BENCHMARKING IDENTITY-CENTRIC EMOTION TRACKING IN CONTINUOUS VIDEO CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Current multimodal conversational emotion understanding relies on a convenient but unrealistic premise: pre-segmented utterances with aligned video, audio, and text, where models reason only about the current speaker. In reality, conversations are continuous and unsegmented, and non-speakers also react emotionally. To address this gap, we introduce Identity-Centric Emotion Tracking (ICET), which removes these assumptions. ICET comprises three interdependent sub-tasks: Utterance Segmentation, Identity-Consistent Emotion Classification, and Emotion-Cause Extraction. Utterance Segmentation is performed over all participants, while the latter two sub-tasks cover both speakers and non-speakers. To support ICET, we contribute EmoTrack, a Chinese-English bilingual dataset with 22,585 utterance-level emotion-cause annotations over 21 hours of conversational videos. Benchmarking Multimodal Large Language Models (MLLMs) on ICET reveals challenges in utterance segmentation, long-range dependency modeling, and visual utilization. We further show that speech-aligned transcripts amplify both gains and errors, while reinforcement learning remains unstable due to sparse rewards and credit assignment challenges. These findings highlight EmoTrack's diagnostic value and show that ICET remains an open challenge.

1 INTRODUCTION

Emotions are central to human social interactions, shaping how we interpret intentions and build relationships Minsky (1986); Lian et al. (2025). In real-world conversations, emotions do not occur as isolated events but evolve continuously over time, often triggered by responses from multiple participants Poria et al. (2019); Zhang et al. (2023). Understanding these emotional dynamics in conversational video is a core challenge in affective computing, with applications ranging from education Schutz & Lanehart (2002) and psychological counseling Liu et al. (2021); Chu et al. (2025) to empathetic embodied robots Paiva et al. (2017); Spezialetti et al. (2020).

Prior work on multimodal conversational emotion understanding Xie et al. (2026); Jing et al. (2026); Lian et al. (2025); Cheng et al. (2024) has largely adopted a speaker-centric, pre-segmented paradigm. As summarized in Table 1, existing datasets provide utterance-level segments where temporal boundaries are manually pre-defined, and models are simply asked to predict the emotion and cause for the current speaker, as illustrated in Fig. 1. While this setup simplifies the problem, it introduces two unrealistic constraints that do not hold in real-world conversations.

First, temporal boundaries are treated as given. In reality, conversations unfold as continuous multimodal streams without explicit segmentation Zhu et al. (2025b). Models must segment utterances on their own, rather than relying on pre-defined boundaries. Second, only the active speaker is considered. Non-speakers also exhibit meaningful affective responses, such as laughter, frowns, or eye rolls. These signals reflect ongoing interaction dynamics and shape the evolution of the conversation.

To bridge this gap, we propose Identity-Centric Emotion Tracking (ICET), a new task that removes both constraints. ICET operates on unsegmented conversational video and requires models to perform three interdependent sub-tasks for a target individual, who may be either a speaker or a non-speaker: (1) Utterance Segmentation (USeg), which predicts the end timestamp of each utterance, with the i-th utterance spanning from the previous end timestamp to its own; (2) Identity-Consistent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

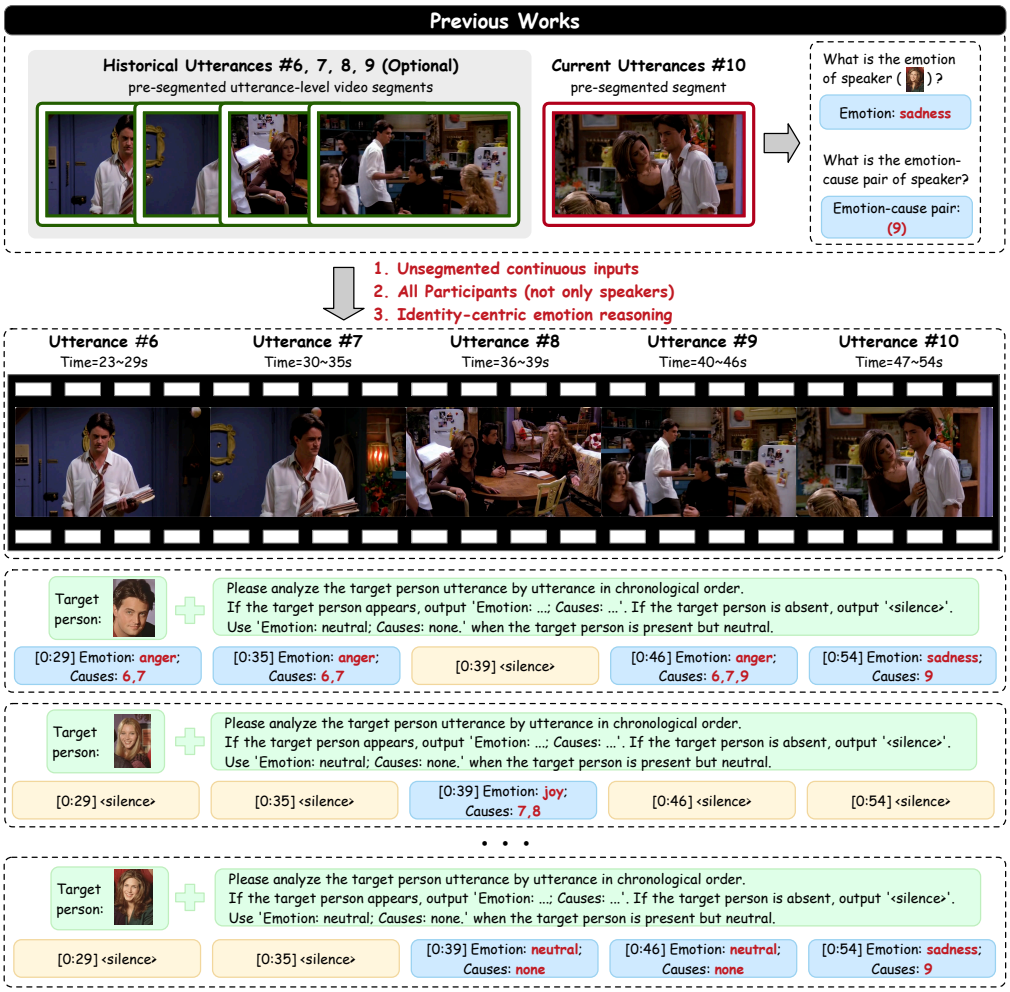


Figure 1: Comparison of existing and our proposed task on the English subset of EmoTrack.

Emotion Classification (EC), which predicts the emotional state of the target individual; and (3) Emotion-Cause Extraction (ECE), which identifies the set of utterances that cause this emotion.

To support this task, we introduce EmoTrack, a large-scale bilingual benchmark of 22,585 fine-grained utterance-level emotion-cause annotations over 21 hours of conversational videos, covering both speakers and non-speakers, on which we evaluate a range of Multimodal Large Language Models (MLLMs). We find that performance degrades with video length, with USeg dropping most sharply and ECE nearly halving on longer English videos, while shorter Chinese videos remain more stable. Yet once a coarse boundary is met, downstream emotion and cause recognition are no longer mainly constrained by boundary errors, suggesting distinct bottlenecks for segmentation and emotion reasoning. Modality utilization is imbalanced: audio provides the most reliable cues (e.g., prosody and pauses), whereas simple prompt-based visual fusion yields limited benefit, especially in multi-party settings where irrelevant participants dominate attention.

We further find that speech-aligned transcript generation acts as an amplifier: accurate transcripts improve performance, whereas noisy transcripts degrade it. Reinforcement learning yields mixed results across all three sub-tasks, likely due to sparse and delayed sequence-level rewards and difficult credit assignment over long sequences: a single scalar reward must be credited to many segmentation, classification, and extraction decisions, making optimization unstable. These findings highlight a significant gap between existing MLLM capabilities and the demands of real-world ICET.

Our contributions are: (1) ICET, a task formulation that removes pre-segmentation and speaker-centrism assumptions, requiring joint utterance segmentation and identity-consistent emotion rea-

Table 1: Summary of multimodal conversational emotion understanding datasets.

Dataset	Emotion	Cause	Unsegmented	Target Setting	Language	Utterances
IEMOCAP Busso et al. (2008)	✓	✗	✗	Speaker	English	7,433
SEMAINE McKeown et al. (2011)	✓	✗	✗	Speaker	English	5,798
MELD Poria et al. (2019)	✓	✗	✗	Speaker	English	13,708
MEISD Firdaus et al. (2020)	✓	✗	✗	Speaker	Chinese	20,000
CMMA Zhang et al. (2023)	✓	✗	✗	Speaker	Chinese	21,795
ConvECPE Li et al. (2023)	✓	✓	✗	Speaker	English	7,433
ECF Wang et al. (2023)	✓	✓	✗	Speaker	English	13,619
ECEM Wang et al. (2025)	✓	✓	✗	Speaker	English	7,273
EmoTrack (Ours)	✓	✓	✓	**All participants**	**English, Chinese**	**22,585**

Note: Single-modal or non-conversational datasets are excluded.

soning for both speakers and non-speakers; (2) EmoTrack, a Chinese-English bilingual dataset with 22,585 utterance-level emotion-cause annotations over 21 hours of conversational videos, covering both speakers and non-speakers; (3) comprehensive benchmarking of open- and closed-source MLLMs, demonstrating EmoTrack's diagnostic value for current MLLM limitations.

