OSMO: Open-vocabulary Self-eMOtion Tracking

Mohamed Abdelfattah[1] Bugra Tekin Fadime Sener Necati Cihan Camgoz
Eric Sauser Shugao Ma Alexandre Alahi[1] Edoardo Remelli

[1] École Polytechnique Fédérale de Lausanne (EPFL)

Abstract

*We introduce the novel task of **egocentric self-emotion tracking**, which aims to infer an individual's evolving emotions from egocentric multimodal streams such as voice, visual surroundings, semantic subtext, and eye-tracking signals. To establish this research direction, we present: (1) **OSMO dataset**, a large-scale annotation effort on 110 hours of existing bilingual smart-glasses recordings, establishing the largest egocentric emotion dataset and the first with subject-wise emotion timelines; (2) **OSMO benchmark**, a suite of five tasks (emotion recognition, sentiment, intensity, localization, and reasoning), that redefine emotion understanding as a continuous, context-aware process rather than discrete classification of trimmed videos; (3) **OSIRIS**, a large multimodal model that tracks emotions over time by reasoning over the user's personal emotion history, current expressions, and egocentric observations. Extensive evaluations show that OSIRIS achieves a state-of-the-art performance, delivering, for the first time, coherent emotion timelines from egocentric data. Project website: https://osmo-emos.github.io.*

1. Introduction

Self-emotion tracking can reduce depression symptoms by 34% [27], anxiety by 20% [58], and help over 85% of people feel more in control of their mood [62]. Yet its adoption remains low, because current solutions, *e.g.,* mobile apps, rely on high-friction manual emotion logging. Therefore, we ask: *can emotion[1] perception and tracking be automated?*

Smart glasses [18] offer a passive, unobtrusive, and continuous means of emotion tracking. Designed for all-day wear, they enable seamless long-term perception in natural settings. Their integrated multimodal sensors capture vocal tone, gaze behavior, and environmental context, providing a rich basis for robust, context-aware emotion recognition.

[1]We focus on **embodied emotions**, observable physiological expressions (*e.g.,* speech, eye movements, body language), detectable via non-intrusive sensors, unlike **internal emotions** that require intrusive sensing.

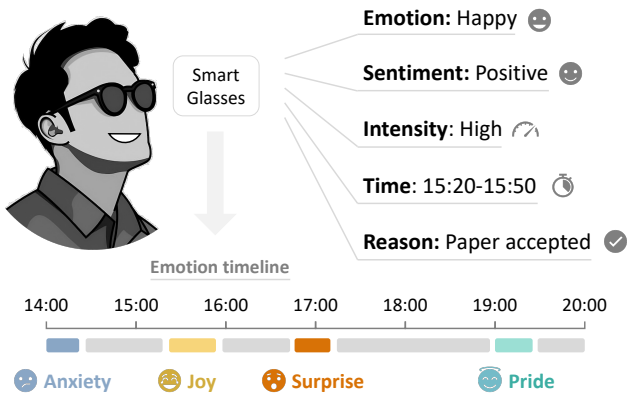


Figure 1. **Self-emotion tracking.** Using multimodal inputs from smart glasses (egocentric video, audio, eyes, and text), our goal is to construct a timeline of the user's affective states.

However, existing emotion recognition datasets are not suitable for training models deployable on smart glasses, as they are typically exo-centric [2, 10, 34, 35, 38] or consist of short, isolated video clips [19], which precludes modeling the continuity of emotions. More critically, their primary sources (lab settings [2, 23], movies [10, 35, 38, 51] and internet vlogs [19]) feature exaggerated and staged expressions, thus failing to capture the subtle and spontaneous emotions of real-world environments, which is far more challenging.

Consequently, recent emotion **Large Multimodal Models (LMMs)** [10, 19, 35, 70], inherit these data flaws and face key limitations: (1) *reliance on facial views*, performing poorly in egocentric settings; (2) *processing utterances in isolation*, misinterpreting context-dependent meaning (*e.g.,* "That's just great" as sincere vs. sarcastic); (3) *ignoring the influence of prior emotions*, overlooking carry-over effects [56, 60]; and (4) *lack interpretable reasoning*, producing spurious and ungrounded outputs.

To address these limitations, this paper introduces a new framework for egocentric emotion tracking from smart glasses (see Fig. 1) featuring three novel contributions:

1) OSMO Dataset. We introduce Open-vocabulary Self-

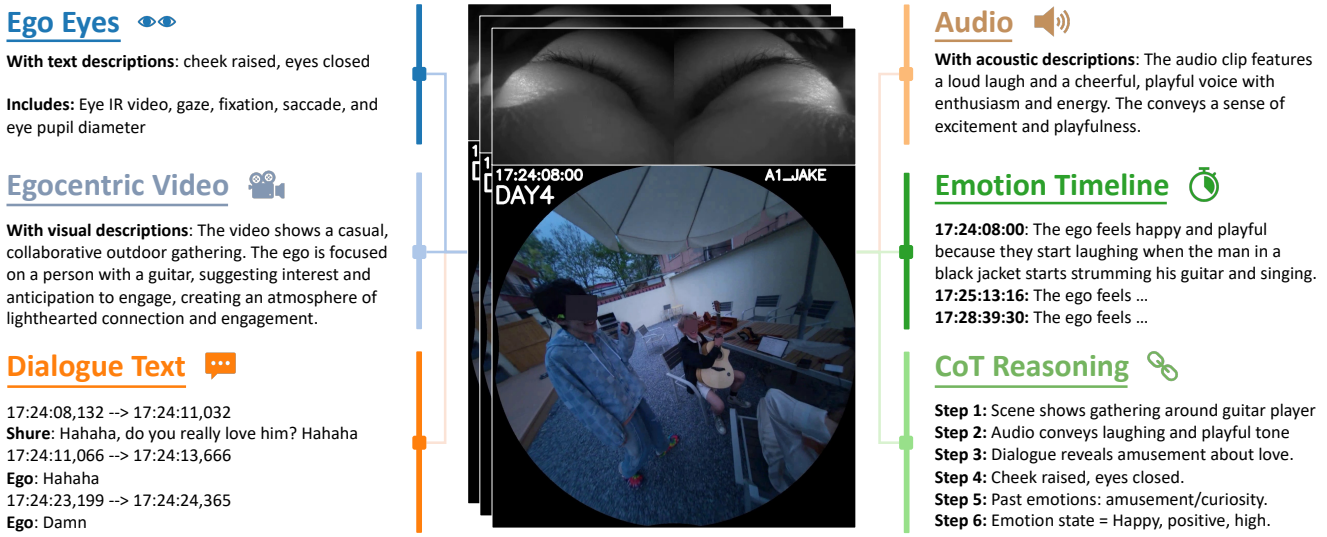


Figure 2. **OSMO dataset sample.** Given synchronized raw egocentric data (video, audio, eyes), OSMO adds human-annotated subject-wise emotion timelines, detailed LMM-generated modality descriptions, and Chain-of-Thought labels. Raw sample source: EgoLife [71].

e**MO**tions (OSMO), a large-scale annotation of 110 hours of smart-glasses recordings from open-source datasets [41, 42, 71], yielding the largest egocentric emotion resource to date and the first with frame-level *emotion timelines* capturing *real-world* affective dynamics. Our annotations enrich the original multimodal data (RGB, audio, eye tracking) with detailed, timestamped labels of the ego wearer's affective states (see Fig. 2). Uniquely, OSMO spans English and Mandarin, enabling cross-lingual emotion analysis.

