EmoS: A High-Fidelity Multimodal Benchmark for Fine-grained Streaming Emotional Understanding

Pengze Guo[1*] **Jingxi Liang**[1*] **Zhiwen Xie**[1,2†] **Qifeng Wang**[1] **Derek F. Wong**[1†]
[1] NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
[2]School of Computer Science, Central China Normal University
nlp2ct.{pengze,jingxi,qifeng}@gmail.com, zwxie@ccnu.edu.cn
derekfw@um.edu.mo

Abstract

In the context of today's high-pressure, aging society, the demand for large-scale emotional models capable of providing empathetic support is more critical than ever. However, existing benchmarks fail to simultaneously achieve ecological validity, signal clarity, and reliable fine-grained labeling. We introduce EmoS, a high-fidelity bilingual benchmark designed to resolve the limitations of ecological validity and noise in existing datasets by combining strictly filtered static slices with a dynamic Streaming Monologue subset. Supported by a rigorous dual-layer human annotation pipeline, EmoS provides trusted ground truth that captures continuous emotional evolution. Empirical results show that fine-tuning MLLMs (multimodal large language models) on EmoS yields significant gains over zero-shot baselines, laying the foundation for the training and evaluation of future emotion recognition models and empathy models. The dataset and code are publicly available at https://github.com/NLP2CT/EmoS.

1 Introduction

With the rapid advancement of artificial intelligence (AI) , there is an increasing societal expectation for AI agents to transcend simple task execution and act as an emotionally intelligent partner (Cheng et al., 2024; Shi et al., 2025; Liu et al., 2022a). In high-pressure environments, such as psychological counseling and aging care, AI systems must not only recognize static emotions but also proactively intervene before a user's state deteriorates. However, the development of such empathetic models is currently bottlenecked by the quality of underlying data. We identify a critical "data quality trilemma" in Multimodal Emotion Recognition (MER), where existing benchmarks struggle

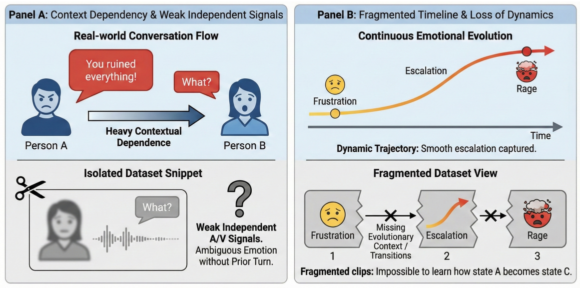

Figure 1: Illustration of limitations in current short-utterance multimodal datasets.

to simultaneously achieve ecological validity, signal clarity, and reliable labeling. Prior datasets often fail to meet these evolving needs. Early lab-controlled datasets like IEMOCAP (Busso et al., 2008) and DAIC-WOZ (Gratch et al., 2014), while clean, lack the spontaneity of real-life interactions. Conversely, Widely used 'in-the-wild' datasets like MELD (Poria et al., 2019) and CH-SIMS v2 (Liu et al., 2022a) provide rich contexts but are fundamentally compromised by unreliable and coarse-grained ground truth. Beyond simple modality noise, MELD suffers from heavy context dependence and ambiguity, while CH-SIMS v2 is limited to sentiment polarity rather than specific emotions, rendering both insufficient for fine-grained emotion recognition tasks. More critically, most datasets fragment continuous dialogues into isolated utterance-level snippets. This fragmentation severs the emotional evolution timeline, making it impossible for models to learn dynamic trajectories, such as the gradual escalation from frustration to rage. While recent synthetic datasets generated by Large Language Models (LLMs) (Cheng et al., 2024; Lian et al., 2025) attempt to scale up data, they often suffer from hallucinated labels, compromising the credibility of the ground truth.

The limitations of existing work underscore the critical necessity of constructing a high-quality

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39074–39089
July 2-7, 2026 ©2026 Association for Computational Linguistics

MER benchmark. An ideal MER benchmark requires high-fidelity multimodal alignment across textual, acoustic, and visual signals. Crucially, it must provide fine-grained emotion labels (e.g., joy, anger) rather than coarse sentiment polarity. Furthermore, it should support continuous temporal modeling over long-form dialogues, moving beyond isolated utterance-level classification. To achieve this, we introduce **EmoS**, a high-fidelity bilingual benchmark spanning 9,403 static samples and 2 hours of streaming monologues. Constructed through a rigorous dual-layer annotation pipeline (Basic-7 (Ekman, 1992) and GoEmotions-28 (Demszky et al., 2020)) validated by the Dawid-Skene algorithm (Dawid and Skene, 1979), EmoS integrates a strictly filtered MELD-Core, fine-grained CH-SIMS v2, and a novel streaming subset to capture continuous emotional evolution.

We benchmarked SOTA MLLMs (e.g., Gemini-3, Qwen-3) on EmoS. Results indicate that Zero-Shot performance is limited, with Gemini-3 achieving only $\sim$61% accuracy due to conservative neutral bias. Conversely, Task-Adaptive Fine-Tuning is essential: fine-tuning Qwen-3 boosted accuracy to 70.3%, dramatically improving recall on long-tail emotions like Disgust (F1 0.30$\rightarrow$0.75). Furthermore, models with ultra-long context windows demonstrated exceptional sensitivity to narrative flow, successfully predicting 82% of emotion turning points in the streaming subset. In summary, our contributions are threefold:

- We present EmoS, a strictly cleaned, human-annotated benchmark ($N = 9,403$ static samples + 2h streaming) that resolves modality noise and introduces a novel streaming subset.
- We establish a high-standard Dual-Layer Annotation Protocol combining basic and fine-grained taxonomies, supported by rigorous annotator style analysis.
- We provide a comprehensive Benchmark of SOTA MLLMs, showing that fine-tuning on high-quality data is prerequisite for mastering dynamic emotional trajectories.

2 Related Work

Research in Multimodal Emotion Recognition (MER) has evolved from laboratory settings to in-the-wild and synthetic data, yet a perfect benchmark remains elusive. Early lab-controlled datasets like IEMOCAP (Busso et al., 2008) offer clean signals but lack ecological validity due to scripted interactions (Dhall et al., 2013). To address this, in-the-wild datasets such as MELD (Poria et al., 2019) and CMU-MOSEI (Zadeh et al., 2018) were introduced; however, they are often plagued by severe modality noise (e.g., canned laughter, shot transitions) and fragmented timelines that disrupt the modeling of emotion dynamics. More recently, LLM-generated datasets (Cheng et al., 2024) have attempted to scale up annotation but frequently suffer from hallucinated labels and a lack of rigorous human verification (Ji et al., 2023).

Consequently, the field faces a "data quality trilemma": existing benchmarks are either ecologically invalid, excessively noisy, or unreliable. EmoS is designed to resolve this trilemma by strictly filtering for signal clarity and introducing human-verified streaming monologues. Due to space constraints, we provide a comprehensive review of existing datasets and their specific limitations in Appendix H.