2 Task Definition

Prior work (Table 1) on multimodal emotion understanding typically focuses on speaker-centric settings, modeling only the current speaker's emotion. All existing approaches assume pre-segmented multimodal utterances as input, thereby bypassing the need for pre-defined temporal boundaries.

In contrast, we introduce a more practical setting, termed **Identity-Centric Emotion Tracking**. It allows any participant in a conversation to be specified as the target, including both speakers and non-speakers, and operates directly on raw, unsegmented conversational video without access to predefined utterance boundaries. This better reflects real-world scenarios, where conversational signals are continuous, multiple participants are involved, and dynamically changing analysis targets.

Formally, the input is a raw, unsegmented multimodal conversation video $\mathcal{V}$ and a designated target participant p. The goal is to model the temporal emotional dynamics of p. The model must output a variable-length sequence of utterance-level predictions:

$$\mathcal{Y} = \{(\hat{e}_i, \hat{y}_i, \hat{\mathcal{C}}_i)\}_{i=1}^{K}, \tag{1}$$

where K is the number of utterances predicted by the model over all participants, $\hat{e}_i$ denotes the end timestamp of the i-th predicted utterance with $\hat{e}_0 = 0$, $\hat{y}_i$ is the emotion state of p associated with the i-th utterance, and $\hat{\mathcal{C}}_i \subseteq \{1, \ldots, K\}$ is the set of predicted utterances that cause this emotion.

3 EmoTrack Dataset

3.1 Data Source

English Source. The English subset is derived from ECF Wang et al. (2023), which extends MELD Poria et al. (2019). MELD provides utterance-level speaker-centric emotion annotations on *Friends* across seven emotions (anger, disgust, fear, joy, neutral, sadness, surprise), and ECF adds speaker emotion-cause annotations. Both are pre-segmented clips that discard transitional segments, so concatenating them would introduce artificial discontinuities. We therefore retrieve continuous segments from the original episodes and re-align them with ECF utterances using episode-level timestamps, which we manually verify and correct. These timestamps serve as ground-truth utterance boundaries for USeg. On the reconstructed video, we retain speaker annotations and additionally annotate non-speaker emotions and causes.

Chinese Source. For the Chinese subset, no existing dataset provides emotion-cause annotations, so we base it on CMMA Zhang et al. (2023), which contains utterance-level emotion annotations for multiple Chinese TV dramas (e.g., *My Own Swordsman*, *The Story of the Cooking Class*) with an ECF-aligned seven-emotion taxonomy. We annotate emotion causes for speakers and both emotions and causes for non-speakers. We also replaced a small number of questionable speaker labels upon

consensus among three annotators. CMMA contains many annotation errors, such as sentence-video misalignment (missing/extra sentences, split/merged sentences) and inconsistent visual-text content; we discarded samples with two or more issues and repaired the rest. Finally, as with the English subset, CMMA clips cannot be directly concatenated, so we return to the original videos and manually verify and correct episode-level timestamps.

3.2 Dataset Annotation

To ensure annotation quality, we recruit graduate-level annotators with research backgrounds in affective computing. Specifically, the Chinese subset is annotated by three native Chinese speakers (two male and one female). The English subset is annotated by three annotators (two female and one male), including a native English speaker and two PhD students working in an English-speaking research environment, ensuring high-quality annotations for English data.

Annotation Strategy. Annotating unsegmented conversational videos is challenging due to complex multi-party interactions, long-range dependencies, and the difficulty of consistently tracking an individual's emotional state. In particular, labeling entire videos directly may lead to annotator fatigue and reduced reliability, especially for fine-grained emotional shifts and their underlying causes. To address this, we first perform annotation at the utterance level and then aggregate the annotations to construct continuous emotion–cause trajectories. Prior to annotation, annotators are trained with detailed guidelines to ensure consistency. During annotation, they first review full conversational transcripts together with synchronized video streams to capture contextual flow and interpersonal dynamics, and then label each utterance with emotion and cause information based on multimodal cues, including facial expressions, vocal tone, and linguistic content. For non-speaker cases, only visual cues are available.

Table 2: Inter-annotator agreement metrics.

Dataset	Annotators	Cohen's Kappa	Fleiss' Kappa
EmoTrack (English)	A&B	0.6059	0.6132
	A&C	0.6019	
	B&C	0.6319	
EmoTrack (Chinese)	A&B	0.6459	0.6385
	A&C	0.6219	
	B&C	0.6479	

To ensure annotation reliability under such conditions, we adopt a conservative filtering strategy: ambiguous visual evidence (e.g., side or occluded facial views) is retained only when full annotator agreement is achieved; otherwise, the instance is discarded. Emotion categories follow Ekman's six basic emotions Ekman (1992) along with a neutral category. For each non-neutral emotion, causes are identified at the utterance level by selecting relevant utterances that explicitly express events or arguments associated with the emotional state Lee et al. (2010); Chen et al. (2010).

Finally, utterance-level annotations are integrated into continuous conversational videos to form temporally coherent emotion–cause trajectories. This ensures clear semantic boundaries while maintaining relatively stable emotional expression within each utterance. All annotations are performed independently by three annotators, with final labels for both emotions and causes determined via majority voting. In cases of disagreement, a senior annotator reviews the instance and makes the final decision. Detailed annotation procedure is provided in Appendix B.

Annotation Quality. We compute inter-annotator agreement using Cohen's Kappa Cohen (1960) and Fleiss' Kappa McHugh (2012), where the former measures pairwise consistency and the latter measures overall agreement across three annotators. Following Wang et al. (2023), emotion and cause labels are binarized at the utterance level, and scores are computed separately and then averaged. In Table 2, all Kappa values exceed 0.6, indicating substantial agreement.

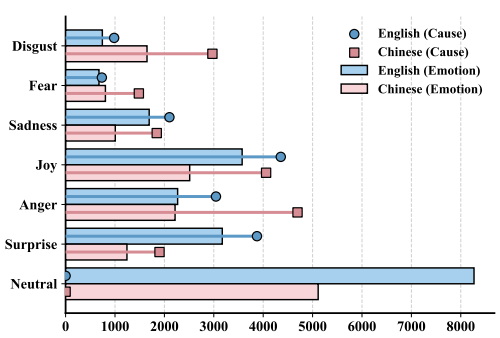
3.3 Dataset Statistics

Overall Statistics. As shown in Table 3 and Fig. 2, EmoTrack contains over 22K utterances across English and Chinese subsets, covering more than 21 hours of video. The two subsets exhibit distinct characteristics: English conversations are generally longer and more verbose, while Chinese dialogues are more compact but involve a larger number of target individuals, leading to denser multi-party interactions. The dataset is partitioned into train/dev/test splits with a ratio of 7:1:2.

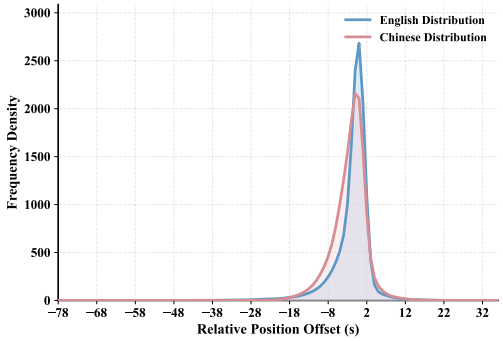
Emotion Distribution. As shown in Fig. 2, neutral dominates both Chinese and English data, while joy is the most frequent non-neutral emotion and fear is the least frequent. We also observe clear cross-linguistic differences: surprise is more prevalent in English, whereas disgust is

Table 3: Statistics of the EmoTrack dataset.