2) OSMO Benchmark. We define five benchmark tasks for self-emotion tracking: (i) open-vocabulary emotion recognition, (ii) sentiment analysis, (iii) intensity prediction, (iv) temporal localization, and (v) emotion reasoning. Collectively, these tasks move beyond categorical classification toward a holistic understanding of emotional dynamics.

3) OSIRIS Model. We present **OSIRIS** (**O**mnimodal **S**elf-emotion **I**nference with **R**easoning on **I**nput **S**ignals), the first LMM to jointly process egocentric video, audio, dialogue, and eye infrared feed for emotion tracking. Key innovations include a memory module for modelling emotion carry-over effects [56, 60] and the **SENSE** framework (**S**tructured **E**motional reaso**N**ing from **SE**nsory inputs) for generating Chain-of-Thought (CoT) supervisory labels to force multimodal analysis, *i.e., thinking*, before emotion inference. Extensive experiments show OSIRIS substantially outperforms state-of-the-art (SOTA) LMMs.

2. Related Works

Emotion Datasets. We categorize emotion datasets by their sources into two groups: **(1) Internet-sourced**, including datasets from movies [10, 34, 35, 38, 51] and social media [5, 19, 28, 69, 73]; **(2) Lab-controlled**, collected in controlled environments [2, 6, 7, 23, 39, 45, 54]. While these datasets have significantly advanced affective computing, they are primarily designed to capture elicited or performed emotions, which can differ from naturally occurring affective responses in everyday settings. Moreover, they are all exocentric except for E3 [19], which suffers from 1) noisy, handheld vlog recordings, 2) lack of eye-tracking or subject-level emotion timelines, and 3) closed-set emotion labels. In contrast, the OSMO dataset captures *real-world, unscripted emotions* with timestamped, subject-wise, open-vocab annotations paired with rich multimodal signals from smart glasses.

Emotion Modeling. Emotion recognition models are typically categorized by input modality. **(1) Unimodal** models infer emotions from a single source, *e.g.,* videos [32, 63, 65], eyes [22, 68, 76], audio [13, 20, 40], or text [8, 33, 52], but often lack context from other modalities (*e.g.,* text without tone). **(2) Multimodal** models [1], especially recent LMMs, achieve richer emotion understanding by integrating complementary cues. AffectGPT [35] uses pre-fusion for audio–video, E3 Emotion-LlaMA [19] combines visual–acoustic signals in first-person videos, and Emotion-LLaMA [10] employs adaptive fusion for human-centric affect analysis. Yet, they overlook conversation history (e.g., emotion in "unbelievable!" depends on prior context), emotional continuity, and interpretability. In contrast, OSIRIS models contextual emotions with explicit CoT reasoning.

3. OSMO Dataset

How can we build an emotion tracking dataset? We begin by defining five key requirements: **(1)** *In-the-wild*: capturing

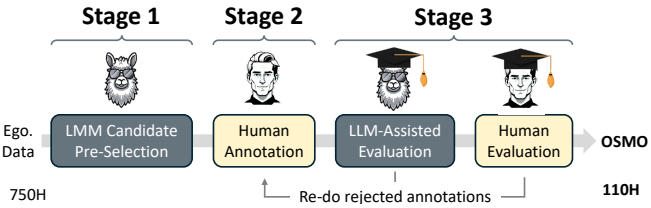


Figure 3. OSMO Human-LMM collaboration approach.

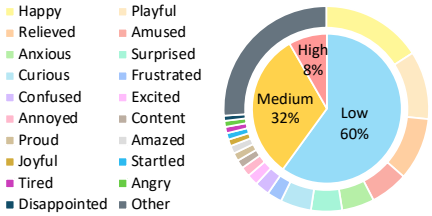


Figure 4. OSMO Emotions and Intensities.

Metric	Value
#Hours	110 h
#Ann.	116,100
#Segments	23,200
#Subjects	288
#Emotions	1,027
#Words/cap.	31.4 avg.
Languages	ZH / EN
Emotion freq.	2.1 min
Emotion dur.	7.7 s avg.

Table 1. Statistics.

natural emotions in real-world scenarios; **(2)** *Longitudinal*: with timestamped data to model emotional dynamics; **(3)** *Ego identities*: enabling subject-wise emotion timelines and personalized tracking; **(4)** *Smart glasses recordings*: providing rich multimodal signals (egocentric video, eye/hand tracking); and **(5)** *Open-vocab annotations*, with labels for emotions, sentiment, intensity, temporal span, and causes.

OSMO meets all five requirements by leveraging a key insight: instead of collecting new data, we annotate three existing datasets, **EgoLife** [71], **Nymeria** [42], and **Aria Everyday Activities (AEA)** [41]. This decision is strategic for two reasons: (i) these datasets already satisfy the first four requirements and only lack emotion labels, and (ii) their long-term, real-world recordings were captured *without instructing participants to display emotions*, yielding spontaneous, unscripted affect largely absent from prior emotion datasets. For high-quality annotations, we introduce a novel three-stage human–LMM collaborative pipeline (Fig. 3).

3.1. Stage 1: LMM Pre-selection

Although our source datasets [41, 42, 71] collectively provide 750 hours of egocentric recordings, emotional expressions are sparse [46, 59, 74], making exhaustive manual annotation impractical. To address this, we design a multi-stage *pre-selection pipeline* that leverages SOTA LMMs to efficiently narrow down our search for non-neutral segments.

We first use Whisper [53] to generate timestamped speech transcriptions and segment the recordings into 200K utterance-level clips. Each clip is then pseudo-labeled with one of Ekman's six basic emotions [16] (happiness, sadness, fear, anger, disgust, surprise) or neutral using four SOTA LMMs: Emotion-LLaMA [10], AffectGPT [35], DeSTA2.5-Audio [40], and Qwen-Audio2 [13]. Majority voting (≥ 3) mitigates individual model noise, retaining clips with consistent non-neutral predictions and balanced emotion coverage. This produces 17.8K high-confidence clips, expanded into 30-second segments for context, yielding a curated 125-hour subset for human verification and annotation.

3.2. Stage 2: Human Annotations

While Stage 1 effectively crosses-out neutral segments, it remains unreliable in identifying the exact emotion type, duration, and trigger. To refine this, we recruit 41 annotators (gender-balanced, educationally diverse). Each annotator underwent a rigorous multi-phase training program, including detailed protocols, hands-on practice, iterative feedback, and consistency calibration. In total, the OSMO annotation process required more than **8,000 hours of human effort**. Importantly, *all LMM predictions were hidden to prevent bias.* Below we describe the annotation process:

Open-Vocabulary Emotion Recognition. Conventional closed-set emotion taxonomies [14, 50] fail to capture the richness and overlap of real-world emotions, forcing discrete labels (*e.g.,* anger, surprise) onto continuous states (*e.g.,* shyness, hesitation). To address this, we adopt an *open-vocabulary* paradigm [35], allowing annotators to freely describe emotions. Annotators labelled each segment with primary and secondary emotions guided by Plutchik's Wheel [50], a widely-adopted model of eight basic emotions, alongside open-vocab terms for fine-grained distinctions. Annotators used extended wheels [55, 67], examples, and calibration sessions to ensure consistency.

