EmoThinker: Advancing Visual-Acoustic Emotion Analysis via Structural Token Selection and Chain-of-Thought Reasoning

Supplementary Material

1. Dataset

To validate the effectiveness of our EmoThinker model, we conduct experiments on different MEA datasets :DFEW [8], IEMOCAP [4], MELD [11], UR-FUNNY [7], and MUStARD [5, 12]. The dataset split and distribution of IEMOCAP and MELD are shown in Table 2 and Table 3. We provide detailed descriptions and statistics of the three datasets as below.

DFEW (Dynamic Facial Expressions in the Wild) [8] is a large-scale video-based facial expression dataset with over 16,000 movie clips covering diverse real-world challenges (e.g., illumination changes, occlusions, and large pose variations), accompanied by extensive benchmarks and a dedicated EC-STFL baseline for dynamic FER in the wild.

IEMOCAP. Interactive Emotional dyadic Motion CAPture database (IEMOCAP) is a multi-annotated MEA dataset, which has both 2-way version (*Positive* and *Negative*) and 6 labels (*Happy*, *Sad*, *Neutral*, *Anger*, *Excited*, *Frustrated*). As there are ambiguous labels in 6-way dataset, many works merged them and formed 4-way version. It contains more than 12 hours of videos from 10 speakers.

MELD. Multimodal Emotion Lines Dataset (MELD) is a multi-party dialogue dataset consisting of 13707 samples collected from the TV show *Friends*. It has two annotation versions: 2-way dataset (*Positive* and *Negative*) and 7-way dataset (*Neutral*, *Surprise*,*Fear*, *Sad*, *Joy*, *Disgust*, *Angry*).

UR-FUNNY. UR-FUNNY [7] is a diverse multimodal dataset for humor detection that captures text, visual gestures, and acoustic cues from real social interactions, providing a benchmark framework for studying multimodal language in humorous communication.

MUStARD. MUStARD [5] is a multimodal sarcasm detection dataset of audiovisual utterances from popular TV shows, each paired with dialogue context, designed to enable and benchmark multimodal sarcasm detection methods beyond text alone, with initial results showing that combining modalities significantly improves performance. Moreover, MUStARD++ [12] extends MUStARD into a larger, carefully re-annotated multimodal sarcasm–emotion dataset with corrected and doubled emotion labels, added valence and arousal scores, and four sarcasm types, along with benchmark fusion models for fine-grained emotion recognition in sarcastic utterances.

Dataset	Train	Test	Overall
DFEW	6360	2340	11700
UR-FUNNY	3156	1375	4531
MUStARD	435	102	537
MELD	9989	2610	13707
IEMOCAP	5354	1650	7532

Table 1. The details of MEA dataset statistics.

Split	Happy	Sad	Neutral	Anger	Excited	Frustrated
train	431	762	1,187	832	703	1,322
dev	21	77	137	101	39	146
test	299	245	384	170	143	381
All	751	1,084	1,708	1,103	885	1,849

Table 2. Class distribution of IEMOCAP dataset.

Split	Neutral	Surprise	Fear	Sad	Joy	Disgust	Angry
train	4,710	1,205	268	271	683	1,743	1,109
dev	470	150	40	22	111	163	153
test	1,256	281	50	68	208	402	345
all	6,436	1,636	358	361	1,002	2,308	1,607

Table 3. Class distribution of MELD dataset.

2. CoET Construction

Uni-modal Captioning. Given the raw video and audio samples, we first extract the frame-wise visual descriptions and general audio captions. For video, we employ Qwen3-VL [2] to perform batch inference with sampled video frames. The generated captions are then processed with Qwen3 [15] to clean useless format and redundancy. Additionally, for long captions, we summarize the content to control the overall length. Finally, to control the diversity of video descriptions, we compare each frames to filter the redundant captions with same semantic. For audio samples, we send them into Qwen3-Omni-Captioner [15] to generate initial audio captions. Then we obtain the final audio descriptions by refining the content and cleaning the format. The used human prompts are provided in Figure 1.

Question-Answer Pair Generation. Based on the refined uni-modal captions, we next construct MEA-oriented question–answer pairs. For each video clip, we concatenate the filtered frame-wise visual captions and the corresponding audio caption, and feed them into GPT-oss-120B [1] together with our emotion label space. We design a two-stage prompting scheme that mirrors the reasoning process

described in the main paper. In the first stage, the model is asked to summarize the key visual and acoustic evidence separately and then explicitly compare them, highlighting both consistent cues and potential conflicts (e.g., calm voice with tense posture). In the second stage, conditioned on this comparison, the model is required to synthesize a final judgment and output: (1) an MEA-focused question that targets either the discrete emotional state of a character or higher-level affective properties of the scene, (2) several candidate answers in a fixed format, and (3) the correct answer accompanied by a short natural-language justification. To ensure the quality of automatically generated pairs, we apply simple rule-based filters to remove QA pairs that violate the required structure, contain no explicit emotional concept, or express uncertainty. The remaining candidates are then inspected by human annotators, who correct ambiguous wording and discard pairs with incorrect labels or modality-biased reasoning. Finally, we collect 27,617 question-answer pairs for 20,872 videos.

Multimodal Emotion Thought Annotation. Finally, we augment each QA instance with multimodal emotion thoughts that explicitly decompose and integrate the underlying affective cues. For the visual modality, we first use the DeepFace toolkit [13] to detect all faces in the sampled frames and record their bounding boxes. For every detected face, OpenFace [3] is applied to extract the intensities of facial Action Units (AUs), which are kept as low-level numeric annotations. These AU patterns are then translated into short textual descriptions (e.g., brow raising, lip corner pulling) and combined with LLM-generated sentences about body posture, gaze direction, and interpersonal interaction to form human-centered cues. In addition, the model is asked to describe background attributes such as brightness, color saturation, and clutter, yielding a complementary set of atmosphere-related cues.

