LONGEMO: TOWARDS EMOTION UNDERSTANDING AND REASONING IN LONG VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

While recent Multimodal Large Language Models (MLLMs) have shown promise in affective computing, their reasoning capabilities are largely confined to short video clips with limited interactions. However, real-world emotions are not merely isolated instantaneous reactions but dynamic and cumulative processes deeply shaped by past experiences and ongoing events. To bridge this gap, we introduce LongEmoBench, a benchmark dedicated to emotion understanding and reasoning in long videos. It assesses progressive capabilities scaling from continuous scene interactions to complex episodic developments. Furthermore, we propose LongEmo, a novel memory-augmented agentic framework designed to tackle the immense challenges of long-range affective reasoning. LongEmo processes continuous video streams to construct an Event Memory Graph, explicitly modeling long-range dependencies and capturing emotional dynamics across discrete events. Given a question, the agent retrieves a query-relevant event stream from the graph, iteratively integrating multimodal memories and relational dependencies to deduce the final answer. Extensive evaluations of 17 representative methods reveal that they struggle significantly with emotion understanding and reasoning in long videos. In contrast, LongEmo achieves state-of-the-art performance, demonstrating the efficacy of its event-centric memory architecture.

1 INTRODUCTION

Understanding human emotions is a cornerstone for developing human-centric AI. Recent advances in Multimodal Large Language Models (MLLMs) (Cheng et al., 2024; Lian et al., 2025a; Zhao et al., 2025a) have spurred substantial progress in affective computing. Beyond recognizing isolated expressions (Lian et al., 2025d) from facial muscle movements, vocal acoustics, and lexical semantics, modern models can explain the immediate causes of emotional reactions (Zhang et al., 2025b), and infer affective shifts within short local dialogues (Hu et al., 2026a).

However, existing methods largely confine reasoning to extremely short, isolated video clips. In real-world scenarios, emotion is rarely a simple slice of a transient state, but a dynamic process accumulated over time (Scherer, 2009). A character's present feelings are deeply influenced by past experiences, may be deliberately masked by superficial behavior, and are continuously redefined as events unfold. Accurately inferring such complex states requires traversing the temporal dimension to integrate multimodal clues scattered across long-range contexts. Existing affective benchmarks (Lian et al., 2025b; Hu et al., 2026b; Zhang et al., 2026) almost exclusively remain at the level of short clips, single scenes, or single conversations, where the emotional dynamics they capture are restricted to, at most, immediate expressive reactions. This naturally raises the question: *are current MLLMs and video agents capable of performing long-range emotion understanding and reasoning across extended video narratives?*

Although general long-video benchmarks (Fu et al., 2025; Wang et al., 2025; Yang et al., 2026) successfully scale inputs to hour-long durations, they primarily assess physical realities and factual plot points rather than characters' internal states, leaving long-term emotion understanding and reasoning largely unaddressed. To bridge this gap, we introduce **LongEmoBench**, a comprehensive benchmark dedicated to emotion understanding and reasoning in long videos. It comprises 1,975 high-quality QA pairs across approximately 70 hours of video footage, with individual videos spanning from short contextual scenes to nearly hour-scale continuous narratives, far exceeding the temporal

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

Table 1: Comparison of LongEmoBench with existing affective video benchmarks. **Avg. Duration** denotes the average duration of the video provided for each question, and **Total Hours** denotes the total duration of annotated videos.