Sentiment Analysis. Sentiment captures coarse affective polarity: positive, negative, or neutral. Following [15], favorable emotions (*e.g.,* joy, excitement) were labeled as positive, undesired ones (*e.g.,* anger, fear) as negative, and unassigned segments as neutral. For ambiguous emotions (*e.g.,* surprise, confusion), we used LlaMa3 [61] to infer sentiment polarity from the full annotation text by prompting it to classify the overall affect as positive, negative, or neutral.

Emotion Intensity. Annotators rated emotional strength on a three-level scale (following [7, 19, 64]): *low* (subtle, minimal cues like a calm or neutral voice), *medium* (noticeable but controlled affect with clear expressions), and *high* (dominant emotions with strong, overt manifestations like shouting or crying), based on audiovisual and contextual cues.

Emotion Localization. Annotators marked the start and end seconds of each emotional episode, defined by observable cues such as speech, environmental events, or sudden reactions, and ending once the emotion visibly subsided.

Emotion Reasoning. For each segment, annotators explained the Ego wearer's emotion using only observable

audiovisual cues. These concise rationales identified triggers such as interactions, environmental changes, or social events, ensuring evidence-based and interpretable annotations.

3.3. Stage 3: Annotation Quality Assessment

We implement a strict two-tiered quality control process:

- **LLM-assisted Validation:** Our annotation pipeline checks all annotation jobs for missing information, abnormal durations (<1s or >25s), brief captions (<7 words), and overlapping segments (<1s gap), flagging any anomalies. Then, we utilize a *LLaMA-as-a-judge* approach, where LLaMA3 [61] assigns a quality score (1-10) to each annotation based on the annotation rubric (rejecting < 8).
- **Human Evaluation:** A review team manually evaluates annotations along three dimensions: (1) *Category correctness*: whether the primary and secondary emotion categories align with the observable cues; (2) *Localization accuracy*: whether the temporal boundaries of the annotation are within 4 s of the correct start and end times; and (3) *Reasoning validity*: whether the description includes clear, observable triggers for the selected emotion.

Flagged samples at any stage are iteratively re-annotated until passing all quality checks, resulting in a highly consistent final OSMO dataset with human review agreement of 87.0% (category), 91.2% (localization), and 82.6% (reasoning). Our evaluation highlights a complementary division of labor: LMM majority voting excels at narrowing down candidate regions (88.0% of which were retained by humans) but is substantially less effective at precise classification, achieving an overlap rate [35] of only 48.6 against human labels. This confirms the synergy in our pipeline: LMMs for filtering with human expertise for reliable annotations.

3.4. OSMO Dataset Statistics.

Table 1 summarizes the dataset statistics, and Table 2 compares it with existing emotion datasets. Remarkably, OSMO is (1) the largest egocentric emotions dataset; (2) the first to feature real-world emotions; (3) the first to provide subject-wise emotion timelines and eye-tracking data; and (4) the richest in annotation, including over 1000 open-vocab emotions (see Fig 4), sentiment, intensity, localization, and reasoning. Furthermore, the dataset includes English (41.3%) and Mandarin (58.7%) subsets, offering rich cross-cultural variation in language, interaction style, and environment.

4. OSMO Benchmark

Open-Vocab Emotion Recognition (OVER): evaluates whether generated captions reflect the ego's emotional state. Following [35], we report two metrics: the Set Overlap Score (SOS), reflecting predicted/ground-truth open-vocab emotion sets overlap, and the Hit Rate (HR), indicating whether any ground-truth emotion appears in the prediction.

Table 2. **Comparison between OSMO and prior emotion recognition datasets. Hrs.**: #hours; **Ann.**: #annotations; **Mod.**: modalities (v:video, a:audio, t:text, e:eye); **Ego**: collected with egocentric smart glasses; **TL**: contains long-term emotion timeline; **Un.**: unscripted real-world emotions; **OV**: open-vocabulary emotions; **Loc.**: temporal localization; **Int.**: intensity levels. Abbrev.: MERR-C/F = MERR-Coarse/Fine; MER-C = MER-Caption. † only emotion subset is considerd in Seamless Interactions [2].

Dataset	Hrs.	Ann.	Mod.	Ego	TL	Un.	Loc.	OV	Int.
OV-MERD [34]	0.4	332	v,a,t	✗	✗	✗	✗	✓	✗
MERR-F [10]	3.9	4.5K	v,a,t	✗	✗	✗	✗	✓	✗
SeamInt† [2]	4.7	5.9K	v,a,t	✗	✗	✗	✗	✗	✓
InterAct [23]	10.0	241	v,a,t	✗	✗	✗	✗	✗	✓
MAFW [38]	11.1	10K	v,a,t	✗	✗	✗	✗	✗	✓
MELD [51]	13.7	13.7K	v,a,t	✗	✗	✗	✓	✗	✓
MERR-C [10]	24.8	28.6K	v,a,t	✗	✗	✗	✗	✓	✗
MER-C+ [35]	28.8	31.3K	v,a,t	✗	✗	✗	✗	✓	✗
MER-C [35]	106	115.6K	v,a,t	✗	✗	✗	✗	✓	✗
E3 [19]	72	81.2K	v,a,t	✗	✗	✗	✓	✗	✓
OSMO (Ours)	**110**	**116.1K**	**v,a,t,e**	✓	✓	✓	✓	✓	✓

Sentiment Analysis (SA): assesses whether the model captures the ego's overall tone across positive, negative, neutral. We report accuracy and Weighted F-score (WAF) for balanced evaluation under class imbalance.

Intensity Prediction (IP): measures how well the model predicts emotion strength across three levels {*low, medium, high*}, using the WAF for balanced evaluation and accuracy as a simpler complementary metric.

Emotion Localization (EL): tests whether the model correctly identifies when an emotion occurs. Predicted start–end times are compared with ground truth using mIoU (mean temporal overlap) and R_n,U@m.

Emotion Reasoning (ER): assesses whether the model correctly explains emotions. In addition to quality metrics (BLEU [49], ROUGE-L [36], and METEOR [4]), we use a LlaMa-as-a-judge (following [19, 43]) scoring 1–100 on information correctness (IC), detail orientation (DO), contextual understanding (CU), and temporal consistency (TUC).

Evaluation Protocols. We evaluate OSMO under four protocols assessing different types of generalization:

- **XSub** (Cross-Subject): Tests generalization across participants using disjoint subject sets (*train/val:* 223/57 subjects; 15.4K/3.8K clips).
- **XTime** (Cross-Time): Measures cross-day robustness (*train/val:* 106/25 days; 15.7K/3.4K clips).
- **XLang** (Cross-Language): Evaluates transfer between Mandarin and English, including in-language (`zh/zh`, `en/en`) and cross-language (`zh`→`en`, `en`→`zh`) setups.
- **XSet** (Cross-Set): Assesses generalization to unseen held-out set from annotated AEA [41] (8 subjects; 4.1K clips).