3 Dataset Construction

In the context of Multimodal Emotion Recognition (MER), the primary goal is to recognize and understand emotional expressions from multimodal data, such as speech, facial expressions, and text. Traditional MER tasks focus on recognizing emotions from short, static data slices (e.g., individual sentences or clips), without considering the temporal evolution of emotions. However, in "Streaming MER", the task evolves to incorporate temporal dynamics, where the model must recognize and track emotional transitions across continuous streams of data (e.g., monologues, conversations). To address the needs of the MER task, we proposed the EmoS dataset by referring to the labeling process shown in Figure 2, which illustrates the basic information and processing procedures of our dataset.

3.1 Data Collection

Our dataset, EmoS, consists of three independent subsets. It integrates carefully selected segments from MELD (English) and CH-SIMS v2 (Chinese), and is further complemented by our newly collected long-form monologue subset. The original MELD dataset (Poria et al., 2019) is derived from the sitcom Friends. It contains a large number of very short utterances (e.g., "yes", "okay") and is heavily contaminated by canned laughter. While MELD is suitable for modeling conversational context, these characteristics introduce sub-

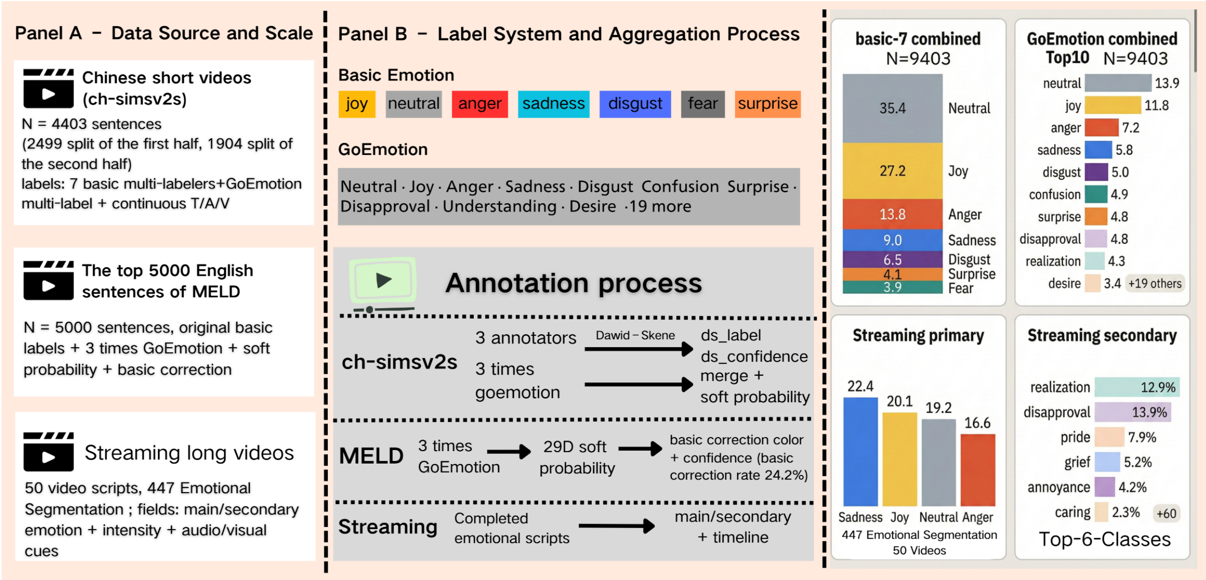

Figure 2: The basic information and processing procedures of our dataset

stantial noise for fine-grained emotion recognition, as such utterances often lack standalone acoustic or visual cues. To address this issue, we perform strict filtering on MELD by removing segments shorter than one second, as well as low signal-to-noise segments caused by laughter tracks or severe visual occlusions (e.g., shot transitions that fail to maintain focus on the speaker). This yields a core subset of 5,000 high-quality samples, termed **MELD-Core**. Meanwhile, we incorporate the full CH-SIMS v2 dataset (Liu et al., 2022a), which provides high-quality multimodal alignment and applies video cropping for a portion of the samples to capture the speaker's visual information. This ensures that the visual focus remains on the speaker's facial expressions, offering a robust benchmark for Chinese multimodal emotion analysis.

Existing datasets often rely on sentence-level slicing or short multi-party dialogue snippets, which breaks the temporal continuity of emotional evolution. To capture dynamics such as buildup, transition, and climax, we collect 50 continuous monologues (about 2 hours total) from movies and TV dramas (e.g., The Legend of 1900 and Hei Bing). We then parse them into roughly 700 consecutive sentences, preserving the temporal coherence needed to model emotion dynamics.

3.2 Data Annotation and Quality Analysis

Given the limitations of current Multimodal LLMs in zero-shot emotion recognition (we tested Gemini-3 and Qwen-3-omni-flash on a pilot set, and both achieved below 70% accuracy on the Basic-7 classification task), we adopt a strict human annotation pipeline. This section details our two-level taxonomy, quality control mechanisms, and modeling of annotator subjectivity.

3.2.1 Annotation Protocol and Taxonomy

To balance standardization and semantic richness, our annotation framework operates at two granularities:

Basic-7 (Discrete Categories). We follow the classic Ekman-style seven-way taxonomy (Anger, Joy, Sadness, Fear, Disgust, Surprise, Neutral). This scheme is widely used in multimodal research and provides a robust benchmark for evaluating basic discriminative capability (Ekman, 1992).

GoEmotions-28 (Fine-grained Multi-label). To capture subtle affective distinctions (e.g., admiration, remorse, confusion), we adopt the 28-category fine-grained taxonomy from GoEmotions (Demszky et al., 2020). This enables further assessment of a model's capability in handling semantic proximity and complex emotion understanding.

Each sample is annotated by three independent annotators. For the CH-SIMS v2 and Streaming subsets, annotators provide both a single Basic-7 label and multi-label GoEmotions annotations. For MELD, we integrate the original Basic labels, additionally collect GoEmotions multi-label annotations, and perform cross-taxonomy consistency checks.

3.2.2 Label Aggregation and Quality Control

To reduce individual bias and estimate the reliability of Ground Truth (Whitehill et al., 2009; Raykar et al., 2010), we implement a multi-stage aggregation and cleaning pipeline.

Basic-7 Aggregation (Dawid–Skene). For the single-label task, we use the Dawid–Skene (DS) (Dawid and Skene, 1979) algorithm to estimate the inferred label y_{ds} and its posterior confidence c_{ds}.

Specifically, the DS algorithm employs an Expectation-Maximization (EM) framework to jointly estimate the latent class priors and the confusion matrix (reliability) of each annotator. Upon convergence, the algorithm yields a posterior probability distribution over the class set K for each sample i. The confidence score $c_{ds}^{(i)}$ is defined as the maximum value of this posterior distribution:

$$c_{ds}^{(i)} = \max_{k \in K} P(y_i = k \mid x_i^{(1)}, x_i^{(2)}, x_i^{(3)}) \quad (1)$$

where $x_i^{(m)}$ denotes the label provided by the m-th annotator. This metric reflects the model's certainty in the inferred label after weighting the reliability of different annotators. Based on c_{ds} and annotator agreement, we divide the dataset into three quality tiers (statistics are shown in Table 1):

- **High-Quality (76.6%):** all three annotators agree (unique_labels=1) or $c_{ds} \geq 0.9$.
- **Medium-Quality (14.0%):** $0.8 \leq c_{ds} < 0.9$, typically involving mildly ambiguous boundaries (e.g., Sadness vs. Neutral).
- **Low-Quality / Hard (9.4%):** $c_{ds} < 0.8$, often where all three annotators assign different labels. These samples reflect inherent subjectivity and uncertainty in human emotion perception.