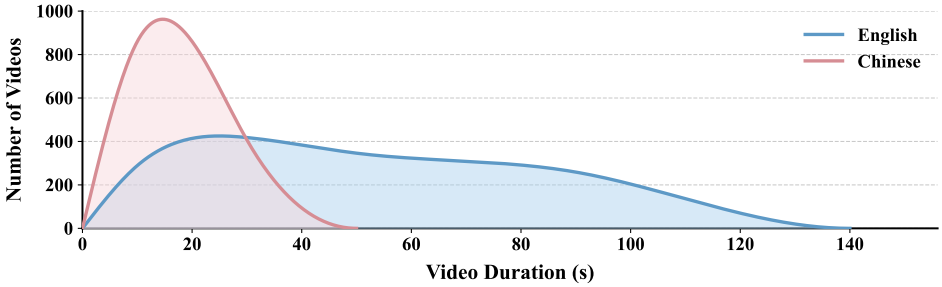
Item	English Subset			Chinese Subset			Overall		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
# Conversations	947	112	253	887	119	261	1,834	231	514
# Utterances	9,733	1,082	2,531	6,466	881	1,892	16,199	1,963	4,423
# Target person	6	6	6	80	62	76	86	68	82
# Words	109,201	12,289	29,103	37,118	5,001	10,734	146,319	17,290	39,837
# Unique words	6,148	1,630	2,743	6,889	1,583	2,896	13,037	3,213	5,639
# Video duration	10.74h	1.19h	2.76h	4.43h	0.60h	1.32h	15.17h	1.80h	4.08h
Avg. Turns per Conversation	2.8	3.0	2.8	2.3	2.3	2.3	2.6	2.7	2.6
Avg. Utterances per Conversation	10.28	9.66	10.00	7.29	7.40	7.25	8.83	8.50	8.61
Avg. Words per Utterances	11.2	11.4	11.5	5.7	5.7	5.7	9.03	8.81	9.01
Avg. Words per Conversation	115.3	109.7	115.0	41.8	42.0	41.1	79.8	74.8	77.5
Avg. Utterance Duration	3.11s	3.20s	3.09s	2.38s	2.30s	2.34s	2.82s	2.80s	2.77s
Avg. Conversation Duration	40.84s	38.40s	39.29s	17.99s	18.25s	18.19s	29.79s	28.02s	28.58s



(a) Distribution of Emotions and Causes.



(b) Relative Position Distribution of Causes.



(c) Distributions of Continuous Video Durations.

Figure 2: Statistical distributions of the EmoTrack dataset.

more frequently in Chinese. These results reveal both shared distribution patterns and cross-lingual variations, suggesting that emotion recognition models should account for language-dependent expression biases rather than assuming a universal distribution.

Cause Distribution. As shown in Fig. 2, most non-neutral utterances in both languages are linked to explicit causal utterances, indicating that emotions are largely grounded in observable context. Causal utterances concentrate near corresponding utterance, peaking at 0s with a long-tailed distribution, suggesting that emotions are mainly driven by nearby events with limited long-range dependencies, consistent with prior findings Li et al. (2023).

4 Evaluation Benchmark

We benchmark diverse MLLMs (see Appendix D) on EmoTrack across the following three tasks.

Table 4: Model performance (%) comparison of open-source and closed-source MLLMs on EmoTrack. Models marked with ⋆ have undergone supervised fine-tuning.

Model	Utterance Segmentation				Emotion Classification		Emotion-Cause Extraction	
	IoU≥0.1	IoU≥0.3	IoU≥0.5	mIoU	Acc.	W-F1	Acc.	W-F1
English Subset								
⋆ Qwen2.5-Omni-7B Xu et al. (2025)	**47.37**	**37.91**	**32.38**	**29.76**	**40.99**	**41.98**	**9.61**	**16.09**
⋆ Qwen2.5-VL-7B Bai et al. (2025)	23.43	17.64	11.91	11.67	10.56	15.39	4.62	7.60
⋆ Qwen2.5-VL-32B Bai et al. (2025)	25.79	18.92	12.21	12.28	10.34	13.86	5.18	8.12
⋆ InternVL3-8B Zhu et al. (2025a)	30.52	21.53	14.34	13.92	36.00	29.75	3.31	6.02
⋆ Emotion-LLaMA-7B Cheng et al. (2024)	30.69	20.97	13.52	14.01	35.78	29.45	3.08	4.99
⋆ InternVL3-38B Zhu et al. (2025a)	24.14	18.09	13.11	11.60	32.78	29.72	1.82	2.49
GPT-4o OpenAI et al. (2024)	12.21	8.26	6.05	6.00	28.85	25.94	0.42	1.58
GPT-5.2 Singh et al. (2026)	**27.05**	**18.81**	**13.71**	**13.26**	14.06	20.15	0.61	0.00
Gemini 2.5 Pro Comanici et al. (2025)	17.63	10.52	6.50	7.41	16.10	22.63	0.09	0.00
Qwen3.8 Omni Flash Qwen Team (2026)	25.93	13.87	8.47	10.11	31.71	38.28	**3.59**	**8.10**
Gemini 3.1 Pro Comanici et al. (2025)	17.84	10.47	7.19	7.88	**35.40**	**40.24**	3.40	6.73
Seed 2.0 Pro Seedance et al. (2026)	25.27	14.79	9.52	10.36	12.47	18.15	0.84	0.96
Chinese Subset								
⋆ Qwen2.5-Omni-7B Xu et al. (2025)	**76.20**	**72.16**	**63.04**	**55.43**	**36.84**	33.29	3.22	11.30
⋆ Qwen2.5-VL-7B Bai et al. (2025)	54.62	47.64	37.11	32.42	24.63	28.87	**5.62**	14.16
⋆ Qwen2.5-VL-32B Bai et al. (2025)	50.10	44.34	33.58	30.13	22.21	26.07	4.85	14.14
⋆ InternVL3-8B Zhu et al. (2025a)	57.87	50.30	38.35	34.47	34.51	**36.24**	4.65	**15.84**
⋆ Emotion-LLaMA-7B Cheng et al. (2024)	62.79	54.89	39.52	34.95	30.05	23.84	1.89	3.85
⋆ InternVL3-38B Zhu et al. (2025a)	59.99	51.61	39.99	35.22	30.98	30.09	2.09	4.67
GPT-4o OpenAI et al. (2024)	16.39	14.53	10.80	8.45	24.10	23.83	1.43	3.34
GPT-5.2 Singh et al. (2026)	31.72	26.13	18.85	16.20	13.80	18.25	1.69	3.11
Gemini 2.5 Pro Comanici et al. (2025)	26.60	17.76	10.76	11.60	15.16	20.91	0.41	0.78
Qwen3.8 Omni Flash Qwen Team (2026)	39.72	25.04	12.34	15.26	**31.38**	36.62	4.70	10.73
Gemini 3.1 Pro Comanici et al. (2025)	21.41	13.31	8.82	9.33	31.02	**36.82**	**4.90**	**11.98**
Seed 2.0 Pro Seedance et al. (2026)	**57.08**	**43.17**	**25.96**	**25.47**	18.78	24.69	2.09	2.61

Utterance Segmentation. This sub-task evaluates the model's ability to predict the end timestamp of each utterance in continuous conversational video, where the i-th utterance is defined as the interval $(\hat{e}_{i-1}, \hat{e}_i]$. Following prior work Chen et al. (2024); Xia et al. (2025) on streaming and temporal event modeling, we adopt metrics that capture temporal alignment quality under different overlap criteria. Specifically, we report performance under multiple temporal Intersection-over-Union (tIoU) thresholds, including tIoU ≥ 0.1, tIoU ≥ 0.3, and tIoU ≥ 0.5. Each segment is defined by consecutive boundary timestamps, i.e., $(e_{i-1}, e_i]$ for ground truth and $(\hat{e}_{j-1}, \hat{e}_j]$ for predictions. The tIoU between two segments is computed as:

$$\mathrm{tIoU}(i,j) = \frac{\max\left(0, \min(e_i, \hat{e}_j) - \max(e_{i-1}, \hat{e}_{j-1})\right)}{(e_i - e_{i-1}) + (\hat{e}_j - \hat{e}_{j-1}) - \max\left(0, \min(e_i, \hat{e}_j) - \max(e_{i-1}, \hat{e}_{j-1})\right)}. \tag{2}$$

For each threshold τ, a prediction is considered correct if it matches the ground-truth segment with the same utterance index and $\mathrm{tIoU} \geq \tau$; a missing index has tIoU 0. We report Accuracy@τ, defined as the proportion of correctly matched ground-truth segments. Additionally, we compute mean IoU (mIoU) over all ground-truth utterances to measure overall temporal alignment quality.

Emotion Classification. This task evaluates the model's ability to predict the emotional state of a target individual given the conversational context. Following prior work Tu et al. (2026); Shi & Huang (2023), we report Accuracy (Acc.) and weighted F1 score (W-F1) as evaluation metrics, excluding silence utterances.

Emotion-Cause Extraction. This task aims to identify the utterances that trigger the target individual's emotional state. Following prior work Li et al. (2023); Wang et al. (2023), we report Accuracy and weighted F1 score for cause identification, excluding neutral and silence utterances. Notably, we adopt a stricter criterion: a predicted cause is considered correct only if the associated emotion is also correctly predicted, and Accuracy further requires the cause-index set to match exactly.