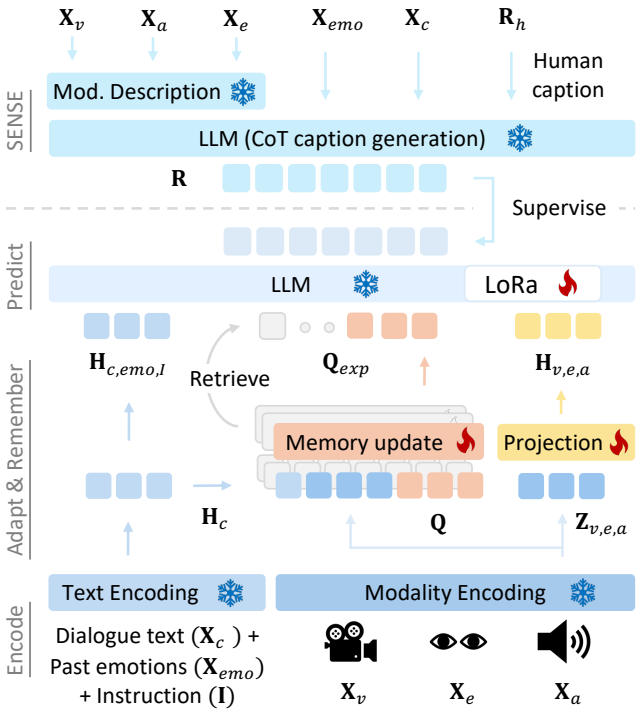


Figure 5. **Overview of OSIRIS.** OSIRIS integrates multimodal inputs, video ($\mathbf{X}_v$), eyes ($\mathbf{X}_e$), audio ($\mathbf{X}_a$), text ($\mathbf{X}_c$), and prior emotions ($\mathbf{X}_{emo}$), to infer affective states in five steps: (1) *Encode* multimodal signals using frozen expert encoders, (2) *Adapt/project* representations into a shared language space, (3) *Remember* past expressions by retrieving recently stored queries and updating the memory with the current state, (4) *SENSE* generation of CoT supervisory caption $\mathbf{R}$, and (5) *Predict* $\mathbf{R}$ through the LLM.

5. OSIRIS Model

Overview. OSIRIS aims to perceive the world through the wearer's eyes, enabling contextual understanding and continuous tracking of evolving emotional states. Building on recent advances in LMMs [3, 11, 12, 26, 77], it fuses egocentric modalities from smart glasses to infer and explain affect. Its key novelty lies in modeling emotion continuity (how feelings evolve over time) and multimodal reasoning grounded in perceptual and conversational evidence. Formally, for an incoming sample j, the multimodal input $\mathbf{X}^j$ includes $\mathbf{X}_v^j$ (egocentric video), $\mathbf{X}_e^j$ (eye video), $\mathbf{X}_a^j$ (audio), $\mathbf{X}_c^j$ (dialogue textual context), and $\mathbf{X}_{emo}^j$ (prior emotions). Conditioned on an instruction $\mathbf{I}$, OSIRIS generates a textual response $\mathbf{R}^j$ describing the wearer's emotional state through the following five-step process.

5.1. Step 1: Encode

Egocentric Video. The egocentric video stream $\mathbf{X}_v^j$ offers rich contextual information crucial for understanding emotions. For instance, among other cues, a lively social gathering (*e.g.,* Figure 2) may convey happiness, whereas a cluttered or dim environment might evoke feelings of sadness or anxiety. We process $\mathbf{X}_v^j$ using a frozen, off-the-shelf visual encoder [57], which outputs latent representations $\mathbf{Z}_v^j \in \mathbb{R}^{N_v \times C_d}$ across N_v frames, each represented by C_d-dimensional feature vectors.

Eyes. Prior works [22, 68, 76] show that eye dynamics alone can reveal affective states. In OSMO, we similarly find that ocular cues, like widened eyes in surprise or closed eyes in laughter (*e.g.,* Figure 2), strongly correlate with emotions. Hence, OSIRIS encodes the eye infrared video $\mathbf{X}_e^j$ with a frozen encoder [57] into $\mathbf{Z}_e^j \in \mathbb{R}^{N_v \times C_d}$, making it the first model to explicitly integrate eyes into the multimodal modeling of emotions.

Dialogue Context. Emotional meaning is often ambiguous in isolation; an utterance like "Wonderful" may convey either joy or frustration depending on the prior exchange. To disambiguate such cases, OSIRIS incorporates dialogue history containing N_u textual utterances from all speakers. Dialogue is encoded in two forms: 1) the dialogue audio stream $\mathbf{X}_a^j$ is encoded using a frozen acoustic model [53] into $\mathbf{Z}_a^j \in \mathbb{R}^{N_a \times C_a}$ of N_a acoustic frames and C_a channels, and 2) the dialogue text $\mathbf{X}_c^j$ is embedded by the LLM's text embedding layer [61] to obtain $\mathbf{H}_c^j \in \mathbb{R}^d$ with the LLM's embedding dimension d.

5.2. Step 2: Adapt

To unify heterogeneous modalities, each representation $\{\mathbf{Z}_m^j\}_{m\in\{v,e,a\}}$ is fed to a separate learnable adapter $G_m(\cdot)$ that maps to the LLM's embedding dimension d:

$$\mathbf{H}_m^j = G_m(\mathbf{Z}_m^j), \quad \mathbf{H}_m^j \in \mathbb{R}^d. \tag{1}$$

This step unifies the dimensions of diverse modalities, enabling cross-modal reasoning within the LLM.

5.3. Step 3: Remember

Personal Emotion History. Emotions are not discrete or instantaneous but evolve gradually over time with inertia [21, 29–31, 72]. For instance, a joyful surprise can leave lingering warmth and optimism. To model this temporal continuity, **OSIRIS** maintains a personalized *emotion log*:

$$\mathbf{L} = \{E^{(1)}, E^{(2)}, \ldots, E^{(j-1)}\}, \tag{2}$$

where each entry E^i represents an *emotion event*, a bounded episode capturing *(i)* what was felt, *(ii)* how it was expressed, and *(iii)* when (and for how long) it occurred.

(i) What (Semantic content): Each emotion event stores an open-vocabulary textual description $\mathbf{O}^i$ (*e.g.,* "happy", "disappointed") that semantically characterizes the felt emotion in natural language, providing an interpretable and flexible anchor aligned with human emotion vocabularies.

(ii) How (Multimodal expression): OSIRIS represents each event through a multimodal expression signature. Embeddings $\{\mathbf{Z}_m^i\}_{m\in\{v,e,a\}}$ and $\mathbf{H}_c^j$ are projected, pooled, and normalized into descriptors $\{\tilde{\mathbf{z}}_m^i\}_{m\in\{v,e,a,c\}} \in \mathbb{R}^d$, each weighted by a modality gate $\alpha_m = \sigma(g_m)$. The gated descriptors are concatenated and refined via cross-attention using N_{ms} learnable queries $\mathbf{Q} \in \mathbb{R}^{N_{ms}\times d}$:

$$\hat{\mathbf{z}}_m^i = \alpha_m \tilde{\mathbf{z}}_m^i, \quad \hat{\mathbf{z}}^i = [\hat{\mathbf{z}}_v^i, \hat{\mathbf{z}}_e^i, \hat{\mathbf{z}}_a^i, \hat{\mathbf{z}}_c^i],$$
$$\mathbf{Q}^i = \texttt{CrossAttention}(\mathbf{Q}, \hat{\mathbf{z}}^i), \quad \mathbf{Q}^i \in \mathbb{R}^{N_{ms}\times d}. \quad (3)$$

The resulting compact representation $\mathbf{Q}^i$ serves as a multimodal code summarizing expressive cues across channels.