GoEmotions Soft-Label Modeling. Given the polysemy of fine-grained emotions, we avoid hard-label voting and instead model Ground Truth as a soft probability distribution. For each class k among $K = 29$ categories, we compute

$$p_k = \frac{n_k}{3}, \quad p_k \in \{0, 0.33, 0.66, 1.0\},$$

where n_k is the number of annotators who select class k. This probabilistic representation preserves annotator disagreement and supports label distribution learning.

Metric	First 2,499	Latter 1,904	Total
N	2,499	1,904	4,403
High (≥0.9)	1,702	1,670	3,372
Medium (0.8–0.9)	496	122	618
Low (<0.8)	301	112	413
unique = 1	1,097	904	2,001
unique = 2	1,119	917	2,036
unique = 3	283	83	366

Table 1: CH-SIMS v2 basic-7 classes quality tiers and unique-label counts.

Metric	First 2,499	Latter 1,904
N	2,499	1,904
Annotator label counts	1.84 / 1.42 / 1.41	1.93 / 1.25 / 1.51
Union mean	3.68	3.60
Union ≥ 4	1,458	1,521
Exact match (%)	7.6%	24.4%

Table 2: CH-SIMS v2 GoEmotions multi-label statistics.

Basic Label Correction on MELD. To account for label noise in MELD, we re-aggregate Basic-7 labels and measure the correction confidence and change rate. Furthermore, we conducted a manual audit of the corrected labels in MELD. We observed that in addition to corrections arising from subjective annotator disagreements, a small fraction of the revisions addressed inherent annotation errors in the original dataset. Specifically, due to the massive scale of MELD, certain clips were duplicated but assigned inconsistent emotion labels. These discrepancies have been fully rectified in our processed subset. Table 3 summarizes the correction statistics and the corresponding GoEmotions multi-label properties.

Streaming Monologue Subset Annotation. The annotation of the Streaming Monologue subset follows the same core multi-annotator pipeline as the

Metric	Value
Basic corrected confidence ≥ 0.7	1,993
Basic corrected confidence 0.5–0.7	2,101
Basic corrected confidence < 0.5	401
Basic labels changed (basic_changed)	1,031 (24.2%)
GoEmotions union mean	2.70
Union ≥ 4	979
Union ≥ 5	502
Exact match (3 annotators)	686

Table 3: MELD-Core basic correction and multi-label statistics (Excluding "unsure").

Dataset / Split	Annotator	k_a	Dominant emotion tendencies	Pairwise Jaccard	Exact match
CH-SIMS v2 First 2,499	A1	1.84	Joy, Disgust, Sadness, Anger, Desire, Realization (fine-grained)	0.34 – 0.38	∼8%
	A2	1.42	Neutral / Unsure dominant (conservative)		
	A3	1.41	Disapproval, Confusion, Caring (cognitive/social)		
CH-SIMS v2 Latter 1904	A4	∼1.93	Joy, Optimism, Caring, Excitement + Disapproval/Annoyance (expansive)	0.41 – 0.45	∼24%
	A5	∼1.25	Near-single-label; Joy / Neutral / Anger / Sadness		
	A6	∼1.51	Joy, Optimism, Anger, Disappointment (bi-label)		
MELD First 3,000	A7	1.61	Joy, Realization, Neutral + Anger / Disgust	0.389 – 0.445	∼31%
	A8	∼1.40	Neutral dominant (conservative)		
	A9	∼1.40	Surprise, Confusion, Realization, Disapproval (Neutral ∼25%)		
MELD Latter 2,000	A10	2.31	Joy / Neutral + Anger / Disgust (most expansive)	0.369 – 0.401	∼9%
	A11	∼1.50	Neutral dominant		
	A12	∼1.50	Joy vs. Disapproval (Neutral ∼25%)		

Table 4: Annotator style summary on CH-SIMS v2 and MELD. We report per-annotator label cardinality (k_a) and dominant label usage. Split-level agreement is measured by pairwise Jaccard (range over three annotator pairs) and exact match (Representing the proportion agreed upon by the three annotators)

static datasets, but it necessitates a more rigorous temporal segmentation protocol. We perform a coarse temporal segmentation based on pauses occurring in the video. Subsequently, the three annotators discuss and annotate through a meeting-based approach to collectively determine the precise segment boundaries and the overarching direction of emotion evolution, retaining only high-confidence samples. To enrich the contextual metadata of this subset, we leverage Large Language Models (LLMs) as an auxiliary tool to draft initial descriptions for emotion causes, which are subsequently verified and corrected as necessary.

3.2.3 Annotator Modeling and Stylistic Analysis

A key finding is that annotator styles significantly affect annotation consistency (Fleiss, 1971; Krippendorff, 2011). From GoEmotions annotation statistics (Table 4), we observe that cross-subset differences in agreement are not solely driven by data noise, but also by differences in annotation style. We characterize annotator behavior using average label cardinality (Label Cardinality, k_a) and semantic preference, and summarize four typical styles, as shown in (Figure. 3) :

Expansive Style. k_a is significantly above the group median. Such annotators tend to decompose complex emotions into multiple components (e.g., decomposing "remorse" into Sadness + Disappointment + Remorse), leading to a large label union per sample (Union ≥ 4).

Parsimonious Style. k_a is below the median. They focus on the dominant emotion, often yielding higher Exact Match agreement, but potentially missing secondary affective nuances.

Safety-Seeking / Neutral-Fallback. These annotators assign Neutral or Unsure at a notably higher rate, reflecting a conservative strategy under ambiguity to reduce the risk of mistakes.

Socio-Cognitive Sensitive. Such annotators are more sensitive to cognitive labels (e.g., Confusion, Realization) and social-evaluative labels (e.g., Caring, Disapproval). This pattern is especially evident in dialogue data (MELD), suggesting deeper attention to interpersonal cues.

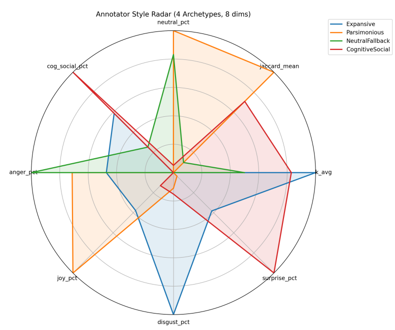

Figure 3: The performance of four annotator styles across eight representative dimensions.