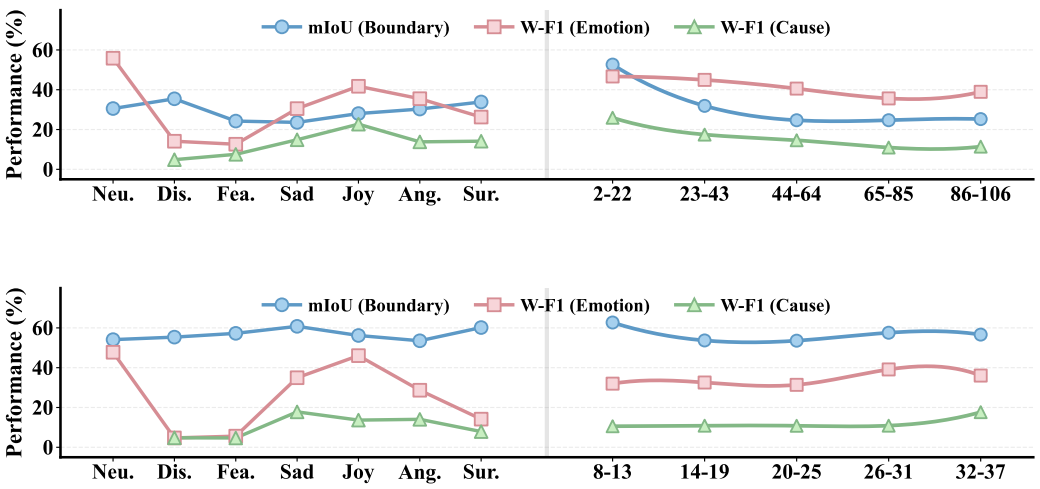


Figure 3: Model performance (%) on three tasks (USeg, EC, ECE) across emotion categories and video durations. Top: English subset; Bottom: Chinese subset.

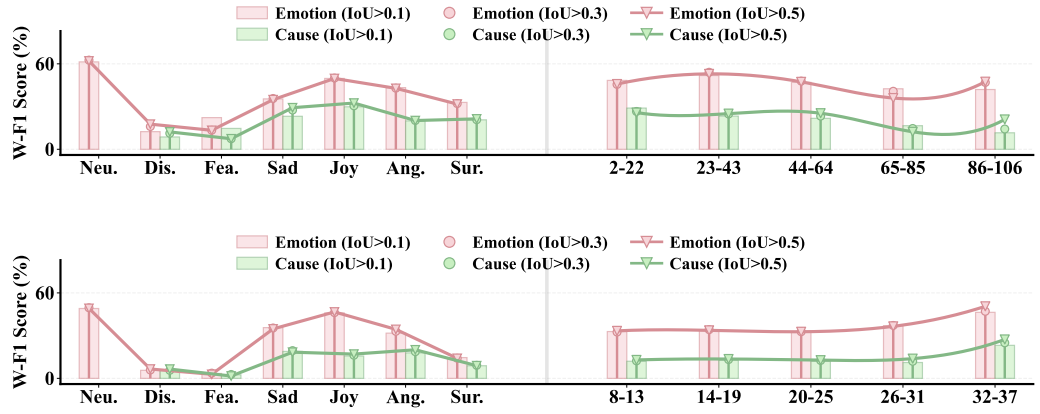


Figure 4: Model performance (%) on EC and ECE across emotion categories and video durations under different IoU thresholds. Top: English subset; Bottom: Chinese subset.

5 Results and Discussion

In this section, we present the experimental results and provide an in-depth analysis. Detailed implementation information can be found in Appendix E.

Main Results. Table 4 reports the performance of open-source and closed-source MLLMs on EmoTrack across three progressively challenging tasks: Emotion Classification (EC), Utterance Segmentation (USeg), and Emotion-Cause Extraction (ECE).

For USeg, all models perform better on Chinese than English, consistent with shorter videos reducing cumulative temporal drift. Qwen2.5-Omni performs best, indicating that utterance boundaries are primarily acoustic-prosodic events: pauses and prosody provide sharper boundary evidence than visual semantics. While frame-based models such as GPT-5.2 suffer from sparse sampling, video-capable models like Gemini 3.1 Pro do not consistently outperform them and can even lag behind. This suggests that continuous video input alone does not guarantee temporal grounding: many video-capable models apply coarse temporal compression, let irrelevant visual content in multi-party scenes dominate attention, and underuse audio.

Table 5: Model Performance (%) of Qwen2.5-Omni-7B under different modality configurations, with transcript generation, and with GRPO-based reinforcement learning on EmoTrack.

Model	Utterance Segmentation				Emotion Classification		Emotion-Cause Extraction	
	IoU≥0.1	IoU≥0.3	IoU≥0.5	mIoU	Acc.	W-F1	Acc.	W-F1
English Subset								
Qwen2.5-Omni-7B	**47.37**	**37.91**	32.38	**29.76**	40.99	41.98	9.61	16.09
w/o Audio	23.54	17.88	13.13	11.56	36.60	30.03	3.64	4.91
w/o Video	45.90	37.71	**32.60**	29.59	41.38	41.28	9.65	16.25
w/ Transcript	44.24	36.82	31.85	29.00	36.54	39.86	9.47	16.24
w/ GRPO	46.75	37.51	31.83	29.24	**44.16**	**43.62**	**11.52**	**18.31**
Chinese Subset								
Qwen2.5-Omni-7B	76.20	72.16	63.04	55.43	36.84	33.29	3.22	**11.30**
w/o Audio	65.15	58.02	43.82	38.69	35.37	29.97	3.06	8.34
w/o Video	63.34	59.31	51.09	44.59	35.24	32.73	**3.63**	11.12
w/ Transcript	**77.99**	**74.57**	**67.93**	**59.77**	**37.97**	**34.14**	3.47	9.82
w/ GRPO	77.27	73.35	64.26	56.41	37.13	32.61	3.12	10.57

For EC, the language trend is mixed: Qwen2.5-Omni and Gemini 3.1 Pro favor English, while several SFT models favor Chinese. ECE shows a clearer pattern: most models perform better on Chinese, similar to USeg, suggesting that cause extraction benefits from shorter, more focused Chinese videos. Overall, SFT models outperform non-SFT ones on both EC and ECE, highlighting the value of task-specific tuning; however, strong general models like Gemini 3.1 Pro and Qwen3.8 Omni Flash surpass many SFT models, showing that general capabilities also help.

Further Analysis. We further analyze **Qwen2.5-Omni-7B** as a case study, as it achieves the best average performance across the three sub-tasks and both languages in Table 4.

For **modality contribution**, Table 5 analyzes the effect of different modalities. For USeg, removing video leads to a slight drop on English but a substantial drop on Chinese, whereas removing audio causes a significant drop in both languages. This indicates that audio provides essential boundary cues, while visual inputs fused via simple prompting provide limited benefit that varies by language. For EC and ECE, removing audio also causes clear drops, while removing video has little effect, suggesting that visual information is largely unused by simple prompt-based fusion.

Across **emotion categories**, as shown in Fig. 3, USeg is relatively stable and less sensitive to category distribution than EC and ECE, which fluctuate more substantially. In both Chinese and English settings, low-frequency categories such as disgust and fear consistently yield lower performance, while high-frequency categories such as neutral and joy generally achieve better results. This suggests that EC and ECE depend heavily on category frequency, whereas USeg relies more on acoustic-prosodic boundary cues. Regarding **video length**, all three tasks degrade as videos become longer in the English setting. USeg shows the most pronounced drop, decreasing sharply and then plateauing. EC declines more gradually, while ECE nearly halves across length buckets. In contrast, Chinese videos are shorter and show no clear degradation. While part of the Chinese–English gap is attributable to video length, the sharp degradation on longer English videos still points to uncertainty and error accumulation in utterance segmentation.

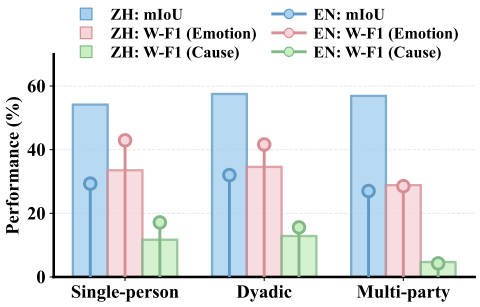


Figure 5: Model performance (%) across different interaction scenarios (Single-person, Dyadic, Multi-party) on EmoTrack.

We further analyze how USeg quality affects EC and ECE. As shown in Fig. 4, stricter IoU thresholds bring limited gains: EC remains nearly flat, and ECE improves only slightly. Only the longest English videos show a partial EC recovery. This suggests that once a coarse boundary is met, emotion and cause recognition are no longer mainly constrained by utterance boundary errors.



(a) Single-person segment. (b) Dyadic segment. (c) Multi-party segment.

Figure 6: Attention heatmap for the target person (Chandler) on the English subset.

Attention Visualization. Given the weak performance of the video modality, we visualize attention patterns under different interaction settings. In Fig. 6, attention to the target person decreases noticeably from single-person to dyadic settings and nearly disappears in multi-party scenarios, suggesting that additional participants introduce substantial visual noise. We further analyze model performance across these settings, as shown in Fig. 5. USeg remains relatively stable across settings, with only a slight drop in the multi-party English case, likely because it relies more on audio signals. In contrast, EC and ECE perform similarly in single-person and dyadic settings, but degrade substantially in multi-party scenarios, especially ECE, which drops sharply in both English and Chinese. This indicates that identifying the target person's emotional state and causes becomes increasingly difficult as more participants are involved, with cause extraction suffering the most.