(iii) When (Temporal metadata): Each event includes a timestamp t^i and a duration D^i:

$$E^i = \{\mathbf{O}^i, \mathbf{Q}^i, t^i, D^i\}. \quad (4)$$

This allows OSIRIS to jointly model evolving emotions within their *semantic*, *expressive*, and *temporal* contexts.

Memory Retrieval. At inference time t^j, OSIRIS retrieves the N_p latest prior textual emotions $\mathbf{X}_{emo}^j$ and N_q multimodal queries $\mathbf{Q}_{exp}^j$:

$$\mathbf{X}_{emo}^j = \{\mathbf{O}^{(j-N_p)}, \ldots, \mathbf{O}^{(j-1)}\},$$
$$\mathbf{Q}_{exp}^j = \{\mathbf{Q}^{(j-N_q)}, \ldots, \mathbf{Q}^{(j)}\}, \quad (5)$$

each paired with temporal metadata $(\Delta t^i, D^i)$, where $\Delta t^i = t^j - t^i$ denotes elapsed time. Semantic emotions are included with textual inputs, while multimodal codes are directly inserted into LLM tokens, allowing OSIRIS to interpret emotions as part of a continuous temporal trajectory.

5.4. Step 4: SENSE

Current models [13, 19, 40] typically infer emotions in a single, opaque step, relying on spurious correlations like "tears equal sadness." This ignores the nuance that tears can signal joy or sorrow depending on context, and it fails to leverage the autoregressive reasoning that LLMs excel at.

Inspired by recent advances in chain-of-thought reasoning [9, 25, 47, 48, 66, 75], we reframe emotion recognition as a structured reasoning problem, where OSIRIS first interprets perceptual cues before inferring emotions. For instance, in Fig. 2 sample, it must first reason over cues such as the *guitar, laughter, and closed eyes* before predicting happy.

Unfortunately, manually annotating such detailed multimodal cues is prohibitively expensive. However, we observe that human captions $\mathbf{R}_h$ correctly capture emotions but lack exhaustive perceptual details, while multimodal captioning models provide fine-grained sensory details yet miss emotional depth. Motivated by this observation, we introduce **SENSE** (**S**tructured **E**motional reaso**N**ing from **SE**nsory inputs), a data-generation framework that enriches the human affective captions with the descriptive precision of LMMs.

We first feed raw signals $(\mathbf{X}_v, \mathbf{X}_a)$ to SOTA video [71] and audio [40] captioning models to extract detailed visual $(\mathbf{R}_v)$ and acoustic $(\mathbf{R}_a)$ descriptions, and derive eye-related cues $(\mathbf{R}_e)$ by mapping emotions to eye action units [17, 44]. Next, we feed $\mathbf{R}_h$, $\mathbf{R}_v$, $\mathbf{R}_a$, $\mathbf{R}_e$, $\mathbf{X}_c$, and $\mathbf{X}_{emo}$, to LLaMA3 [61] which acts as a *cognitive proxy*, linking diagnostically relevant signals and producing a chain of reasoning $\mathbf{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_6\}$, covering step-by-step analysis: (1) visual, (2) audio, (3) dialogue, (4) eye, (5) prior emotions, and (6) final inference. By fine-tuning OSIRIS on $\mathbf{R}$, we teach the model not only *what* to predict but also *how* to reason, shifting the task from direct classification to a cognitively grounded process that mirrors human introspection.

5.5. Step 5: Predict

Training Objective. Given the multimodal context $\mathcal{X} = \{\mathbf{X}_v, \mathbf{X}_e, \mathbf{X}_a, \mathbf{X}_c, \mathbf{X}_{emo}, \mathbf{Q}\}$ and an instruction $\mathbf{I}$, OSIRIS maximizes the likelihood of generating $\mathbf{R} = \{r_l\}_{l=1}^{L_r}$:

$$\theta^* = \arg\max_\theta \prod_{l=1}^{L_r} P_\theta(r_l \mid \mathcal{X}, \mathbf{I}, r_{<l}), \quad (6)$$

where $r_{<l}$ are previously generated tokens. We fine-tune the base LLM [61] using LoRA [24], which inserts low-rank trainable adapters into the attention and feed-forward layers. This allows efficient optimization for multimodal emotional reasoning while keeping most pretrained weights frozen.

6. Results

Below we discuss quantitative results:

XSub and XTime Results. Table 3 shows that OSIRIS establishes a new SOTA, outperforming the zero-shot LLaMa3 baseline [61] by **+35.1** on XSub and **+35.6** on XTime. As expected, zero-shot emotion-specialized LMMs [10, 19, 35] underperform on OSMO due to the differences in training data and task setup. This highlights that models trained on performed or exaggerated emotion datasets *may not directly transfer to the more natural, unscripted expressions* captured in OSMO. Under the finetuning setup, OSIRIS achieves the best overall performance, surpassing the previous SOTA finetuned AffectGPT [35] by an average of **+10.7** (XSub) and **+10.1** (XTime) across all metrics. This demonstrates that the reliance of prior works on isolated, out-of-context prediction is ill-suited for modeling emotional continuity. In contrast, OSIRIS dynamically tracks emotions by integrating emotion history with current contextual cues. Notably, it achieves the largest gains in reasoning, outperforming AffectGPT by an average **+14.1** on LLaMa-Judge metrics, validating the effectiveness of our SENSE reasoning strategy.

XLang Results. Table 4 presents the cross-lingual performance on OSMO-XLang. Interestingly, all models transfer

Table 3. **OSIRIS outperforms SOTA LMMs on OSMO-XSub and OSMO-XTime.** Tasks: OVER = open-vocab emotion recognition; SA = sentiment analysis; IP = intensity prediction; EL = emotion localization; ER = emotion reasoning. Bl = BLEU4, RL = ROUGE-L. Mean Δ is the average gain over zero-shot Meta-Llama-3-8B-Instruct [61] baseline†. *Highlighted* rows are finetuned; others are zero-shot.