Benchmark	# Tasks	# Videos	# QA	Avg. Duration	Total Hours
OV-MER (Lian et al., 2025d)	1	332	332	4.0 s	0.4 h
EmoTrans (Hu et al., 2026a)	4	1,000	3,274	5.5 s	1.5 h
MtMeUR (Hu et al., 2025)	5	1,451	5,101	17.3 s	6.8 h
MME-Emotion (Zhang et al., 2026)	8	6,500	6,500	4.8 s	8.6 h
EmoBench-M (Hu et al., 2026b)	13	5,642	5,646	7.3 s	11.9 h
LongEmoBench (Scene)	5	1,213	1,417	**1.3 min**	**23.9 h**
LongEmoBench (Episode)	3	141	558	**20.0 min**	**45.9 h**

span of existing benchmarks (See Table 1). We systematically decompose the evaluation into two complementary granularities: Scene-level and Episode-level. Scene-level assesses intra-scene affective comprehension, evaluating standard MLLMs within limited context windows. Episode-level specifically challenges long-context models and advanced video agents, requiring them to synthesize temporally distant evidence for emotion tracking and complex reasoning.

To tackle the immense challenges posed by long-term emotion reasoning, we propose **LongEmo**, a novel memory-augmented agentic framework. Motivated by the observation that a video fundamentally unfolds as a dynamic event stream within a persistent world, we formulate long-term emotion reasoning as modeling affective dynamics across discrete events. Accordingly, LongEmo distills the untrimmed video into structured event memories, encapsulating characters' emotional states and supporting multimodal evidence. To capture long-range dependencies, we construct a global Event Memory Graph by interconnecting these event memories based on the temporal and causal logic driving emotional dynamics. Given a user query, LongEmo retrieves a relevant event stream from the graph to establish a character-centric emotional context. Subsequently, the reasoning agent executes a progressive access loop, seamlessly integrating multimodal memories within individual events and relational dependencies across the event stream to logically deduce the final answer.

Using LongEmoBench, we evaluate 17 approaches on LongEmoBench, including 9 standard MLLMs and 8 long-video understanding methods (long-video models and agent-based methods). Evaluations reveal that while standard MLLMs capture immediate emotional cues, they struggle with complex causal reasoning and subtle state transitions. Furthermore, although long-video methods handle general tasks adequately, they fall short in emotion reasoning, failing to model long-range emotional dynamics. In contrast, LongEmo achieves state-of-the-art (SOTA) performance, validating the effectiveness of its event-centric memory architecture.

2 Related Work

Multimodal Emotion Benchmarks. The rapid advancement of MLLMs has fundamentally reshaped benchmarking in affective computing. This evaluation paradigm has continuously evolved, progressing from early closed-set classification (Lian et al., 2024b; Hu et al., 2026b) to open-vocabulary description (Lian et al., 2024a; 2025d), and extending to causal explanation (Lian et al., 2025a; Zhang et al., 2026). To further challenge modern models, recent pioneering works have expanded into more complex interactive scenarios, introducing benchmarks that evaluate emotion transitions (Hu et al., 2026a), multi-turn and multi-party affective dialogues (Hu et al., 2025; Sasu et al., 2025), and even model robustness against emotional hallucinations (Xing et al., 2025). However, the video inputs in these studies remain restricted to brief scenes or single interactions, providing only immediate context and thus failing to evaluate a model's ability to capture long-range dependencies and perform complex affective reasoning. In contrast, LongEmoBench transcends these temporal limitations, connecting fine-grained audiovisual cues with extended temporal contexts to push the boundaries of affective computing into the true long-video regime.

Long-Video Understanding Methods. To process hour-level audiovisual sequences, recent approaches typically tackle long-context bottlenecks through two primary paradigms: agentic frameworks and architectural optimizations. The first paradigm circumvents context constraints by for-