For the audio modality, we start from the refined audio captions and instruct the LLM to disentangle the paralinguistic properties into four dimensions: rhythm (speaking rate and pauses), tone (pitch height and contour), quality (timbre, tenseness, breathiness), and noise (environmental sounds, overlapping speech). A subsequent prompt refines these attributes over time to obtain temporally coherent audio cues for the whole clip.

Given the structured visual and audio cues, we again employ GPT-oss to produce the final Chain-of-Emotion-Thought. The model is prompted to: (1) restate the key cues from each modality, (2) explicitly reason about cross-modal consistency or conflict, and (3) derive the final emotion label used in CoET while explaining the decision in a step-by-step manner. We store both the intermediate cues (face boxes, AUs, human-centered and background descriptions, acoustic attributes) and the generated reasoning text as the multimodal emotion thought annotation for each QA pair.

Setting	DFEW		MELD	
	UAR	WAR	Acc	w-F1
w/o CoET	57.35	66.47	60.51	58.43
w/o STS	62.27	72.82	64.72	64.20
w/o Dilation	64.85	75.64	66.48	64.85

Table 4. Ablation study of EmoThinker with different settings on DFEW and MELD datasets. STS represent the structual token selection method, and Dilation denotes the facial dilating operation.

3. Ablation Study

To further verify the effectiveness of EmoThinker, we conduct more ablation experiments to identify the contribution of each proposed method. The results are shown in Table 4.

Effect of CoET. We first analyze the impact of the Chain-of-Emotion-Thought (CoET) framework by comparing the performance of the model with and without CoET. As shown in Table 4, removing CoET results in a significant drop in performance on both the DFEW and MELD datasets. This demonstrates that the explicit, structured reasoning process provided by CoET is crucial for improving the model's ability to perform affective reasoning and emotion recognition.

Effect of STS. Next, we examine the effect of the Structural Token Selection (STS) method, which aims to refine the model's ability to focus on relevant emotional cues from the input data. As shown in Table 4, removing STS leads to a noticeable reduction in performance, which suggests that STS plays an important role in improving the model's attention to emotional features and reducing the influence of irrelevant semantic, thereby enhancing overall performance.

4. Case Study

We provide more cases to demonstrate the inference performance between existing LVLMs [6, 9, 10, 14] and our EmoThinker. As shown in Figure 2, we first conduct the classification inference. In the case shown, we present a scenario with a video frame where a man wearing glasses is interacting with another individual. For the classification task, existing LVLMs, such as Video-ChatGPT [10], LLaMA-VID [9], and Emotion-LLaMA [6], all provide emotional labels based on their respective analyses of the video. Video-ChatGPT [10] labels the emotion as *anger* based on the man's posture and gaze direction. LLaMA-VID [9] suggests a *sad* emotion, interpreting his interaction with the other man as a conversation about a somber subject. Emotion-LLaMA [6], on the other hand, describes the man with glasses as *puzzled* attributing his facial expression to dissatisfaction or confusion about the discussion. In contrast, EmoThinker provides a more detailed interpretation. By carefully analyzing the visual and contextual cues

Base Prompt:

You are an expert of visual semantic analysis. Given a frame of a video in emotion analysis dataset. Please describe the visual details from the following aspects + [VISUAL COMPONENT PROMPTS] Output one paragraph without any format.

Human-centered: ***facial expression, head pose, eye gaze, mouth movement, and other human-centered cues.***

Object-focused: ***salient objects, visual composition, background, and other high-level visual cues.***

Visual Low-level: ***Overall brightness, color palette, and lighting conditions.***

Figure 1. Human prompts used for multi-level visual cue extraction.

Sampled Frames

Question: *What emotion is the man with glass?*

Video-ChatGPT: The man with glass is in **anger** emotion state, while looking up at the other man.

LLaMA-VID: **Sad**. The man with glass is asking the other man with a sad mood.

Emotion-LLaMA: In this video, The man with glass seems to be **puzzled** with the other man. His head was raised and his eyes were narrowed, as if he was dissatisfied with what was being discussed.

EmoThinker: Let's think step by step. This video shows two men seated in a waiting room, casually reading magazines. The man in green speaks while looking down; as he talks, the man with glasses gradually lifts his gaze, his brows knitting and eyes narrowing. His head tilts and his mouth tightens, signaling confusion and mild skepticism rather than anger. The visual cues form a narrative of social awkwardness. Therefore, the man with glasses is puzzled by the other man's remarks.

Figure 2. Human prompts used for multi-level visual cue extraction.

in the scene, EmoThinker generates a step-by-step reasoning process. It describes the scenario as two men sitting in a waiting room, with the man in glasses gradually lifting his gaze, knitting his brows, and narrowing his eyes—signals of confusion or skepticism, rather than anger. EmoThinker concludes that the man with glasses is *puzzled* by the other man's remarks.

References

[1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. 1

[2] Alibaba Cloud / Qwen Team. Qwen3-vl. https://huggingface.co/collections/Qwen/qwen3-vl, 2025. Software framework. 1

[3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018. 2

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008. 1

[5] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*, 2019. 1

[6] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024. 2

[7] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*, 2019. 1

[8] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020. 1

[9] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 2

[10] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2

[11] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018. 1

[12] Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. A multimodal corpus for emotion recog-

nition in sarcasm. *arXiv preprint arXiv:2206.02119*, 2022. 1

[13] Sefik Ilkin Serengil. Deepface: A facial analysis framework. https://github.com/serengil/deepface, 2020. GitHub repository, accessed November 13, 2025. 2

[14] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 2

[15] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1