Model	Modality				OVER		SA	IP	EL	ER						Mean Δ
	V	A	T	E	HR	SOS	WAF	WAF	mIoU	IC	DO	CU	TUC	Bl	RL	
OSMO XSub Evaluation Protocol (non-overlapping subjects in train/val):																
Baseline LLaMa3† [61]			✓		45.4	36.4	47.7	32.5	25.4	36.2	36.8	36.1	35.3	8.5	9.6	-
Qwen2-Audio [13]		✓	✓		48.3	38.7	51.3	34.9	27.0	40.2	38.7	39.6	37.1	9.5	10.6	2.4
DeSTA2.5-Audio [40]		✓	✓		50.4	41.2	52.4	37.6	28.4	42.4	40.3	42.2	39.6	9.8	11.0	4.1
Emotion-LLaMa [10]	✓	✓	✓		53.4	45.2	51.5	37.6	19.4	43.8	44.1	42.8	38.9	7.8	11.0	4.2
Emotion-LLaMa [10]	✓	✓	✓		66.0	53.2	64.4	46.4	43.0	69.6	72.2	69.0	62.7	19.5	28.2	22.2
E3-LLaMa [19]	✓	✓	✓		54.8	42.6	52.8	40.9	35.3	40.2	43.8	43.3	37.1	10.8	14.4	6.0
E3-LLaMa [19]	✓	✓	✓		65.2	53.2	64.4	49.3	42.5	73.1	76.8	73.5	65.1	20.8	28.2	23.8
AffectGPT [35]	✓	✓	✓		51.4	42.0	54.0	37.6	28.7	42.8	40.7	42.6	40.8	10.2	11.5	4.8
AffectGPT [35]	✓	✓	✓		66.7	53.2	67.5	47.6	43.5	71.3	74.0	76.1	69.2	20.8	28.2	24.4
OSIRIS (Ours)	✓	✓	✓		75.3	60.5	74.7	56.7	47.3	86.5	89.2	86.4	80.9	25.1	35.0	33.5
OSIRIS (Ours)	✓	✓	✓	✓	**77.6**	**62.6**	**76.7**	**58.0**	**51.2**	**87.0**	**90.3**	**88.5**	**81.4**	**26.0**	**36.1**	**35.1**
OSMO XTime Evaluation Protocol (non-overlapping days in train/val):																
Baseline LLaMa3† [61]			✓		43.0	34.3	48.7	31.2	24.4	35.0	39.6	39.3	31.1	8.7	10.2	-
Qwen2-Audio [13]		✓	✓		47.8	37.7	52.4	33.5	26.0	37.2	43.1	42.2	33.8	9.3	11.0	2.6
DeSTA2.5-Audio [40]		✓	✓		50.9	41.9	57.0	33.9	27.4	38.5	45.8	43.6	36.9	9.3	11.3	4.6
Emotion-LLaMa [10]	✓	✓	✓		55.3	44.7	52.5	35.8	16.0	44.4	46.2	46.1	40.5	6.3	10.3	4.8
Emotion-LLaMa [10]	✓	✓	✓		65.1	52.6	65.7	44.7	40.1	74.0	72.2	74.3	66.3	20.3	27.1	23.3
E3-LLaMa [19]	✓	✓	✓		52.7	45.8	53.8	37.1	34.9	43.7	41.6	41.1	39.0	10.9	15.1	6.4
E3-LLaMa [19]	✓	✓	✓		65.9	54.6	67.2	44.2	41.1	74.0	73.1	70.8	67.2	21.4	27.4	23.7
AffectGPT [35]	✓	✓	✓		51.9	41.9	57.0	35.3	28.5	40.5	46.3	45.4	37.6	9.8	11.9	5.5
AffectGPT [35]	✓	✓	✓		67.4	55.9	71.2	45.3	42.6	72.3	78.5	77.0	67.2	21.4	27.8	25.5
OSIRIS (Ours)	✓	✓	✓		74.9	62.4	77.6	53.8	47.5	85.0	87.7	87.5	80.9	26.2	34.8	33.9
OSIRIS (Ours)	✓	✓	✓	✓	**78.4**	**64.2**	**79.1**	**55.2**	**50.1**	**87.1**	**90.2**	**88.5**	**81.9**	**26.7**	**35.6**	**35.6**

Table 4. **OSIRIS shows strong XLang transfer.** *LAS* reflects LLaMa Average Score over IC, DO, CU, and TUC.

Model	OVER HR	SA WAF	IP WAF	EL mIoU	ER LAS
English→ English performance					
Emotion-LLaMa [10]	57.1	63.4	48.9	35.9	71.3
E3-LLaMa [19]	59.5	60.9	44.9	35.5	71.7
AffectGPT [35]	62.2	64.8	50.5	35.0	74.2
OSIRIS (Ours)	**69.9**	**72.8**	**60.8**	**43.2**	**88.9**
English→ Mandarin transfer					
Emotion-LLaMa [10]	68.6	65.8	46.6	33.6	74.9
E3-LLaMa [19]	67.2	65.2	44.8	33.2	72.7
AffectGPT [35]	70.8	66.6	45.7	36.3	74.8
OSIRIS (Ours)	**78.7**	**74.8**	**54.4**	**43.2**	**88.5**
Mandarin→ Mandarin performance					
Emotion-LLaMa [10]	73.4	66.9	45.9	37.9	74.9
E3-LLaMa [19]	67.4	70.4	45.9	34.8	70.6
AffectGPT [35]	74.0	69.6	47.6	39.2	77.0
OSIRIS (Ours)	**85.1**	**80.9**	**54.7**	**47.2**	**90.6**
Mandarin→ English transfer					
Emotion-LLaMa [10]	56.4	53.4	45.1	31.6	74.2
E3-LLaMa [19]	53.5	51.7	46.5	29.7	72.8
AffectGPT [35]	55.5	56.3	45.8	31.6	77.3
OSIRIS (Ours)	**65.3**	**64.7**	**55.2**	**38.5**	**90.6**

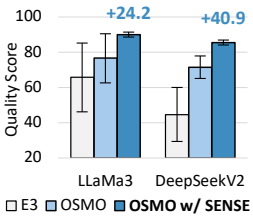


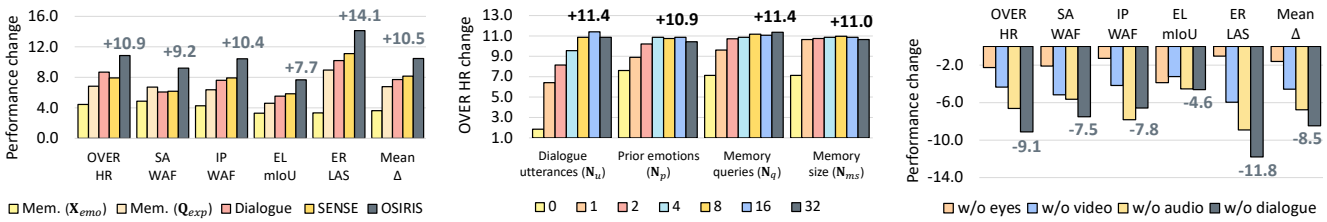
Figure 6. **Ann. quality.**

Model	E3	OSMO	
		XSub	XSet
Trained on E3 dataset [19]			
E3-LLaMa [19]	54.9	42.6	39.0
OSIRIS (Ours)	**60.4**	**46.8**	**44.4**
Trained on OSMO-Xsub			
E3-LLaMa [19]	44.2	53.2	52.8
OSIRIS (Ours)	**52.0**	**62.6**	**60.7**

Table 5. **SOS performance.**

better from English→Mandarin than the reverse, despite comparable data durations. This asymmetry stems from the higher subject diversity in the English set (282 vs. 6), underscoring that *diversity outweighs scale* in cross-cultural emotion modeling. While OSMO's 288-subject emotion timelines establish solid foundations for cultural transferability, these results highlight the need to further expand OSMO across subjects and cultures for a more universal emotion tracking. Notably, OSIRIS achieves superior cross-lingual generalization across all metrics compared to existing emotion LMMs [10, 19, 35], demonstrating the effectiveness of combining the SENSE strategy with personal emotion history for more robust and interpretable transfer.

Dataset–Model Evaluation. We compare captions from E3 [19] and OSMO using LLaMa3 [61] and DeepSeek-V2 [37] as judges, each scoring emotional clarity and ground-



(a) Component contribution over AffectGPT [35]. (b) Effect of personal history over AffectGPT [35]. (c) Effect of removing modalities.