This typology does not assume any style is "more correct"; rather, it explains agreement differences. When an annotation group mixes expansive and conservative styles, the label union and

Emotion	CH-SIMS v2	MELD-Core	Combined
Neutral	27.9%	46.4%	35.4%
Joy	31.2%	21.4%	27.2%
Anger	14.0%	13.4%	13.8%
Sadness	10.4%	6.9%	9.0%
Disgust	8.8%	3.1%	6.5%
Surprise	3.3%	5.4%	4.1%
Fear	4.3%	3.4%	3.9%
Total	4,403	5,000	9,403

Table 5: Distribution of Basic Emotions across dataset subsets (percentages). Percentages are computed over **non-Unsure** samples; the total row reports the **total number of samples** (including Unsure). We additionally introduce an Unsure category for both the Basic and GoEmotions taxonomies to allow annotators to abstain under genuine ambiguity; Unsure labels are retained in the raw annotation records for potential future re-auditing.

disagreement (low Exact Match, low Jaccard) increase substantially. In contrast, when conservative annotators share similar decision rules, pairwise agreement and three-way agreement can remain high even with one expansive annotator present. Overall, this suggests that different populations exhibit different emotion perception tendencies; we argue that a general-purpose model should be evaluated and trained on the most broadly representative subset with maximal coverage and generality.

3.3 Statistical Analysis

3.3.1 Dataset Statistics and Distributions

The final dataset contains 9,403 samples and additional streaming data. We adopt the Dawid–Skene aggregation method, along with additional post-processing procedures, to ensure label reliability; details are provided in a later section.

Basic Emotion Distribution. As shown in Table 5, Neutral (35.4%) and Joy (27.2%) are relatively balanced in the combined dataset. Notably, MELD exhibits a strong neutral bias (46.4%), reflecting the prevalence of functional utterances in everyday conversations, while CH-SIMS v2 contributes a substantial portion of positive emotion samples. Negative emotions (Anger, Sadness) remain stably represented in both languages.

Fine-grained (GoEmotions) Analysis. The label distribution follows a long-tail pattern. In the Chinese subset, cognition-related labels such as "Disapproval" and "Optimism" are more prominent, whereas the English subset tends to favor labels like "Surprise" and "Realization", indicating a higher reliance on contextual inference. The dataset exhibits a pronounced long-tail distribution: labels such as Grief, Relief, and Embarrassment are extremely rare, with Grief accounting for only(0.1%) in MELD. In addition, complex social emotions like Remorse and Gratitude generally occur at frequencies below 1.0%, Due to space limitations, detailed data tables can be found in the Appendix C.

Dynamics in Streaming Data. The streaming subset is characterized by high emotional volatility. Across the 50 clips, we identify 401 emotion turning points, averaging 8.0 transitions per clip (see Table 6). Overall, the two emotion taxonomies exhibit highly consistent yet hierarchically differentiated structural patterns in the streaming subset. Neutral emotion dominates in both schemes (Basic-7: 32.8%; GoEmotion: 31.6%), indicating that emotional states in continuous contexts tend to fluctuate around an affective "baseline" rather than remaining in highly activated states, which provides a buffer for frequent emotional shifts. Meanwhile, negative emotions are both prevalent and highly differentiated: in Basic-7, they are mainly reflected by Anger, Disgust, and Sadness, whereas GoEmotion further decomposes similar negative experiences into multiple high-frequency categories such as anger, disgust, disapproval, annoyance, and confusion. This coarse-to-fine correspondence suggests that negative affect in streaming contexts is not characterized by isolated emotional outbursts, but by the alternation of semantically related yet distinguishable emotional states. In contrast, positive emotion (Joy) maintains a moderately high but non-dominant proportion in both taxonomies (Basic-7: 19.6%; GoEmotion: 23.7%), implying that positive experiences tend to appear as brief insertions rather than sustained states. Collectively, these patterns form a dynamic emotional landscape characterized by a neutral background and high-frequency transitions between positive and negative states. Due to the limitation of space, we have placed more statistical results in the Appendix C.

4 Experiments

4.1 Basic Emotion Recognition Capability

To systematically evaluate the emotion recognition capabilities of current multimodal large language

Metric	Value
Number of Segments	50
Total Turning Points	401
Avg. Turning Points / Segment	8.0
Max Turning Points (Sample)	25 (Any Given Sunday)

Table 6: Statistics of the **Streaming Subset**. High frequency of turning points indicates strong emotional dynamics.

models (MLLMs), we compared a set of representative multimodal models. The closed-source model is Gemini-3 (Gemini Team, 2023; Google, 2025). The open-source baselines include Qwen-2.5-Omni-7B (hereafter abbreviated as Qwen-2.5), Qwen3-omni-flash (denoted as Qwen-3 (Xu et al., 2025a,b) in Table 7), and EmotionLLaMA (Cheng et al., 2024). In addition, we report a task-adapted fine-tuned variant of Qwen-3, namely **Qwen-3 (FT)**, to quantify the gain brought by task adaptation (we attempted to include AffectGPT (Lian et al., 2025), but its official implementation could not be reproduced on local devices due to environment issues). Due to space constraints, the specific settings for finetuning are presented in Appendix B.

Except for Qwen-3 (FT), all models were evaluated under the zero-shot setting. During prompt engineering, we observed that some models could not reliably follow the instruction of 28-class GoEmotions, often collapsing to 7 classes or producing invalid labels. Therefore, to ensure cross-model comparability, we adopted the Basic-7 label space. The test set was further split into a high-confidence subset ($N = 800$) for reporting the main results, and a low-confidence subset ($N = 400$) for additional analyses such as model self-awareness and confidence calibration. Given the imbalanced class distribution (Neutral 37.1%, Joy 26.0%, while Surprise/Disgust/Fear are each below 6%), we report Accuracy, Macro-F1, and Weighted-F1. Results on the high-confidence subset are shown in Table 7. For model predictions that failed to follow the instruction and thus fell outside the predefined set, we uniformly counted them as incorrect samples.

Overall, models are more stable on frequent classes, while their performance diverges more substantially on minority classes. This is also reflected by the fact that most systems achieve higher Weighted-F1 than Macro-F1. Under the zero-shot evaluation, Gemini-3 achieves the best overall performance (Acc 0.611 / Macro-F1 0.528 / Weighted-F1 0.605). The Qwen family shows similar Accuracy and Weighted-F1 (e.g., Qwen-3: 0.580/0.582; Qwen-2.5: 0.571/0.549), but a lower Macro-F1, indicating weaker robustness on long-tail categories. EmotionLLaMA performs the worst overall (Acc 0.369 / Macro-F1 0.289 / Weighted-F1 0.368), suggesting limited reliability on general emotion classification tasks. Task-adaptive fine-tuning can substantially improve class-balanced performance. Qwen-3 (FT) increases Accuracy from 0.580 to 0.703 and Macro-F1 from 0.517 to 0.628 (Weighted-F1: 0.582→0.681), and yields notable gains on minority classes (e.g., the F1 score of Disgust reaches 0.757, whereas zero-shot Qwen-3 attains only 0.304). At the class level, Joy and Neutral are generally easier to recognize (e.g., Gemini-3 achieves Joy F1 of 0.735; Qwen-3 (FT) achieves Neutral F1 of 0.793), while low-frequency negative emotions remain the primary bottleneck.