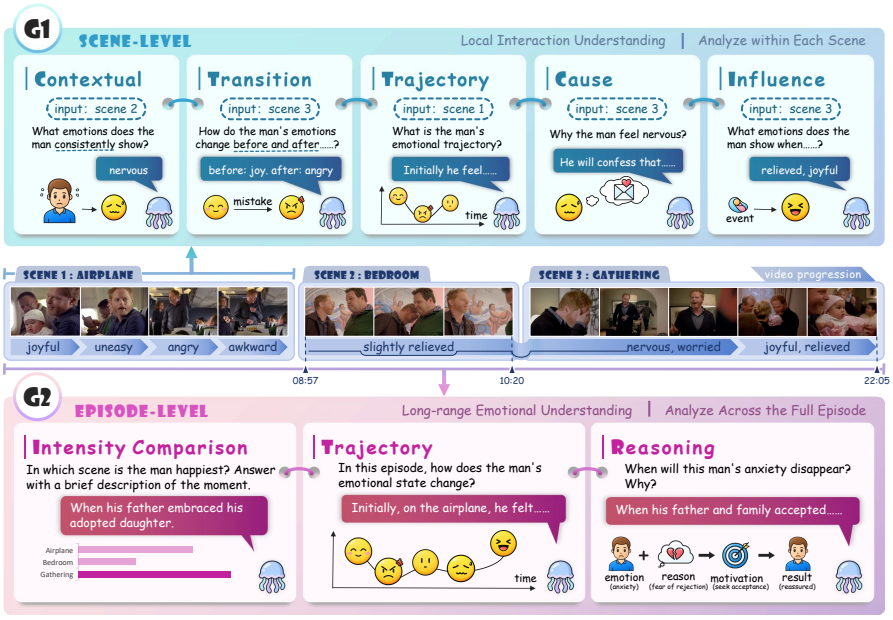


Figure 1: **Overview of our LongEmoBench.** The G1 (Scene-Level) includes five tasks for localized emotional interactions. The G2 (Episode-Level) includes three tasks for long-range emotional reasoning. A timeline illustrates the temporal scope distinguishing these two granularities.

mulating temporal modeling as an active, agent-driven exploration process. Memory-augmented systems like M3-agent (Long et al., 2025), WorldMM (Yeo et al., 2026), and Light-Omni (Nie et al., 2026) construct explicit long-term memory banks to iteratively index and recall historical semantic events. Alternatively, autonomous vision agents such as PyVision-RL (Zhao et al., 2026) bypass exhaustive video processing by selectively sampling temporal context strictly on demand. The second paradigm mitigates computational overhead directly at the structural level. Token compression strategies (Cao et al., 2026; Chen et al., 2026) apply query-guided or retrieval-inspired algorithms to drastically reduce spatial-temporal redundancy, whereas streaming architectures (Xiao et al., 2024; Lin et al., 2026) process inputs sequentially to maintain stable long-term reasoning. However, while these methods excel at general long-video reasoning, they risk compromising affective computing by disrupting continuous emotional trajectories, discarding fine-grained cues, and losing distant emotional anchors. To bridge this gap, we introduce LongEmoBench as a vital testbed to evaluate whether these architectures can overcome these bottlenecks and truly master long-term affective reasoning. Concurrently, we propose a novel event-centric memory method explicitly designed to address these limitations and preserve continuous emotional dynamics.

3 LongEmoBench

Figure 1 presents an overview of LongEmoBench. It is designed to systematically evaluate whether current MLLMs and agentic frameworks can perform precise emotion understanding and reasoning across extended video contexts. Below, we introduce the benchmark's foundational design and evaluation protocol, while extended details are provided in Appendix.

3.1 Task Taxonomy

LongEmoBench structures its evaluation through a hierarchical task taxonomy grounded in narrative granularity. We organize this taxonomy into two primary tiers, Scene-Level (G1) and Episode-Level (G2), to measure progressive reasoning capabilities that scale from contiguous interpersonal interactions to comprehensive cross-scene narratives. Examples for each task are provided in Appendix A.

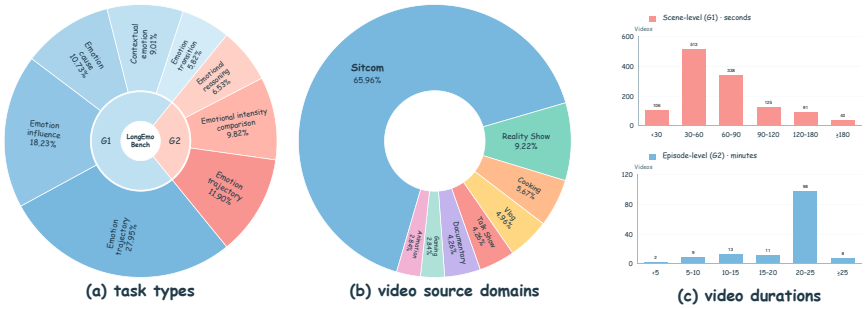


Figure 2: **Statistical overview of LongEmoBench.** (a) Task type distribution. (b) Video source domain distribution. (c) Video duration distributions for scene and episode levels.