Figure 7. **Ablation insights:** (a) All components help, with SENSE contributing most. (b) Longer personal history improves performance but saturates with larger memory size. (c) Removing any modality degrades performance, with dialogue text being the most critical.

ing (1–100). As shown in Fig. 6, both LLMs rate E3 captions as lower and less consistent, while OSMO achieves higher and more stable scores, reflecting its rigorous annotation and review process under natural affect. OSMO CoT captions with SENSE further surpass E3 by **+24.2** and **+40.9** points with LLaMa3 and DeepSeek-V2, respectively, underscoring the effectiveness of structured LLM–human collaboration. Table 5 shows that models trained on OSMO generalize better across datasets, including unseen AEA (OSMO-XSet).

6.1. Ablation Study

We ablate the contributions of OSIRIS on OSMO-XSub:

Component-wise Contributions. Figure 7a shows the performance change brought by each OSIRIS component relative to finetuned AffectGPT [35]. Adding **prior emotions** ($\mathbf{X}_{emo}$) improves the mean by +3.6, showing that modeling emotional continuity (*e.g.,* confusion → frustration) aids temporal reasoning. **Memory** queries ($\mathbf{Q}_{exp}$) add +6.8 mean by modeling emotion carry-over effects. Incorporating **dialogue** brings +7.7 mean, as conversational cues clarify ambiguous tones (*e.g.,* "really?" as surprise vs. anger). The reason-first **SENSE** strategy yields the largest single gain (+8.2 mean), enforcing structured reasoning before emotion inference. Together, these modules produce a **+10.5** mean boost, confirming that integrating temporal, contextual, and reasoning cues is crucial for robust affect understanding.

Dialogue vs. Subtext. Our analysis (Fig. 7b) demonstrates that dialogue context substantially outperforms reliance on ego subtext alone. While using only the current ego utterance (N_u=1) yields a +6.4 gain in OVER HR over the baseline [35], expanding to a fuller dialogue history (N_u=4) increases this to +8.1, with performance peaking at +11.4 for N_u=16. This shows that emotions are often disambiguated by conversational interplay. For example, an ego's statement "This is incredible" could signify either joy or despair, contingent on a partner's preceding remark ("We won!" vs. "We failed."). Consequently, broader dialogue context is critical for accurately interpreting and tracking emotion evolution.

Effect of Personal History. Figure 7b illustrates the effect of the numbers of **prior emotions** (N_p), **memory queries** (N_q), and **memory slots** (N_{ms}) on HR improvement over the baseline [35]. Increasing N_p from 0 to 4 boosts performance from 7.6 to 10.9, indicating that modeling up to four preceding emotions effectively captures temporal affect continuity without overfitting to distant states. Expanding N_q similarly improves results, peaking at 11.4 when $N_q = 32$, as richer contextual retrieval enhances temporal reasoning. In contrast, enlarging N_{ms} beyond one slot yields only minor gains (10.6→11.0), showing that OSIRIS effectively compresses historical affect into concise latent summaries.

Effect of Modality Removal. Figure 7c reveals that excluding **dialogue** text causes the largest mean drop (-8.5), as text provides the most direct emotional evidence. Excluding **audio** also substantially hurts mean performance (-6.8), since vocal cues (*e.g.,* high pitch in anger, slow tempo in sadness) convey critical paralinguistic information. Removing **video** leads to a moderate decline (-4.6), as visual context (*e.g.,* social gatherings for joy, rotten food for disgust) helps situate emotions. While omitting **eyes** shows a smaller mean decrease (-1.6), it specifically impairs fine-grained localization by 3.9 mIoU, highlighting how ocular behaviors, *e.g.,* widened eyes, offer precise spatiotemporal grounding.

7. Conclusion

We introduce **OSMO**, the first large-scale dataset for egocentric emotion tracking, and **OSIRIS**, an LMM that reasons over video, audio, eyes, dialogue text, and emotional context to track evolving affective states. OSMO reframes emotion perception as a temporally coherent, open-vocabulary task, and OSIRIS sets a new SOTA across recognition, localization, and reasoning benchmarks. Together, they establish a new paradigm for continuous, context-aware, and interpretable emotion tracking from wearable sensors.

Limitations. Despite its scale, OSMO focuses on daily social interactions and may underrepresent extreme emotions or cultural diversity. Moreover, benchmark results indicate that emotion understanding from egocentric signals remains far from solved. Future extensions could expand to broader contexts, languages, and physiological modalities to further enrich the spectrum of human emotion understanding.

References

[1] Mohamed O Abdelfattah, Kaouther Messaoud, and Alexandre Alahi. OSKAR: Omnimodal self-supervised knowledge abstraction and representation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[2] Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, Sahir Gomez, Hongyu Gong, Srivathsan Govindarajan, Brandon Han, Sen He, Denise Hernandez, Yordan Hristov, Rongjie Huang, Hirofumi Inaguma, Somya Jain, Raj Janardhan, Qingyao Jia, Christopher Klaiber, Dejan Kovachev, Moneish Kumar, Hang Li, Yilei Li, Pavel Litvin, Wei Liu, Guangyao Ma, Jing Ma, Martin Ma, Xutai Ma, Lucas Mantovani, Sagar Miglani, Sreyas Mohan, Louis-Philippe Morency, Evonne Ng, Kam-Woh Ng, Tu Anh Nguyen, Amia Oberai, Benjamin Peloquin, Juan Pino, Jovan Popovic, Omid Poursaeed, Fabian Prada, Alice Rakotoarison, Alexander Richard, Christophe Ropers, Safiyyah Saleem, Vasu Sharma, Alex Shcherbyna, Jia Shen, Jie Shen, Anastasis Stathopoulos, Anna Sun, Paden Tomasello, Tuan Tran, Arina Turkatenko, Bo Wan, Chao Wang, Jeff Wang, Mary Williamson, Carleigh Wood, Tao Xiang, Yilin Yang, Zhiyuan Yao, Chen Zhang, Jiemin Zhang, Xinyue Zhang, Jason Zheng, Pavlo Zhyzheria, Jan Zikes, and Michael Zollhoefer. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. 2025. 1, 2, 4

[3] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 5

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 4

[5] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018. 2

[6] 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. 2

[7] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 2, 3

[8] Chih-Yao Chen, Tun Min Hung, Yi-Li Hsu, and Lun-Wei Ku. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958, 2023. 2

[9] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1254–1262, 2024. 6

[10] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024. 1, 2, 3, 4, 6, 7

[11] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Li Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *ArXiv*, abs/2406.07476, 2024. 5

[12] Bo-Cheng Chiu, Jen-Jee Chen, Yu-Chee Tseng, and Feng-Chi Chen. Damo: A data-efficient multimodal orchestrator for temporal reasoning with video llms. *arXiv preprint arXiv:2506.11558*, 2025. 5

[13] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. 2, 3, 6, 7

[14] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114 (38):E7900–E7909, 2017. 3

[15] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, 2020. Association for Computational Linguistics. 3

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

[17] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 6

[18] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1

[19] Yueying Feng, WenKang Han, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, Jingyuan Chen, et al. E3: Exploring embodied emotion through a large-scale egocentric video dataset. *Advances in Neural Information Processing Systems*, 37:118182–118197, 2024. 1, 2, 3, 4, 6, 7

[20] Yingxue Gao, Huan Zhao, and Zixing Zhang. Adaptive speech emotion representation learning based on dynamic graph. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1116–11120. IEEE, 2024. 2