We identify two typical failure modes: (1) conservative prediction, where minority classes exhibit very high precision but extremely low recall (e.g., Qwen-2.5 on Fear with P=1.000 and R=0.026; on Surprise with P=1.000 and R=0.130); and (2) over-prediction, where recall is improved at the expense of precision, leading to unstable outputs (e.g., EmotionLLaMA fails completely on Disgust and shows a “high-recall, low-precision” pattern on Sadness). In sum, Table 7 indicates that zero-shot emotion recognition in current MLLMs is mainly constrained by instruction-following stability and long-tail emotion recognition capability, while fine-tuning provides a direct path toward more reliable and more class-balanced emotion understanding.

4.2 Fine-grained Emotion Recognition Capability (GoEmotions)

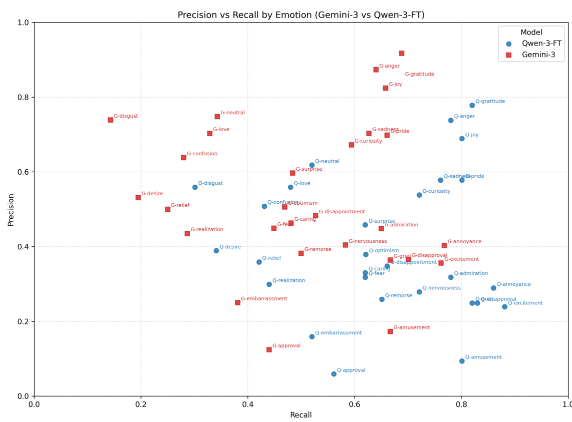

Figure 4: Precision vs. Recall per emotion label. The comparison highlights distinct preference behaviors

Metric	Gemini-3	Qwen-3	Qwen-3 (FT)	Qwen-2.5	EmotionLLaMA
Accuracy	0.611	0.580	**0.703**	0.571	0.369
Macro-F1	0.528	0.517	**0.628**	0.419	0.289
Weighted-F1	0.605	0.582	**0.681**	0.549	0.368
Anger (P)	0.669	**0.708**	0.670	0.648	0.512
Anger (R)	**0.770**	0.680	0.759	0.648	0.536
Anger (F1)	**0.716**	0.694	0.712	0.648	0.524
Disgust (P)	0.217	0.700	**0.819**	0.444	0.000
Disgust (R)	0.143	0.194	**0.704**	0.133	0.000
Disgust (F1)	0.172	0.304	**0.757**	0.205	0.000
Fear (P)	0.383	0.449	0.720	**1.000**	0.375
Fear (R)	**0.590**	0.550	0.571	0.026	0.050
Fear (F1)	0.465	0.494	**0.637**	0.051	0.088
Joy (P)	0.678	0.551	0.611	**0.836**	0.664
Joy (R)	**0.802**	0.787	0.533	0.298	0.393
Joy (F1)	**0.735**	0.648	0.569	0.439	0.494
Neutral (P)	**0.777**	0.756	0.760	0.584	0.517
Neutral (R)	0.422	0.429	**0.830**	0.865	0.322
Neutral (F1)	0.547	0.547	**0.793**	0.697	0.397
Sadness (P)	0.533	0.620	0.560	**0.689**	0.176
Sadness (R)	**0.721**	0.647	0.670	0.636	0.679
Sadness (F1)	0.613	0.633	0.610	**0.661**	0.280
Surprise (P)	0.309	0.209	0.209	**1.000**	0.157
Surprise (R)	**0.808**	0.519	0.650	0.130	0.531
Surprise (F1)	**0.447**	0.298	0.316	0.231	0.242

Table 7: Emotion classification results on 800 high-confidence samples. Class distribution (%): Anger 15.5, Disgust 4.4, Fear 5.1, Joy 26.0, Neutral 37.1, Sadness 8.7, Surprise 3.3.

We conducted a systematic comparison of the multi-label emotion classification performance of Gemini-3 and Qwen-3 (Fine-tuned). The detailed results are presented in the Appendix D. Qwen-3 (Fine-tuned) exhibited higher overall recall, which was particularly pronounced for fine-grained cognitive and social emotion labels; however, this improvement was accompanied by a systematic decline in precision, leading to an increase in false positives. In contrast, Gemini-3 maintained higher and more stable precision, adopting a relatively conservative prediction strategy, but its recall on long-tail labels remained insufficient. This divergence reflects different preferences in the precision–recall trade-off: Qwen-3 (Fine-tuned) tends to expand label coverage, whereas Gemini-3 places greater emphasis on prediction reliability and false-positive control, at the expense of broader label coverage.

4.3 Confidence Calibration and Robustness Analysis

To examine whether the models are able to "recognize their own uncertainty," we compare their output confidence scores on the high-confidence and low-confidence subsets. As shown in Table 8, Gemini 3 and Qwen 3 Flash exhibit a decrease in average confidence on low-confidence samples (from 0.817 to 0.804 and from 0.873 to 0.849, respectively), indicating that they can, to some extent, perceive input uncertainty. In contrast, Qwen2.5, constrained by its parameter scale, lacks the ability to effectively discriminate confidence levels, showing nearly identical scores across the two subsets (0.850 vs. 0.852). Qwen3 (FT), while achieving the highest overall confidence, also exhibits the largest confidence drop ($\Delta = -0.032$).

4.4 Evaluation on Streaming Data

To evaluate continuous dynamics, we designed a Streaming Utterance-Level Annotation task where models must process videos to summarize emotional trajectories and annotate the main speaker's utterances using our dual-layer taxonomy. Detailed contents can be referred to Appendix I. Given that Gemini-3 is currently the unique model capable of processing ultra-long multimodal context windows, our evaluation on the Streaming subset is exclusively conducted on this architecture. We observe that as the context length expands, the model's robust text understanding capabilities significantly enhance its grasp of the narrative flow and emotional progression. Gemini-3 demonstrates exceptional sensitivity to dynamic changes, achieving an accuracy of 82% in predicting emotion turning points (state transitions), a task that typically challenges models limited to short-context windows. Furthermore, on the task of segmented utterance emotion classification within the continuous stream,

Model	High Conf.	Low Conf.	Δ
Gemini 3	0.817	0.804	-0.013
Qwen 3 Flash	0.873	0.849	-0.024
Qwen2.5	0.850	0.852	+0.002
Qwen3(FT)	0.987	0.955	-0.032

Table 8: Average confidence scores on the High-Confidence and Low-Confidence subsets. Δ denotes the difference (Low Conf. − High Conf.).

the model attains a 71% accuracy on the Basic-7 taxonomy and a Macro-F1 score of 0.55 on the fine-grained GoEmotions taxonomy.

5 Conclusion

In conclusion, we introduce EmoS, a high-precision benchmark that elevates emotion understanding from a static to a dynamic level. By rigorously filtering noise and introducing a streaming subset, EmoS resolves the limitations of fragmented datasets. Our benchmarking highlights distinct model behaviors: while Gemini-3 prioritizes precision, fine-tuned models excel in coverage, and long-context capabilities prove essential for tracking narrative shifts . Ultimately, EmoS lays the foundation for training and evaluating future MLLMs capable of mastering emotional variations.