Effect of Transcript Generation. Due to the dominant role of the audio modality, we explore speech-aligned transcript generation as an auxiliary output (Table 5). Its effect is language- and task-dependent, with the language gap being most pronounced: it improves USeg and EC on Chinese but not English, and barely helps ECE. English transcripts are far noisier (WER 13.5%, 24% utterance-level match) than Chinese (CER 5.5%, 60% match). Grouping by transcript correctness, correct transcripts yield much higher USeg mIoU (78.66 vs. 36.41) and better EC/ECE W-F1 (35.80/11.62 vs. 31.90/7.65). These results indicate that transcripts act as an amplifier: accurate transcripts amplify useful audio cues, while noisy ones amplify errors.

Effect of Reinforcement Learning. Due to the difficulty of utterance segmentation and multi-step decision making, we further explore reinforcement learning with Group Relative Policy Optimization (GRPO) Shao et al. (2024). In Table 5, GRPO yields mixed results. It improves EC and ECE on English but slightly degrades USeg, while on Chinese it slightly improves USeg and EC accuracy but degrades EC W-F1 and ECE. We attribute this to the reward structure of ICET: sequence-level and delayed rewards must be credited to many segmentation, classification, and extraction decisions over long conversations, making credit assignment difficult. Discrete tasks such as EC and ECE can still exploit such coarse rewards, whereas USeg requires fine-grained temporal localization, where sequence-level rewards are too coarse to resolve boundary errors. These results suggest that GRPO does not uniformly benefit all sub-tasks and motivate task-specific reward design.

6 Conclusion

We introduced ICET, which removes the unrealistic assumptions of pre-segmented utterances and speaker-centrism, requiring joint utterance segmentation, identity-consistent emotion classification, and emotion-cause extraction in continuous conversational video. To support ICET, we contributed EmoTrack, a bilingual dataset with 22,585 fine-grained utterance-level emotion-cause annotations over 21 hours of conversational videos, covering both speakers and non-speakers. Benchmarking diverse MLLMs yields three findings. First, video length is a major bottleneck with a systematic language gap: all three tasks degrade on longer English videos, with USeg dropping most sharply and ECE nearly halving, while shorter Chinese videos remain more stable; yet downstream tasks are no longer mainly constrained by boundary errors once a coarse boundary is met. Second, modality utilization is imbalanced: audio provides essential cues, whereas simple prompt-based visual fusion yields limited benefit, suggesting that visual information is not effectively utilized. Third, speech-aligned transcripts amplify both gains and errors, while reinforcement learning yields mixed results across all three sub-tasks due to sparse and delayed sequence-level rewards and difficult credit assignment over long conversations. These findings highlight that ICET remains an open challenge.

Ethics statement

EmoTrack is built by extending publicly released conversational datasets, without new recordings of private individuals. The English subset follows ECF Wang et al. (2023), derived from MELD Poria et al. (2019) and the television series *Friends*; the Chinese subset follows CMMA Zhang et al. (2023), drawn from Chinese television dramas. This work uses only publicly available datasets and does not involve new human subjects; therefore, IRB approval was not required. Annotations were produced by trained graduate annotators in affective computing, who were compensated and provided informed consent, following the protocol in Section 3.2 and Appendix B. The conversations are scripted performances, often stylized for an audience, so the labels describe portrayed affect for benchmark research. They are not clinical or psychological assessments, and should not be treated as a universal measure of human emotion, or as evidence that a system generalizes to spontaneous interaction, clinical populations, or groups underrepresented in these programs. The source programs may not represent diverse populations, and the emotion labels reflect scripted portrayals rather than spontaneous behavior; results should be interpreted with these biases in mind.

We release EmoTrack for scientific research. The underlying programs remain copyrighted, and use of the release is subject to the licenses of the source datasets. We do not redistribute the original videos; we release only annotations, splits, and derived features. Identity-centric tracking in continuous video could be misused for surveillance, profiling, or any application that tracks individuals without consent. To reduce misuse, we release annotations without raw videos or identity-recognition tools, and we discourage any use for surveillance, profiling, or tracking individuals without consent. In high-risk applications such as mental health diagnosis, emotional crisis intervention, or judicial interrogation, these technologies should serve only as assistive tools and must not be the sole basis for decision-making.

Reproducibility statement

We specify the ICET formulation, including the designated target participant and the three sub-tasks (USeg, EC, and ECE), in Section 2. Data sources, cross-modal alignment filtering, the 7:1:2 train/dev/test partition, and corpus statistics are provided in Section 3, Table 3, and Fig. 2. Annotation guidelines, the labeling interface, and inter-annotator agreement are described in Section 3.2 and Appendices B and C. Metric definitions for USeg, EC, and ECE are given in Section 4. Evaluated models and their versions are listed in Appendix D. Hardware, fine-tuning and inference configurations, random seeds, and closed-source API settings are reported in Appendix E. All reported results are averaged over multiple runs. Fig. 11 shows the prompt template used for EmoTrack. Our implementation code for training, inference, and evaluation is provided as supplementary material. Due to copyright and license restrictions, we do not redistribute the original videos; we release annotations, splits, and derived features. An anonymized sample of EmoTrack is available at `https://kaggle.com/datasets/3a4f8e57a41c5ab40d122ac0e1a6cbb6c1a862b018c98c17c1aba305cb00dd83`, and the full annotation package will be released upon acceptance.

References

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

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(4):335–359, 2008.

Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10761–10770, 2023.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 179–187, 2010.

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

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*, 2025.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Zixiang Ding, Rui Xia, and Jianfei Yu. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3574–3583, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.290. URL https://aclanthology.org/2020.emnlp-main.290/.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. Transition-based directed graph construction for emotion-cause pair extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3707–3717. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.342. URL https://aclanthology.org/2020.acl-main.342/.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4441–4453, 2020.

James J Gross and Lisa Feldman Barrett. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion Review*, 3(1), 2011.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2594–2604, 2018.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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. doi: 10.1109/TASLP.2021.3122291.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Ran Jing, Geng Tu, Yice Zhang, and Ruifeng Xu. Causal-erc: A multimodal framework with causal prompting for emotion recognition in conversations with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31383–31391, 2026.

Xincheng Ju, Dong Zhang, Junhui Li, Shoushan Li, and Guodong Zhou. Enhanced generative framework with llms for multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Multimedia*, 2025.

Dohee Kang, Daeha Kim, Donghyun Kang, Taein Kim, Bowon Lee, Deokhwan Kim, and Byung Cheol Song. Beyond superficial emotion recognition: Modality-adaptive emotion recognition system. *Expert Systems with Applications*, 235:121097, 2024.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. A text-driven rule-based system for emotion cause detection. In *Proceedings of the North American Chapter of the Association for Computational Linguistics 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 45–53, 2010.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *CoRR*, abs/2309.11911, 2023. doi: 10.48550/ARXIV.2309.11911. URL https://doi.org/10.48550/arXiv.2309.11911.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(11):1–19, 2025.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2852–2861, 2017.

Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. ECPEC: Emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(03):1754–1765, 2023.

Zheng Lian, Bin Liu, and Jianhua Tao. Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000, 2021.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 9610–9614, 2023a.

Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable multimodal emotion reasoning. *CoRR*, 2023b.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*. ML Research Press, 2025.

Qiao Liang, Ying Shen, Tiantian Chen, and Lin Zhang. M3hg: Multimodal, multi-scale, and multi-type node heterogeneous graph for emotion cause triplet extraction in conversations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11416–11431, 2025.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 3469–3483, 2021.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2011.

Marvin Minsky. *Society of mind*. Simon and Schuster, 1986.

Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. *arXiv preprint arXiv:2311.04507*, 2023.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, et al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3):1–40, 2017.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

Qwen Team. Qwen3.8-omni-flash: Omni senses. agentic delivery., September 2026. URL https://qwen.ai/blog?id=qwen3.8-omni-flash.

Paul A Schutz and SL Lanehart. Emotions in education. *Educational Psychologist*, 37(2):67–68, 2002.

Team Seedance, De Chen, Liyang Chen, Xin Chen, et al. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

R Selvi and C Vijayakumaran. An efficient multimodal emotion identification using fox optimized double deep q-learning. *Wireless Personal Communications*, 132(4):2387–2406, 2023.

HM Shahzad, Sohail Masood Bhatti, Arfan Jaffar, Muhammad Rashid, and Sheeraz Akram. Multimodal cnn features fusion for emotion recognition: A modified xception model. *IEEE Access*, 11:94281–94289, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1551–1560, 2021.

Tao Shi and Shao-Lun Huang. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of ACL*, pp. 14752–14766, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, et al. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:532279, 2020.

Geng Tu, Jun Wang, Zhenyu Li, Shiwei Chen, Bin Liang, Xi Zeng, Min Yang, and Ruifeng Xu. Multiple knowledge-enhanced interactive graph network for multimodal conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3861–3874, 2024.

Geng Tu, Ran Jing, Xuan Luo, Erik Cambria, Wenjie Li, and Ruifeng Xu. Is multimodal conversational emotion recognition satisfactory? exploring the gaps in performance, generalization, and confidence. *Pattern Recognition*, pp. 113087, 2026.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(03):1832–1844, 2023.

Lin Wang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Zhitao Zhang. Generative emotion cause explanation in multimodal conversations. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 1394–1403, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3731715.3733348. URL https://doi.org/10.1145/3731715.3733348.