[21] 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):8–16, 2011. 5

[22] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. Eyemotion: Classifying facial expressions in vr using eye-tracking cameras. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1626–1635. IEEE, 2019. 2, 5

[23] Leo Ho, Yinghao Huang, Dafei Qin, Mingyi Shi, Wangpok Tse, Wei Liu, Junichi Yamagishi, and Taku Komura. Interact: A large-scale dataset of dynamic, expressive and interactive activities between two people in daily scenarios. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(4):1–27, 2025. 1, 2, 4

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[25] Ziyuan Huang, Kaixiang Ji, Biao Gong, Zhiwu Qing, Qinglong Zhang, Kecheng Zheng, Jian Wang, Jingdong Chen, and Ming Yang. Accelerating pre-training of multimodal llms via chain-of-sight. *Advances in Neural Information Processing Systems*, 37:75668–75691, 2024. 6

[26] Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, et al. Storm: Token-efficient long video understanding for multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5830–5841, 2025. 5

[27] Sylvia Deidre Kauer, Sophie Caroline Reid, Alexander Hew Dale Crooke, Angela Khor, Stephen John Charles Hearps, Anthony Francis Jorm, Lena Sanci, and George Patton. Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. *Journal of medical Internet research*, 14(3):e1858, 2012. 1

[28] D Kollias and S Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. arxiv 2018. *arXiv preprint arXiv:1811.07770*, 2, 2018. 2

[29] Georgia Kouri, Nathalie Meuwly, Marianne Richter, and Dominik Schoebi. Attachment insecurities, emotion dynamics and stress in intimate relationships during the transition to parenthood. *BMC psychology*, 12(1):200, 2024. 5

[30] Philip A Kragel, Ahmad R Hariri, and Kevin S LaBar. The temporal dynamics of spontaneous emotional brain states and their implications for mental health. *Journal of cognitive neuroscience*, 34(5):715–728, 2022.

[31] Peter Kuppens and Philippe Verduyn. Emotion dynamics. *Current Opinion in Psychology*, 17:22–26, 2017. 5

[32] Bokyeung Lee, Hyunuk Shin, Bonhwa Ku, and Hanseok Ko. Frame level emotion guided dynamic facial expression recognition with emotion grouping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5681–5691, 2023. 2

[33] Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. Skier: A symbolic knowledge integrated model for conversational emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13121–13129, 2023. 2

[34] Zheng Lian, Haiyang Sun, Licai Sun, Lan Chen, Haoyu Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, et al. Open-vocabulary multimodal emotion recognition: Dataset, metric, and benchmark. *ICML*, 2024. 1, 2, 4

[35] 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. *ICML (Oral, Top* 1%*)*, 2025. 1, 2, 3, 4, 6, 7, 8

[36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 4

[37] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024. 7

[38] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 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*, pages 24–32, 2022. 1, 2, 4

[39] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 2

[40] Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, et al. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*, 2025. 2, 3, 6, 7

[41] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria everyday activities dataset, 2024. 2, 3, 4

[42] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *the 18th European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[43] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 4

[44] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial

expressions collected. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 881–888, 2013. 6

[45] 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. 2

[46] Matthias R Mehl. The electronically activated recorder (ear) a method for the naturalistic observation of daily social behavior. *Current directions in psychological science*, 26(2): 184–190, 2017. 3

[47] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 6

[48] Fei Ni, Jianye Hao, Shiguang Wu, Longxin Kou, Jiashun Liu, Yan Zheng, Bin Wang, and Yuzheng Zhuang. Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13991–14000, 2024. 6

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 4

[50] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980. 3

[51] 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*, pages 527–536, Florence, Italy, 2019. Association for Computational Linguistics. 1, 2, 4

[52] Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13492–13500, 2023. 2

[53] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 3, 5

[54] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 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. IEEE, 2013. 2

[55] Vera Sacharin, Katja Schlegel, and Klaus R Scherer. Geneva emotion wheel rating study. *Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences. Aalborg University, Aalborg*, 2012. 3

[56] Stephen R Schmidt and Constance R Schmidt. The emotional carryover effect in memory for words. *Memory*, 24(7):916–938, 2016. 1, 2

[57] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 5

[58] Joshua M Smyth, Jillian A Johnson, Brandon J Auer, Erik Lehman, Giampaolo Talamo, and Christopher N Sciamanna. Online positive affect journaling in the improvement of mental distress and well-being in general medical patients with elevated anxiety symptoms: A preliminary randomized controlled trial. *JMIR mental health*, 5(4):e11290, 2018. 1

[59] Stefan Stieger, Selina Volsa, David Willinger, David Lewetz, and Bernad Batinic. Laughter in everyday life: an event-based experience sampling method study using wrist-worn wearables. *Frontiers in Psychology*, 15:1296955, 2024. 3

[60] Arielle Tambini, Ulrike Rimmele, Elizabeth A Phelps, and Lila Davachi. Emotional brain states carry over and enhance future memory formation. *Nature neuroscience*, 20(2):271–278, 2017. 1, 2

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 3, 4, 5, 6, 7

[62] Alberta SJ Van der Watt, W Odendaal, K Louw, and Soraya Seedat. Distant mood monitoring for depressive and bipolar disorders: a systematic review. *BMC psychiatry*, 20(1):383, 2020. 1

[63] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17958–17968, 2023. 2

[64] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European conference on computer vision*, pages 700–717. Springer, 2020. 3

[65] Yan Wang, Shaoqi Yan, Yang Liu, Wei Song, Jing Liu, Yang Chang, Xinji Mai, Xiping Hu, Wenqiang Zhang, and Zhongxue Gan. A survey on facial expression recognition of static and dynamic emotions. *arXiv preprint arXiv:2408.15777*, 2024. 2

[66] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025. 6

[67] Gloria Willcox. The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Transactional Analysis Journal*, 12(4):274–276, 1982. 3

[68] Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. Emo: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In *Proceedings of the 18th*

International Conference on Mobile Systems, Applications, and Services, pages 448–461, 2020. 2, 5

[69] Xuecheng Wu, Heli Sun, Junxiao Xue, Jiayu Nie, Xiangyan Kong, Ruofan Zhai, Danlei Huang, and Liang He. Towards emotion analysis in short-form videos: A large-scale dataset and baseline. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, page 1497–1506, New York, NY, USA, 2025. Association for Computing Machinery. 2

[70] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[71] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28885–28900, 2025. 2, 3, 6

[72] Yuting Yang, Jingyi Wang, Haijiang Lin, Xiaoxiao Chen, Yun Chen, Jiawen Kuang, Ye Yao, Tingting Wang, and Chaowei Fu. Emotion dynamics prospectively predict depressive symptoms in adolescents: findings from intensive longitudinal data. *BMC psychology*, 13(1):1–12, 2025. 5

[73] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018. 2

[74] Thea Zander-Schellenberg, Isabella Mutschler Collins, Marcel Miché, Camille Guttmann, Roselind Lieb, and Karina Wahl. Does laughing have a stress-buffering effect in daily life? an intensive longitudinal study. *Plos one*, 15(7): e0235851, 2020. 3

[75] Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in neural information processing systems*, 37:110935–110971, 2024. 6

[76] Haiwei Zhang, Jiqing Zhang, Bo Dong, Pieter Peers, Wenwei Wu, Xiaopeng Wei, Felix Heide, and Xin Yang. In the blink of an eye: Event-based emotion recognition. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 5

[77] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901, 2025. 5