Limitations

While EmoS prioritizes label fidelity through a rigorous dual-layer human annotation pipeline, this focus inevitably constrains the scale of the newly introduced streaming dataset relative to massive web-crawled corpora. Specifically, the Streaming Monologue subset, comprising 50 videos, serves primarily as a high-standard evaluation benchmark rather than a large-scale training source. We are committed to iteratively updating and expanding this Streaming Monologue subset; however, this process requires time, primarily due to the limited availability of suitable movie clips and video content. Subsequent progress will be synchronously updated on our GitHub.

Furthermore, the current iteration is restricted to Chinese and English; expanding to a broader spectrum of languages remains necessary to enhance cross-cultural generalization. Moreover, the inherent long-tail distribution of real-world emotions—where categories such as *Grief* account for only 0.1% of samples—similarly manifests within this dataset. This indicates a critical need for the development of more robust imbalanced learning strategies in future work.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST), the Tencent AI Lab Rhino-Bird Research Program (Grant No. EF2023-00151-FST), the Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), the Macao Young Scholars Program (Grant No. AM2024015), and the National Natural Science Foundation of China (Grant No. 62266013). We thank Pengyu Chen, Zirui Chen, Yanxiu Liu, Fangfei Ren, Siqi Chen, and all the annotators for their valuable contributions to building the EmoS dataset.

References

Ana Aguilera, Diego Mellado, and Felipe Rojas. 2023. An assessment of in-the-wild datasets for multimodal emotion recognition. *Sensors*, 23(11):5184.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

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

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28. Original Dawid–Skene model for aggregating multiple annotator labels via EM.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. 2013. Emotion recognition in the wild challenge (EmotiW): Challenge and workshop summary. In *Proceedings of the 2013 ACM International Conference on Multimodal Interaction (ICMI '13)*, pages 371–372, Sydney, Australia. ACM.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3/4):169–200. Classic argument for discrete basic emotions (e.g., anger, joy, fear, sadness, disgust, surprise).

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Google. 2025. Gemini 3 pro model card. Model card listed on Google's official Model Cards site.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. 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, New York, NY, USA. ACM.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Technical report, University of Pennsylvania.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 2025. 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*, volume 267 of *Proceedings of Machine Learning Research*, pages 36993–37014.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022a. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 24th International Conference on Multimodal Interaction*.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022b. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, New York, NY, USA. ACM.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. 2012. 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.

Amir Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*, pages 527–536.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Shanghai, China. IEEE.

Yemin Shi, Yu Shu, Siwei Dong, Guangyi Liu, Jaward Sesay, Jingwen Li, and Zhiting Hu. 2025. Voila: Voice-language foundation models for real-time autonomous interaction and voice role-play. *arXiv*.

Jacob Whitehill, Ting-Fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*.

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. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of ACL*, pages 2236–2246.

A Copyright and Data Usage

This project strictly adheres to the copyright and license terms of all source datasets.

For MELD and CH-SIMS v2, we do not redistribute any original data, including but not limited to videos, audio tracks, or utterance transcripts. Instead, we only release re-annotation files in CSV format. These files contain video-level or utterance-level IDs and the corresponding newly assigned labels, and are designed to be directly compatible with the original datasets. Users can simply replace the original annotation CSV files with our provided CSV files after obtaining the raw data from the official sources. Access to the original MELD and CH-SIMS v2 datasets must be requested and downloaded through their respective official channels, in accordance with their original licenses.

In contrast, for our self-collected streaming dataset, we hold the rights to distribute the data. Therefore, both the annotations and the corresponding video files are packaged together and publicly released via an open online repository for research purposes.

By clearly separating re-annotations of third-party datasets from our own data, we ensure full compliance with existing dataset licenses while facilitating reproducible research.

The streaming video clips used in this dataset are sourced from openly available segments on YouTube, TikTok, and Bilibili, and the original video URLs are provided in the accompanying CSV files for transparency and traceability.

B Finetuning Details

For the fine-tuning of **Qwen-3** on the EmoS dataset, we conducted instruction tuning using a mixture of high-confidence and low-confidence samples, with 80% of the dataset being used for training and 20% reserved for testing. To reduce computational costs while maintaining performance, we employed Low-Rank Adaptation (LoRA) with a rank of 16 and a scaling factor (alpha) of 32, setting the dropout rate to 0.0. LoRA adapters were applied to all linear layers, including the attention module's q_proj, k_proj, v_proj, o_proj, and the feed-forward network's gate, up, and down projection layers.

During training, we used the AdamW optimizer with an initial learning rate of 5×10^{-5}, paired with a cosine learning rate decay strategy and 10% warmup steps. We set the batch size to 1 on a single H800 GPU and applied gradient accumulation over 8 steps. The training was conducted for 3 epochs under BF16 mixed precision until the loss converged.

C Additional dataset statistics information

The following table contains additional statistical information of the dataset. Due to the length constraints of the paper, this information cannot be included in the main text. Please refer to table9 10 11.

Emotion	CN (%)	EN (%)	ALL (%)
admiration	7.06	1.70	5.05
amusement	6.24	3.23	5.11
anger	16.14	14.43	15.50
annoyance	13.26	8.10	11.33
approval	2.92	3.00	2.95
caring	11.90	5.10	9.35
confusion	13.38	14.57	13.83
curiosity	5.50	11.03	7.58
desire	10.34	9.47	10.01
disappointment	9.76	3.17	7.29
disapproval	16.06	11.67	14.41
disgust	15.46	14.43	15.07
embarrassment	1.52	4.40	2.60
excitement	10.26	4.60	8.14
fear	5.34	4.13	4.89
gratitude	3.22	1.23	2.48
grief	1.70	0.20	1.14
joy	28.64	23.03	26.54
love	7.54	8.50	7.90
nervousness	8.18	7.10	7.78
neutral	24.04	45.93	32.25
optimism	17.20	2.67	11.75
pride	9.64	2.97	7.14
realization	11.84	25.53	16.98
relief	3.80	2.47	3.30
remorse	3.06	2.50	2.85
sadness	14.22	11.43	13.18
surprise	11.14	14.60	12.44

Table 9: Statistics of the 28-class emotion dataset (CN/EN/ALL). Excluding "unsure"

Emotion	%	Emotion	%
admiration	9.84	amusement	6.26
anger	9.17	annoyance	12.75
approval	7.38	caring	10.51
confusion	6.71	curiosity	5.37
desire	5.82	disappointment	9.17
disapproval	21.92	disgust	6.94
embarrassment	1.34	excitement	8.95
fear	1.79	gratitude	1.12
grief	13.20	joy	0.89
love	4.03	nervousness	5.37
neutral	2.24	optimism	7.83
pride	15.44	realization	22.60
relief	4.25	remorse	9.17
sadness	3.36	surprise	0.22

Table 10: Emotion distribution of the streaming subset (GoEmotions, 28 classes).