Zhen Wu, Xinyu Dai, and Rui Xia. Pairwise tagging framework for end-to-end emotion-cause pair extraction. *Frontiers of Computer Science*, 17(2):172314, 2023.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. *arXiv preprint arXiv:2512.21334*, 2025.

Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1003–1012, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1096. URL https://aclanthology.org/P19-1096/.

Bo Xie, Junhao Wang, Haixia Guo, Pengliang Chen, Hua Zhang, Bo Jiang, Ye Wang, and Liwen Chen. Mspf: A multi-semantic prompting fusion framework for emotion-cause pair extraction in conversations. *Information Processing & Management*, 63(1):104356, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pp. 277–292. Springer, 2024.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3718–3727, 2020.

Yazhou Zhang, Yang Yu, Qing Guo, Benyou Wang, Dongming Zhao, Sagar Uprety, Dawei Song, Qiuchi Li, and Jing Qin. Cmma: benchmarking multi-affection detection in chinese multi-modal conversations. *Advances in Neural Information Processing Systems*, 36:18794–18805, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Xianxun Zhu, Heyang Feng, Erik Cambria, Yao Huang, Ming Ju, Haochen Yuan, and Rui Wang. Emvas: end-to-end multimodal emotion visualization analysis system. *Complex & Intelligent Systems*, 11(8):374, 2025b.

A Related Work

Multimodal Emotion Reasoning (MER). MER aims to infer human affect by integrating signals from multiple modalities such as text, audio, and vision. However, emotion perception is often inherently ambiguous, and different modalities may convey conflicting affective cues Busso et al. (2008); Yu et al. (2020). To improve annotation consistency, prior work has largely relied on constrained label spaces and aggregation strategies such as majority voting or probabilistic filtering Li et al. (2017); Lian et al. (2023a); Kang et al. (2024); Selvi & Vijayakumaran (2023); Shahzad et al. (2023). While effective for improving reliability, these approaches are still limited to predefined emotion categories and may fail to capture more nuanced emotional expressions. To address this limitation, Lian et al. Lian et al. (2023b) proposed Explainable Multimodal Emotion Reasoning (EMER), which leverages MLLMs to generate open-form natural language explanations, shifting emotion modeling from classification to generation-based reasoning.

Multimodal Emotion Recognition in Conversations (MERC). The emotion generation theory Gross & Barrett (2011) highlights the important role of contextual information in emotion recognition. Early studies mainly focused on textual modality, modeling conversational context and speaker dependencies Shen et al. (2021). With the increasing importance of multimodal signals in real-world scenarios, MERC has attracted growing attention, where effective modality fusion becomes a key challenge Shi & Huang (2023). Existing approaches can be broadly divided into simple aggregation-based methods Hazarika et al. (2018); Lian et al. (2021) and more advanced graph-based fusion strategies Chen et al. (2023); Nguyen et al. (2023); Tu et al. (2024). Recently, Large Language Models (LLMs) have been introduced into ERC and have demonstrated strong performance Lei et al. (2023); Xue et al. (2024). To better adapt to multimodal scenarios, Jing et al. Jing et al. (2026) incorporate multimodal fusion and contextual modeling into LLMs.

Multimodal Emotion–Cause Pair Extraction in Conversations (MECPEC). Emotion–cause pair extraction (ECPE) was first formally introduced by Xia and Ding Xia & Ding (2019), which formulates the task as joint emotion detection and cause identification. Early approaches explored pipeline and end-to-end paradigms, including sliding-window methods Ding et al. (2020), pairwise classification Wu et al. (2023), and unified labeling frameworks Fan et al. (2020). More recently, MECPEC extends the task to more realistic conversational settings, where graph-based reasoning Liang et al. (2025); Ju et al. (2025) and emotion transition modeling Li et al. (2025) are commonly explored. However, most existing methods still rely on pre-defined segmentation and fixed interaction units, which limits their ability to model continuous conversational emotion–cause evolution.

Despite recent progress, existing approaches rely on pre-segmented inputs and speaker-centric assumptions, limiting their applicability to real-world continuous, multi-party interactions. They focus primarily on active speakers while overlooking other participants who may exhibit emotionally relevant behaviors. To address these limitations, we propose **ICET**, a task for identity-centric emotion tracking in continuous video. ICET operates on unsegmented inputs and requires joint reasoning over temporal dynamics and cross-modal interactions without predefined segmentation. We further introduce **EmoTrack**, a large-scale bilingual dataset to benchmark this task.

B Annotation Procedure

We adopt a systematic four-step annotation protocol to ensure consistency and reliability. Importantly, all annotations are performed on pre-segmented utterance units, each corresponding to a speaker turn. These units are later recombined into continuous video streams for the final unsegmented benchmark. For each utterance, annotators label the emotion and its cause for the speaker, as well as for non-speakers who exhibit emotional reactions. Example annotations for a Chinese sample are shown in Fig. 7 and for an English sample in Fig. 8.

(1) Before the annotation process began, all annotators (with backgrounds in affective computing) received training on the annotation guidelines and tools. They then reviewed the entire conversation video along with the transcript to obtain a comprehensive understanding of the multimodal context.

(2) In the English subset, annotators labeled emotions and causes for non-speakers only, as speaker emotions and causes were already provided by the ECF dataset Wang et al. (2023). In the Chinese subset, both speaker and non-speaker emotions and causes were fully annotated. Following prior



Figure 7: Example annotation from the Chinese subset of EmoTrack. Both speaker and non-speaker emotions and causes are labeled. Each cause refers to an explicitly expressed event or argument semantically linked to the corresponding emotion. A separate "reason" button is used for uncertain cases (e.g., target in profile or partially occluded), but both are treated as causal annotations.

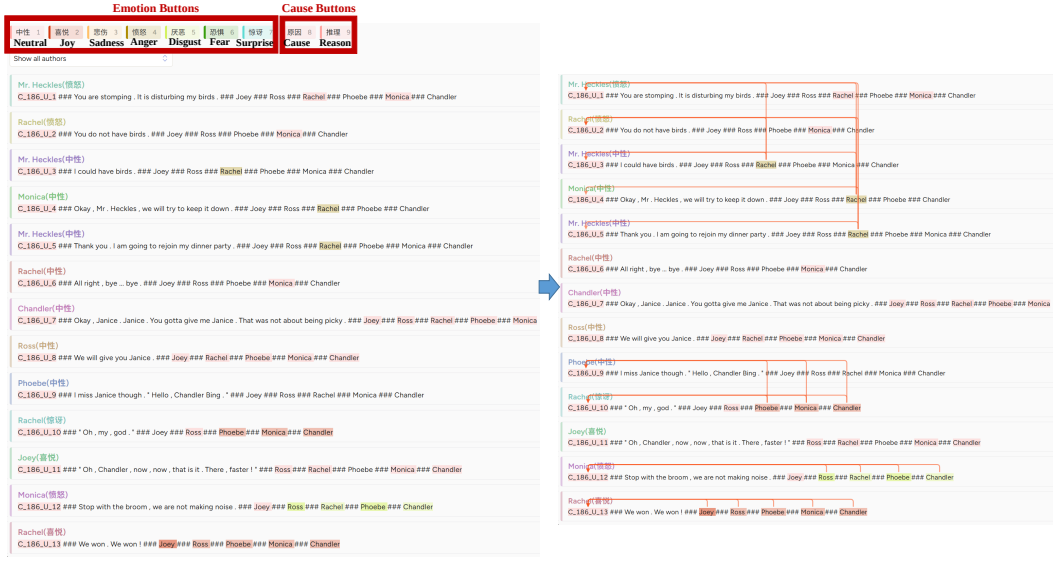


Figure 8: Example annotation from the English subset of EmoTrack. Non-speaker emotions and causes are annotated, while speaker emotions and causes are retained from the original dataset.

definitions Chen et al. (2010), a cause refers to an explicitly expressed event or argument that is semantically linked to the corresponding emotion.

(3) Annotators used a custom annotation interface based on Label Studio[1], an open-source data labeling platform, that displayed the text of each pre-segmented utterance; the corresponding video

[1] https://labelstud.io/



Figure 9: Annotation interface for EmoTrack. The left panel shows the customized CSC code interface; the middle panel illustrates emotion annotation for non-speakers using labeled buttons; the right panel demonstrates cause annotation using Label Studio's relation function.

C_327_U_7: Emotion Annotations


```
{
    "value": {
      "start": "6",
      "end": "6",
      "startOffset": 26,
      "endOffset": 29,
      "text": "吕轻侯",
      "paragraphlabels": [
        "厌恶"
      ]
    },
    "id": "ZaR36wJxV6",
    "from_name": "entity_slot",
    "to_name": "dialogue",
    "type": "paragraphlabels",
    "origin": "manual"
  }
```


C_327_U_7: Cause Annotations


```
{
    "value": {
      "start": "6",
      "end": "6",
      "startOffset": 0,
      "endOffset": 9,
      "text": "C_327_U_7",
      "paragraphlabels": [
        "原因"
      ]
    },
    "id": "omw646utXn",
    "from_name": "entity_slot",
    "to_name": "dialogue",
    "type": "paragraphlabels",
    "origin": "manual"
  }
```

```
{
    "from_id": "ZaR36wJxV6",
    "to_id": "omw646utXn",
    "type": "relation",
    "direction": "right"
  }
```


Figure 10: Example of JSON export from the EmoTrack annotation tool. Left: emotion labels assigned to non-speakers. Right: cause annotations linking emotions to utterance IDs.

segment was viewed simultaneously via a synchronized external player to capture audiovisual cues. Annotators could rewatch the conversation to review and revise initial annotations.

