640
Transformer	A, V, T	0.8853
FBP	A, V, T	0.8855
VAT	A, V	0.8911
Emotion-LLaMA (ours)	A, V	0.8905
Emotion-LLaMA (ours)	A, V, T	**0.9036**

Method	Hap	Sad	Neu	Ang	Sur	Dis	Fea	UAR	WAR
Zero-Shot									
Qwen-Audio [20]	25.97	12.93	67.04	29.20	6.12	0.00	35.36	25.23	31.74
LLaVA-NEXT [56]	57.46	**79.42**	38.95	0.00	0.00	0.00	0.00	25.12	33.75
MiniGPT-v2 [8]	**84.25**	47.23	22.28	20.69	2.04	0.00	0.55	25.29	34.47
Video-LLaVA(image) [55]	37.09	27.18	26.97	58.85	12.97	0.00	3.31	20.78	31.10
Video-LLaVA(video) [55]	51.94	39.84	29.78	58.85	0.00	0.00	2.76	26.17	35.24
Video-Llama [86]	20.25	67.55	**80.15**	5.29	4.76	0.00	9.39	26.77	35.75
GPT-4V [53]	62.35	70.45	56.18	50.69	32.19	10.34	**51.11**	**47.69**	54.85
Emotion-LLaMA (ours)	71.98	76.25	61.99	**71.95**	**33.67**	0.00	3.31	45.59	**59.37**
Fine-tuning									
EC-STFl [39]	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
Former-DFER [91]	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
IAL [45]	87.95	67.21	70.10	76.06	62.22	0.00	26.44	55.71	69.24
MAE-DFER [73]	92.92	77.46	74.56	76.94	60.99	**18.62**	42.35	63.41	74.43
VideoMAE [74]	93.09	78.78	71.75	78.74	63.44	17.93	41.46	63.60	74.60
S2D [10]	**93.62**	**80.25**	**77.14**	81.09	64.53	1.38	34.71	61.82	76.03
Emotion-LLaMA (ours)	93.05	79.42	72.47	**84.14**	**72.79**	3.45	**44.20**	**64.21**	**77.06**

Comparison of multimodal emotion recognition results on EMER, MER2023-SEMI and DFEW.

An Example of Multimodal Emotion Reasoning

Subtitle (En)	Are you really stupid or just pretending? Do I care about spending this money? It's just a joke.
Ground Truth	In the video, although she had a smile on her face, she questioned the other person, "Are you really stupid or just pretending?" expressing her dissatisfaction and anger towards the other person's act of pretending to be foolish. Therefore, it can be inferred that his emotional state is angry.
PandaGPT	The emotional state of this person appears to be positive and relaxed. They are smiling and laughing, indicating that they find the situation amusing or enjoyable. The woman's reaction to the man's joke suggests she is not offended or angry but finds it funny or amusing. This means that the interaction between the two individuals is friendly, not tense or uncomfortable.
VideoChat-Embed	Based on the video, the lady appears to be quite upset and angry. She discusses her financial losses and how her boyfriend seems indifferent to her feelings. She expresses a desire to spend money. The lady's facial expressions and gestures, such as frowning and clutching her purse tightly...
Valley	From the video, we can infer that the woman might be in a neutral or mixed emotional state. In some scenes, she is smiling or making amusing faces, suggesting she feels happy or amused. . .
Emotion-LLaMA	In the video, the woman's facial expressions, including a furrowed brow, lifted eyebrows, and a slightly lowered lip, suggest a mix of emotions. Her tone and intonation convey a sense of frustration, and her words reveal that she is questioning whether the other person is silly, indicating a sense of annoyance. Therefore, it is inferred that her emotional state is anger.

An example of multimodal emotion reasoning comparing Emotion-LLaMA with other MLLMs.

About Us

SZTU-MIPS
Gmail: zebang.cheng@gmail.com
Github: https://github.com/ZebangCheng/Emotion-LLaMA
Demo: https://huggingface.co/spaces/ZebangCheng/Emotion-LLaMA

Paper Code Demo