Emotion	%	Emotion	%
sadness	∼24.2	anger	∼23.0
neutral	∼20.6	joy	∼20.4
disgust	∼6.7	fear	∼3.8
surprise	∼1.3		

Table 11: Emotion distribution of the streaming subset (basic 7 classes).

D Detailed results of multi-classification

Please refer to table12

E Chinese and English distinction

We also recorded the evaluation results for the Chinese and English subsets and have placed them in the appendix for reference. Please refer to table13.

F Data Structure and Annotation Schema

The EmoS dataset is organized into three primary directories, each catering to different modalities and annotation granularities. The detailed structure and schema definitions are provided below.

F.1 Directory Organization

1. Streaming Long-form Video (`streaming/`) This directory contains long-form monologues. Each entry includes both the original full video and segmented clips. The data is paired with two types of annotation files:

- **Interpretation JSON:** Captures the narrative flow. Keys include `summary`, `overall_emotion_trend` (dominant emotion + trajectory description), and a list of `utterances`. Each utterance object contains the time range, bilingual transcripts, primary/secondary emotions, multimodal cues (audio/visual), intensity, and evidence.

Emotion	Gemini-3		Qwen-3 (FT)	
	P	R	P	R
admiration	0.448	0.650	0.320	0.780
amusement	0.173	0.667	0.095	0.800
anger	0.873	0.640	0.740	0.780
annoyance	0.403	0.768	0.290	0.860
approval	0.124	0.440	0.060	0.560
caring	0.463	0.481	0.330	0.620
confusion	0.638	0.280	0.510	0.430
curiosity	0.672	0.594	0.540	0.720
desire	0.531	0.195	0.390	0.340
disappointment	0.483	0.527	0.350	0.660
disapproval	0.366	0.701	0.250	0.830
disgust	0.739	0.143	0.560	0.300
embarrassment	0.250	0.381	0.160	0.520
excitement	0.356	0.762	0.240	0.880
fear	0.449	0.449	0.320	0.620
gratitude	0.917	0.688	0.780	0.820
grief	0.364	0.667	0.250	0.820
joy	0.824	0.658	0.690	0.800
love	0.703	0.329	0.560	0.480
nervousness	0.404	0.583	0.280	0.720
optimism	0.506	0.469	0.380	0.620
pride	0.698	0.661	0.580	0.800
realization	0.435	0.287	0.300	0.440
relief	0.500	0.250	0.360	0.420
remorse	0.382	0.500	0.260	0.650
sadness	0.703	0.627	0.580	0.760
surprise	0.597	0.484	0.460	0.620
neutral	0.748	0.343	0.620	0.520

Table 12: Detailed results of multi-classification (GoEmotions, 28 classes).

- **Timestamp JSON:** Provides strict temporal alignment. The top level includes metadata (language, title, global start/end). The `segments` list details the start/end seconds, text, and the corresponding slice filename for every sentence.

- **Cross-Index:** A sentence-level aggregation JSON that indexes all samples with fixed fields: language, global index, relative source path, title, timestamps, full transcript, and sub-sentence segment arrays.

2. Multi-Annotator Short Video (`ch-simsv2s/`) This subset focuses on short video clips with multiple human annotators per sample. The raw index contains `video_id`, `clip_id`, and multimodal tags. Annotations are split into two batches, each reviewed by three annotators for both Basic-7 (single-label) and GoEmotions (multi-label).

- **Basic-7 Aggregation:** Contains the three individual labels and the aggregated ground truth derived via the Dawid-Skene algorithm (`ds_label`). It includes the aggregation confidence (`ds_confidence`) and a

Model	Lang.	Accuracy	Macro-F1	Weighted-F1
EmotionLLaMA	En	0.333	0.250	0.380
	Zh	0.512	0.332	0.489
Qwen-2.5	En	0.578	0.371	0.556
	Zh	0.564	0.406	0.541
Qwen-3 (Base)	En	0.466	0.368	0.485
	Zh	0.693	0.624	0.686
Gemini	En	0.541	0.441	0.554
	Zh	0.682	0.597	0.668
Qwen-3 (FT)	En	0.668	0.591	0.661
	Zh	0.732	0.653	0.701

Table 13: Performance comparison on high-confidence samples across different models. "En" denotes English and "Zh" denotes Chinese.

unique_labels field to indicate conflict levels.

- **GoEmotions Set:** Stores the set of labels from all three annotators, including derived metrics such as union, intersection, and label distribution summaries.

3. Re-annotated MELD (`newdataset-MELD/`) A subset of MELD re-annotated to bridge the gap between simple sentiment and complex emotions.

- **GoEmotions Soft-Labels:** Contains the label sets from three annotators. It calculates the frequency/probability for each emotion (`Goemotions_prob_*`) and maps these soft labels to predict the best/second-best Basic-7 categories.

- **Basic-7 Hierarchical Table:** Provides a comparison between the original MELD emotion (`basic_original`) and the corrected label (`basic_corrected`), along with the confidence score and the reason for correction. Includes video paths and dialogue IDs for back-tracking.

F.2 Annotation Fields Hierarchy(Part of streaming data)

Video-Level Interpretation Designed for context understanding:

- `summary`: Abstract of the video content.

- `overall_emotion_trend`: Describes the dominant emotion and its evolution.

- `utterances`: A list containing multimodal evidence, intensity, and bilingual transcripts.

Sentence-Level Timestamps Designed for alignment:

- Top-level metadata combined with a `segments` list (index, start/end time, text, slice filename).

Text-Based Multi-Annotation Designed for label reliability:

- **Basic-7:** Aggregates individual annotations into `ds_label` (consensus), `ds_confidence` (reliability score), and `unique_labels` (disagreement indicator).

- **GoEmotions:** Derived fields include label union/intersection, set size, and soft probabilities defined as $P(label) = \frac{\text{count}}{3}$.

- **Correction/Fusion:** Maps GoEmotions probabilities back to Basic-7, generating fields for the best candidate (`basic_Goemotions_best`), the final corrected label, and explicit flags for whether the original label was modified and why.

F.3 Utilization Strategy

To maximize the utility of the EmoS dataset, we recommend the following strategies based on the data structure:

- **Single-Label Training (Basic-7):** Use `ds_label` as the primary supervision signal. The `ds_confidence` score can serve as a sample weight or a threshold for curriculum learning. The `unique_labels` count identifies "hard samples" with high human disagreement.

- **Multi-Label Training (GoEmotions):** Utilize the soft probability vectors (derived from annotator agreement) rather than binary targets to model emotional ambiguity.

- **Multimodal Alignment:** For the `streaming` subset, use the timestamp JSON to align text transcripts with visual/audio slices. The utterance-level interpretations can serve as a test set for long-context understanding.

- **Data Cleaning:** Use `video_id/clip_id` (or `dia/utt` IDs) to join text-level aggregation results with video-level raw files.

- **Stratified Evaluation:** We recommend splitting train/validation sets based on `ds_confidence` levels (High/Medium/Low) to ensure models are tested on samples with varying degrees of difficulty.