(4) Each utterance unit was annotated by three annotators. Final labels were obtained via majority voting, where an annotation was accepted if at least two annotators agreed. In cases of unresolved disagreement, an expert adjudicator provided the final decision. Inter-annotator agreement was computed after every 1,000 utterances to ensure consistency throughout the annotation process.

C Annotation Tools

We customized the *Slot Filling and Intent Classification* template from Label Studio, an open-source data labeling platform, to support our annotation task. As shown in Fig. 9, the interface presents the full conversation text, formatted as: **ID###utterance###speaker###non-speaker1###non-speaker2###...**. Since Label Studio provides only textual information, annotators simultaneously viewed the corresponding video via an external synchronized player to capture audiovisual cues.

Emotion Annotation. Each utterance line lists the speaker and the associated non-speakers. For each target person, annotators select one of seven emotion categories via Buttons 1–7 and assign it by double-clicking the corresponding name, as illustrated in the middle panel of Fig. 9. Multiple

target persons can be annotated for the same utterance, and each annotation contains a single emotion label. A JSON export of the emotion annotations is shown in Fig. 10.

Cause Annotation. To annotate causes, we introduced an additional button (Button 8), which defaults to selecting the utterance ID at the beginning of each line. Annotators then link the emotion-labeled target (e.g., speaker or non-speaker) to the utterance ID using Label Studio's built-in relation annotation function, as shown on the right panel of Fig. 9. For uncertain cases where the target is in profile or partially occluded but their emotion and cause can still be inferred, annotators use the reason label (Button 9) by selecting it and double-clicking the corresponding name. Multiple labels per person are supported. The final annotation is exported in JSON format, as shown in Fig. 10.

D Details about Comparison Methods

To comprehensively evaluate the EmoTrack benchmark, we select a range of state-of-the-art MLLMs spanning different architectural designs and parameter scales.

Closed-source Models. We include leading proprietary models such as **GPT-4o** OpenAI et al. (2024), **GPT-5.2** Singh et al. (2026), **Gemini 2.5 Pro** Comanici et al. (2025), **Gemini 3.1 Pro** Comanici et al. (2025), **Seed 2.0 Pro** Seedance et al. (2026), and **Qwen3.8 Omni Flash** Qwen Team (2026) (accessed via API). These models represent current frontiers in multimodal reasoning and video understanding.

Open-source Models. (1) **Qwen Series**: We evaluate **Qwen2.5-Omni-7B** Xu et al. (2025), **Qwen2.5-VL-7B** Bai et al. (2025), and **Qwen2.5-VL-32B** Bai et al. (2025). Unlike traditional vision-language models, Qwen2.5-Omni-7B is a unified model designed for interleaved audio-visual-text processing, making it uniquely suited for continuous conversational analysis. (2) **InternVL Series**: We include **InternVL3-8B** Zhu et al. (2025a) and **InternVL3-38B** Zhu et al. (2025a) to test models with strong vision-language alignment. (3) **Emotion-LLaMA-7B** Cheng et al. (2024): A specialized baseline adapted for affective computing, used to verify whether task-specific tuning generalizes to our continuous, identity-centric setting.

E Implementation Details

In this section, we outline the training and inference configurations for all evaluated models. All experiments are conducted on 4 NVIDIA RTX A6000 GPUs. For each model, the fine-tuning hyperparameters and inference configurations follow their official repository guidelines. All results are the mean over multiple runs.

Open-source Models. (1) **Qwen2.5-Omni-7B (SFT)**: We employ Low-Rank Adaptation (LoRA) Hu et al. (2022) for parameter-efficient fine-tuning. The visual encoder, audio encoder, and cross-modal projectors remain frozen, and LoRA adapters ($r = 8, \alpha = 32$) are applied only to the linear layers of the language model. The model is trained for 6 epochs with a global batch size of 16 (using gradient accumulation). We use a peak learning rate of 1×10^{-4} with a 0.05 warmup ratio. To process long videos within the 16,384-token context, we sample frames at 1 FPS, capped at 75 frames per video. Spatial resolution is dynamically bounded to a maximum of 25,088 pixels for video frames, while static images are constrained between 3,136 and 602,112 pixels. **Qwen2.5-Omni-7B (GRPO)**: We optimize the SFT policy using GRPO on the validation set. For each prompt, the policy samples $G = 4$ completions (temperature 1.0, nucleus sampling $p = 0.95$). Rewards are computed by comparing parsed outputs to the reference. The policy and frozen reference model are initialized from the same supervised LoRA checkpoint, with Kullback–Leibler (KL) regularization Kullback & Leibler (1951) ($\beta = 0.04$) and group-wise advantage normalization.

(2) **Qwen2.5-VL-7B and Qwen2.5-VL-32B**: The visual encoder and multimodal merger are frozen, while LoRA ($r = 64, \alpha = 128$, dropout 0.05) is injected into the self-attention projections of the language model. We train for 6 epochs with a global batch size of 4 using `bf16` mixed precision. Optimization uses a cosine schedule with peak learning rate 1×10^{-4}, 3% warmup, zero weight decay, and gradient clipping of 1.0. Given the 8,192-token context, we sample at 1 FPS, capped at 16 frames per video. Frame resolution is bounded to a maximum of 602,112 pixels, and static images between 784 and 50,176 pixels. These settings follow the official Qwen2.5-VL repository.

(3) **InternVL3-8B and InternVL3-38B**: Following official recommendations, we fine-tune the InternVL3 multimodal chat model by applying LoRA ($r = 16, \alpha = 32$) to the attention and feed-forward layers of the language backbone. The vision encoder and image-language projector remain frozen. The model is optimized using AdamW (learning rate 4×10^{-5}, weight decay 0.05) with a cosine schedule and 3% warmup. Training spans 6 epochs with a global batch size of 4. To fit within the 8,192 maximum sequence length, video clips are uniformly subsampled to a fixed budget of 8 frames. For static images, we employ dynamic high-resolution tiling coupled with a global thumbnail, preserving fine-grained spatial details while strictly bounding the token patch count.

(4) **Emotion-LLaMa-7B**: We adopt the original hyperparameters from Cheng et al. (2024). The EVA ViT-G vision encoder is frozen, and LoRA ($r = 64, \alpha = 16$) is applied to the attention modules of the language model. Training is performed over 6 epochs with a global batch size of 4. We apply a linear warmup and cosine decay schedule (decaying from 1×10^{-5} to 1×10^{-6}) with a 0.05 weight decay. Images are processed at 448×448 resolution. To comply with the strict 1,024-token supervised text limit, we extract the first frame for video inputs rather than utilizing dense sampling. Auxiliary audio features are extracted at 16 kHz mono, encoded via a frozen Chinese HuBERT-large model Hsu et al. (2021), and mean-pooled over time into a 1,024-dimensional embedding. Frozen face-MAE and VideoMAE features of the face crops are concatenated with this embedding.

Closed-Source Models. We implement two distinct inference pathways for proprietary APIs. The ***sampling-based workflow*** (used for GPT-4o and GPT-5.2) decodes video clips into a maximum of 20 frames (sampled at 1 FPS) to satisfy API payload constraints. A reference face image is required and is placed before these frames. Requests are evaluated at a temperature of $\tau = 0$. The ***native video workflow*** (used for Gemini 2.5/3.1 Pro, Qwen3.8 Omni Flash, and Seed 2.0 Pro) transmits the full video directly, prepends the reference image, and delegates internal frame sampling to the provider's backend. Both pathways are standardized to ensure fair and stable evaluations across different model endpoints.

F Prompt for Methods

We design a unified prompt template for multimodal emotion tracking, as illustrated in Fig. 11. Each input sample contains a reference face image specifying the target person and an unsegmented conversational video. The prompt requires the model to process the video chronologically and generate one structured line for each utterance. If the target person appears in an utterance, the model predicts both the emotion label and the corresponding causal utterance indices; if the target person is absent, it outputs a silence marker. The same output schema is used for both English and Chinese prompts, enabling consistent supervision and evaluation across languages.

G Inter-Emotion Confusion Analysis

As shown in Fig. 12, confusion matrices reveal that Neutral serves as a default "catch-all" category, absorbing most misclassifications, especially from Joy, Anger, and Surprise in English, and from Joy and Disgust in Chinese. Low-frequency emotions such as Disgust and Fear achieve under 10% accuracy in both languages, with errors predominantly falling into Neutral. Joy and Sad show moderate performance (39–51% and 28–31%, respectively), while Surprise is rarely predicted in Chinese, indicating a strong prediction bias. These results underscore that data imbalance and the dominance of Neutral significantly hinder fine-grained emotion recognition.