Scene-Level. This tier models the complete lifecycle of emotional events within contiguous interactive scenes, mirroring how humans progressively perceive, track, and interpret affective dynamics. Rather than isolated categories, we design five interdependent tasks to provide comprehensive coverage of this cognitive process. *Contextual Emotion* serves as the affective baseline, identifying persistent individual states or collective atmospheres. To track temporal evolution, *Emotion Transition* captures immediate affective shifts triggered by specific stimuli, while *Emotional Trajectory* reconstructs continuous, multi-stage emotional developments. Finally, to unravel the underlying mechanisms driving these dynamics, *Emotion Cause* traces retrospectively to explain psychological motives and situational triggers, whereas *Emotion Influence* projects prospectively to assess how explicit actions elicit subsequent interpersonal reactions.

Episode-Level. This tier evaluates long-range reasoning across full episodes or untrimmed videos, designed to assess the ability to connect temporally distant emotional cues and comprehend overarching emotional arcs. It comprises three core tasks: *Emotional Intensity Comparison* requires locating temporal extrema (e.g., peak tension) or contrasting emotional magnitude across different characters, narrative moments, or emotionally relevant target subjects. *Emotional Trajectory* tracks a specific character's prolonged emotional progressions and intensity dynamics spanning multiple disparate narrative scenes. Lastly, *Emotional Reasoning* integrates temporally dispersed audiovisual cues across the video to either deduce discrete factual outcomes (such as reaction frequencies) or provide causal explanations for characters' deep-seated motives and long-term behavioral intentions.

3.2 Dataset Construction and Statistics

To systematically construct LongEmoBench, we employ a human-in-the-loop annotation pipeline beginning with question-driven video collection. Annotators are instructed to curate long-form videos—ranging from narrative episodes to open-domain web videos—that feature rich interpersonal dynamics, pronounced emotional shifts, and sufficient narrative depth to support complex reasoning. Once a video is selected, annotators perform fine-grained audiovisual event segmentation to curate G1 question-answer pairs. To accurately capture diverse emotional dimensions, annotators employ a dual-format annotation strategy. Specifically, for tasks identifying specific emotional states (e.g., Contextual Emotion and Emotion Transition), they strictly map the emotions to the 26-class EMOTIC (Kosti et al., 2019) taxonomy, providing a standardized and granular vocabulary for complex expressions. Conversely, for explanatory and continuous tasks (e.g., Emotion Cause and Emotion Trajectory), annotators craft natural language descriptions to articulate nuanced psychological motives and multi-stage emotional developments. Building upon these G1 annotations, annotators connect distant events across the full video to formulate G2 reasoning queries, such as tracking prolonged trajectories or comparing emotional intensity. To ensure data quality, each query is generated by one annotator and strictly validated by two independent reviewers, with any cases failing to reach a consensus strictly discarded. Further annotation details are provided in Appendix C.

Following this rigorous process, LongEmoBench comprises 1,354 videos totaling 69.8 hours of continuous audiovisual content, yielding 1,975 high-quality reasoning queries as demonstrated in Table 1. To emphasize long-form understanding, the benchmark features extended video contexts where the G2 subset comprises full narrative episodes averaging 19.5 minutes per video. Addi-

tionally, the dataset encompasses diverse real-world content spanning sitcoms, open-domain vlogs, reality shows, live broadcasts, and various other in-the-wild formats. As illustrated in Figure 2, we report the statistical distributions of video durations, source domains, and 8 distinct task types.