G Classification of Goemotions

The classification of Goemotions is 28 categories, with an additional category of "neutral", which is consistent with the original paper. Admiration, Amusement, Approval, Caring, Desire, Excitement, Gratitude, Joy, Love, Optimism, Pride, Relief, Anger, Annoyance, Disappointment, Disapproval, Disgust, Embarrassment, Fear, Grief, Nervousness, Remorse, Sadness, Confusion, Curiosity, Realization, Surprise, Neutral.

H Detailed Related Work

In this section, we provide a detailed analysis of the evolution of MER datasets, categorizing them into laboratory-controlled, in-the-wild, and LLM-generated benchmarks.

H.1 Lab-Controlled Datasets

Early research predominantly relied on datasets collected in controlled laboratory environments (McKeown et al., 2012; Ringeval et al., 2013). A representative benchmark, **IEMOCAP** (Busso et al., 2008), comprises approximately 12 hours of dyadic interaction. Due to its high-quality audio-visual recordings and fixed camera angles, it effectively captures rich facial expressions and has long served as a standard benchmark. Another significant dataset, **DAIC-WOZ** (Gratch et al., 2014), focuses on clinical psychology scenarios through a "virtual interviewer" setup. However, these datasets universally suffer from a lack of ecological validity (Dhall et al., 2013). Constrained by laboratory settings and scripted protocols, the emotional expressions lack spontaneity, limiting model generalization in unconstrained environments.

H.2 In-the-wild Datasets: TV & Social Media

To pursue more naturalistic expressions, the community shifted towards datasets derived from TV shows and social media (Dhall et al., 2013; Mollahosseini et al., 2019). **MELD** (Poria et al., 2019), extended from Friends, introduced multi-party interactions. However, it contains significant noise (Aguilera et al., 2023): audio tracks are contaminated by canned laughter, and cinematic editing results in frequent shot transitions. Such visual discontinuity bottlenecks multimodal feature extraction.

Similarly, **CMU-MOSI** (Zadeh et al., 2016) and **CMU-MOSEI** (Zadeh et al., 2018) provide large-scale monologues from YouTube. However, due to the lack of rigorous cleaning, MOSEI contains many non-emotional segments (e.g., neutral narration), and inconsistent label quality compromises its reliability. Addressing the scarcity of non-English resources, **CH-SIMS v2** (Liu et al., 2022a) provided Chinese multimodal data with high-quality alignment but only offers coarse-grained (Positive/Negative) polarity labels, limiting its application in complex emotion analysis.

H.3 LLM-Generated Datasets & Challenges

With the advancement of LLMs, works such as **Emotion-LLaMA** (Cheng et al., 2024) and **MER-Caption** (Lian et al., 2025) explore instruction tuning using synthetic data. These datasets introduce complex categories and natural language descriptions. However, they present two core pitfalls: (1) open-ended labels complicate standardized evaluation (Huang et al., 2024), and (2) excessive reliance on AI-generated annotations without rigorous human verification leads to inherent hallucination issues (Ji et al., 2023), compromising ground truth credibility.

H.4 Other Dynamic Datasets

Recent works like **DFEW** (Jiang et al., 2020) and **MAFW** (Liu et al., 2022b) have scaled up dynamic facial expression recognition in the wild. However, these datasets primarily focus on visual modalities and lack the rigorous bilingual textual alignment and fine-grained psychological annotation (e.g., GoEmotions) present in EmoS.

I Prompt

Prompt A (Compact JSON Classification). Instructions:

- Examine visual cues (facial expressions, gestures), vocal tone, and the transcript.
- Ignore background characters unless they directly drive the main speaker's emotion.
- Prefer "Neutral" only when no consistent emotion is detectable.

- If the transcript conflicts with the audio tone or visuals, prioritize audio/video evidence.
- In addition to a single primary emotion, estimate multi-label GoEmotions signals. Use **only** these labels: `{Goemotions_labels}`.
- List every GoEmotions label that truly appears; include only plausible entries sorted by confidence.
- Return **JSON only** with the following fields:

Output Format (Text Only). Return **plain text** (not JSON) with the following fields:

- **Predicted emotion:** `<label>`
- **Model confidence:** a float in `[0, 1]`
- **GoEmotions labels:** list all inferred labels (use only the allowed set), each with a confidence score in `[0, 1]`, sorted by confidence
- **Evidence:** one sentence describing the audio/visual/transcript cues used
- **GoEmotions rationale:** briefly explain why the multi-label set was chosen

Summary reminder:

- Report only the GoEmotions labels you genuinely infer; there is no required minimum count.
- When transcripts disagree with tone or visuals, audio and video cues take priority.

Clip metadata:

- language: `{language}`
- transcript: `{transcript}`

Prompt B (Streaming Utterance-Level Annotation). You are an emotion researcher specialized in multimodal analysis. Based on the uploaded video (including both visuals and audio), complete the following tasks:

1. Summarize the clip in 2–3 sentences, highlighting key turning points and the overall emotional atmosphere.
2. Analyze only the absolute main subject character in the visuals and storyline, ignoring all other characters or background figures. Extract all utterances spoken by the main subject (infer/supplement when necessary), and provide structured emotion annotations for each utterance.
   - The primary emotion (`primary_7class`) must be strictly one of the traditional seven classes: `Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise`.
   - The secondary GoEmotions labels (`secondary_Goemotions`) can be multi-label selections from the following list; if nothing matches, use `Unsure`: `admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief, anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness, surprise, confusion, curiosity, realization, neutral, Unsure`.
3. Output a UTF-8 JSON in the following format:

Output Format (Text Only).

1. **Summary (2–3 sentences):** Briefly describe the clip, highlighting key turning points and the overall emotional atmosphere.
2. **Overall Emotions Trend:**
   - **Dominant emotion:** One label from the 7-class set (Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise).
   - **Trajectory:** Describe how emotions evolve over time, citing MM:SS timestamps.
3. **Utterance-Level Annotations (chronological, non-overlapping timestamps):** For each utterance spoken by the main subject, provide:
   - **Timestamp range:** MM:SS–MM:SS
   - **Speaker guess (optional):** May be left empty
   - **Chinese transcript:** The utterance in Chinese

- **English transcript (optional):** English transcription/translation if possible
- **Primary emotion (7-class):** Exactly one label from the 7-class set
- **Secondary emotions (GoEmotions):** A multi-label set chosen only from the allowed GoEmotions list; use `Unsure` if none fit
- **Intensity:** 1–5
- **Audio cues:** Tone, volume, pauses, etc.
- **Visual cues:** Facial/body cues, camera context, etc.
- **Evidence:** One brief sentence citing supporting cues from visuals and/or dialogue

Constraint: Output should be plain text (not JSON). If exact timestamps cannot be identified, explicitly state the assumptions used.

Constraints: Please output **UTF-8 JSON only**. Ensure each utterance's timestamp range does not overlap and is ordered chronologically. If exact timestamps cannot be identified, clearly state the assumptions used.

For Emotion Llama, we used the official sentiment classification prompt.