H Impact of Sampling Frames

As shown in Table 6, we investigate the effect of increasing the video frame sampling rate from 1 FPS to 2 FPS. Results show that higher temporal resolution generally improves performance across all three tasks. For USeg, the gains are more pronounced on Chinese, where mIoU increases from 55.43 to 58.04, while English shows only modest improvements (mIoU 29.76 $\rightarrow$ 30.21). EC and ECE also improve moderately in both languages. This suggests that additional visual frames provide useful information for emotion tracking, though the gains remain modest overall.

English Prompt template

[SYSTEM] You are an emotion tracking assistant. You will be given a reference image of a target person and a video of a multi-turn conversation. Track the target person's state utterance by utterance in chronological order. For each utterance where the target person appears, output exactly one line as '[Utterance N] [MM:SS] Emotion: {emotion}; Causes: {id1,id2,... or none}.' For each utterance where the target person is absent, output exactly one line as '[Utterance N] [MM:SS] <silence>'. Use 'Emotion: neutral; Causes: none.' when the target person is present but neutral. Do not output explanations or any extra text.

[USER] **<img>**\n **<video>**\n Target Person: {name}\n Please analyze the target person utterance by utterance in chronological order. If the target person appears, output 'Emotion: ...; Causes: ...'. If the target person is absent, output '<silence>'. Use 'Emotion: neutral; Causes: none.' when the target person is present but neutral.

[An example assistant response] [Utterance 1] [0:03] Emotion: joy; Causes: 1,3. \n[Utterance 2] [0:06] Emotion: neutral; Causes: none. \n [Utterance 3] [0:09] <silence>.

Chinese Prompt template

[SYSTEM] 你是一个情感追踪助手。你将收到目标人物的参考头像和一段多轮对话视频。请按时间顺序、逐条话语跟踪目标人物的状态。若目标人物在该句出现，严格输出一行 「'[话语 N] [MM:SS] 情感：{emotion}；原因句编号：{id1,id2,... 或 none}。」若目标人物在该句完全缺席，严格输出一行「[话语 N] [MM:SS] <silence>。」；若目标人物出现但情感为中性，输出「情感：neutral；原因句编号：none」。不要输出解释、推理或其他额外文本。

[USER] **<img>**\n **<video>**\n 目标人物：{name}\n 请按时间顺序逐条分析目标人物。若目标人物出现，输出「情感：...；原因句编号：...」；若目标人物完全缺席，输出「<silence>」；若目标人物出现但情感为中性，输出「情感：neutral；原因句编号：none」。

[An example assistant response] [话语 1] [0:03] 情感：joy；原因句编号：1,3。 \n[话语 2] [0:06] 情感：neutral；原因句编号：none。 \n[话语 3] [0:09] <silence>。"

Figure 11: Prompt template for ICET. For models that accept raw video input (e.g., Gemini series), the video is provided as a link. For models that do not (e.g., GPT series), the video is represented as a sequence of sampled frames. The prompt structure is unified across both input formats.

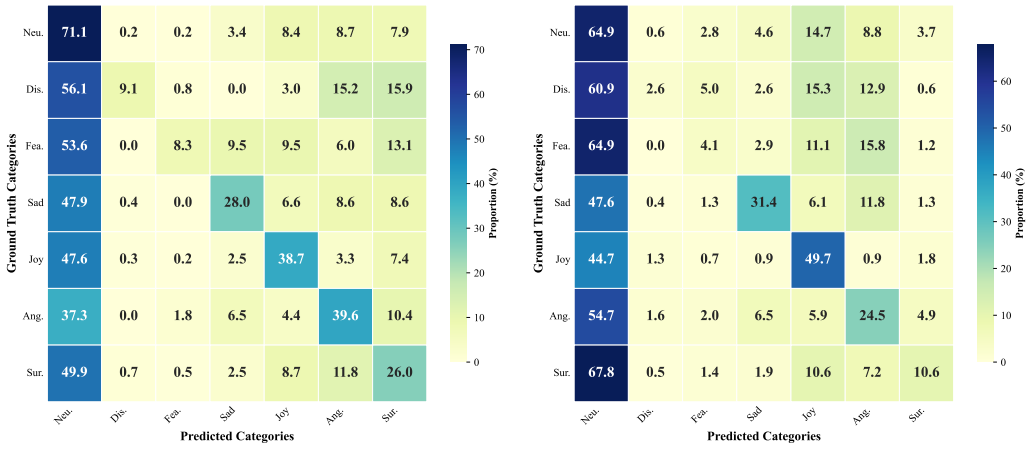


Figure 12: Row-normalized confusion matrices for emotion classification on EmoTrack (values in %, 0–100). Left: English subset; Right: Chinese subset. Each row sums to 100%.

I Impact of Streaming Inference

As shown in Table 6, we evaluate streaming inference, where the model uses only past and current utterances. This reflects real-time applications but removes future context, which is unavailable at inference. Results are mixed: streaming improves mIoU in both languages, but substantially

Table 6: Model performance (%) under different sampling frames and streaming inference.

Model	Utterance Segmentation				Emotion Classification		Emotion-Cause Extraction	
	IoU≥0.1	IoU≥0.3	IoU≥0.5	mIoU	Acc.	W-F1	Acc.	W-F1
English Subset								
Qwen2.5-Omni-7B	**47.37**	37.91	32.38	29.76	40.99	41.98	9.61	16.09
w/ Streaming	34.59	34.59	**34.59**	**34.58**	15.34	22.81	4.66	9.16
w/ FPS=2	47.15	**38.31**	32.75	30.21	**42.09**	**42.36**	**10.59**	**17.27**
Chinese Subset								
Qwen2.5-Omni-7B	76.20	72.16	63.04	55.43	36.84	33.29	3.22	11.30
w/ Streaming	65.10	65.10	65.10	**65.08**	25.66	27.98	**4.80**	**13.73**
w/ FPS=2	**76.92**	**73.40**	**66.22**	58.04	**37.20**	**34.96**	4.14	13.47

Note: FPS refers to frames per second. Streaming inference removes access to future utterances.

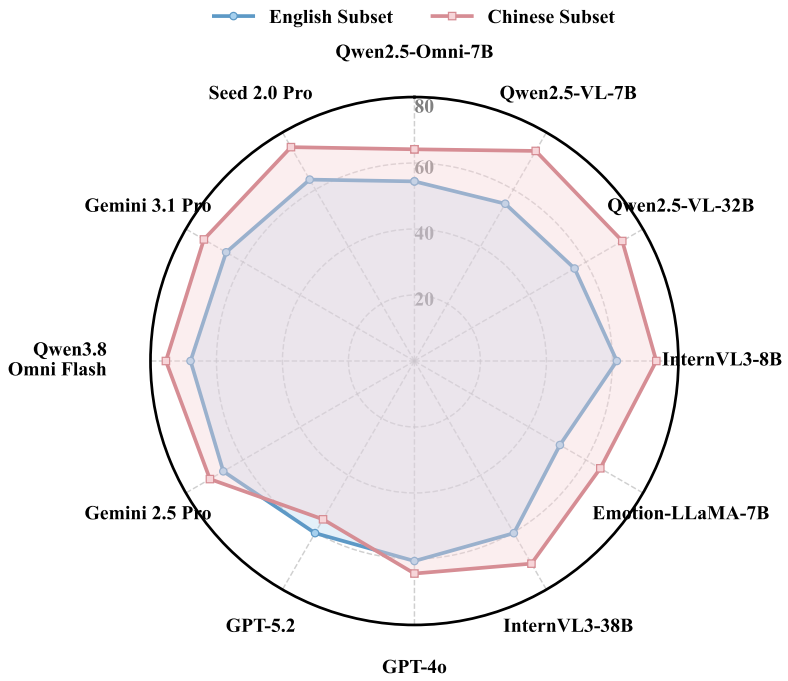


Figure 13: Radar chart of silence prediction weighted F1 scores across different MLLMs on Emo-Track. Blue: English subset; Red: Chinese subset.

degrades EC. For ECE, streaming improves performance on Chinese but degrades it on English. A plausible explanation is that future causes are harder to identify than past ones, and this difficulty may be more pronounced in Chinese, making the removal of future context less harmful and even helpful on Chinese, while hurting English. These results reveal a trade-off between USeg and EC in real-time settings, with ECE showing language-dependent effects.

J Analysis of Silence Prediction

We analyze silence prediction as a proxy for evaluating whether models know when to refrain from outputting an emotion for the target individual. As shown in Fig. 13, the weighted F1 scores reveal two main patterns. First, most models perform better on Chinese than English, indicating that silence recognition is language-dependent. Second, the best English performance is achieved by Qwen3.8 Omni Flash, Gemini 2.5 Pro, Gemini 3.1 Pro, and Seed 2.0 Pro. On Chinese, open-source models

such as InternVL3-8B and Qwen2.5-VL-7B approach the top tier but no longer surpass the best closed-source models. Notably, GPT-5.2 drops notably from English to Chinese and is the lowest on Chinese, suggesting less robustness to cross-linguistic differences. These results suggest that learning when to withhold predictions remains a nontrivial challenge.