3.3 Evaluation Protocol

Label-based Evaluation. For Contextual Emotion, Emotion Transition, and Emotion Influence, we evaluate the predicted emotional categories using the sample-averaged F1-score, following previous MER challenges (Lian et al., 2025c). Notably, for Emotion Transition, the F1-score is computed separately for the initial and final states and then averaged.

Rubric-based Evaluation. For all remaining tasks, we employ DeepSeek-V4-Flash (DeepSeek-AI, 2026) as an LLM judge to assess responses against reference answers and task-specific rubrics. Emotional Intensity Comparison and deterministic questions in Emotional Reasoning receive binary correctness scores, with semantically equivalent answers accepted. Emotional Trajectory questions receive a holistic score from 0 to 4 based on the correctness and completeness of the required emotional stages and their relationships. Questions requiring causal explanations for emotions receive a holistic score from 0 to 3, assessing explanatory correctness, sufficiency, and substantive causal errors. Scoring criteria strictly depend on the question's reasoning requirements rather than response length. To validate the reliability of this automated evaluator, we report a human-LLM correlation (Pearson's $r = 0.872$) on a sampled subset. Full rubrics are provided in Appendix D.

4 Our Method

We introduce LongEmo, a memory-augmented framework that processes long video streams in two stages: (1) event-centric memory construction, and (2) event-stream retrieval and reasoning. As illustrated in Figure 3, we first construct event-centric memory nodes, explicitly grounding distinct emotional states in their supporting audiovisual observations and historical context (Sec 4.1). For a given query, we then retrieve a temporally ordered event stream and progressively access this evidence to execute long-range emotion reasoning and generate the answer (Sec 4.2).

4.1 Event-Centric Memory Construction

Fundamentally, a video unfolds as a dynamic event stream within a persistent world. Human emotional states, in turn, are inherently anchored to these discrete events. Inspired by this observation, we construct our memory at the event level. To achieve this, we first instantiate discrete events as memory nodes to ground localized emotions, and subsequently construct an event graph to model long-range affective dynamics.

Event Memory. We formulate the global event memory as a collection of event nodes, $\mathcal{M} = \{m_1, m_2, \dots\}$, with each node defined as:

$$m_i = (d_i, \mathcal{S}_i, \mathcal{O}_i) \tag{1}$$

where d_i is a textual summary of the event, $\mathcal{S}_i = \{(p_j, s_j, r_j)\}_{j=1}^{N_i}$ stores N_i discrete emotion records, each capturing a participant p_j's inferred emotional state s_j and its initial causal explanation r_j, and $\mathcal{O}_i$ retains the supporting multimodal evidence (e.g., facial expressions, vocal prosody, and dialogue). Preserving these fine-grained raw cues is critical, as it enables the reasoning agent to subsequently reassess initial interpretations under a broader context (detailed in Sec. 4.2).

Memory Generation and Update. Given a long video, we process it sequentially as a series of temporal segments $\{v_t\}_{t=1}^{T}$. At step t, a multimodal perception module processes the current segment v_t by leveraging an accumulated entity registry $\mathcal{P}_{t-1}$ to ensure consistent participant identification, alongside a recent memory subset $\mathcal{H}_{t-1} \subseteq \mathcal{M}_{t-1}$ to maintain event continuity. This process yields K_t candidate event memories paired with their assignment variables $z_{t,k}$:

$$f_{\text{perc}}(v_t, \mathcal{P}_{t-1}, \mathcal{H}_{t-1}) = \{(m_{t,k}, z_{t,k})\}_{k=1}^{K_t} \tag{2}$$

Here, $z_{t,k}$ governs how each candidate is incorporated. If $z_{t,k} = \varnothing$, the candidate $m_{t,k}$ instantiates a new event node. Conversely, if $z_{t,k} = i$, it triggers an update to an ongoing event $m_i \in \mathcal{H}_{t-1}$.

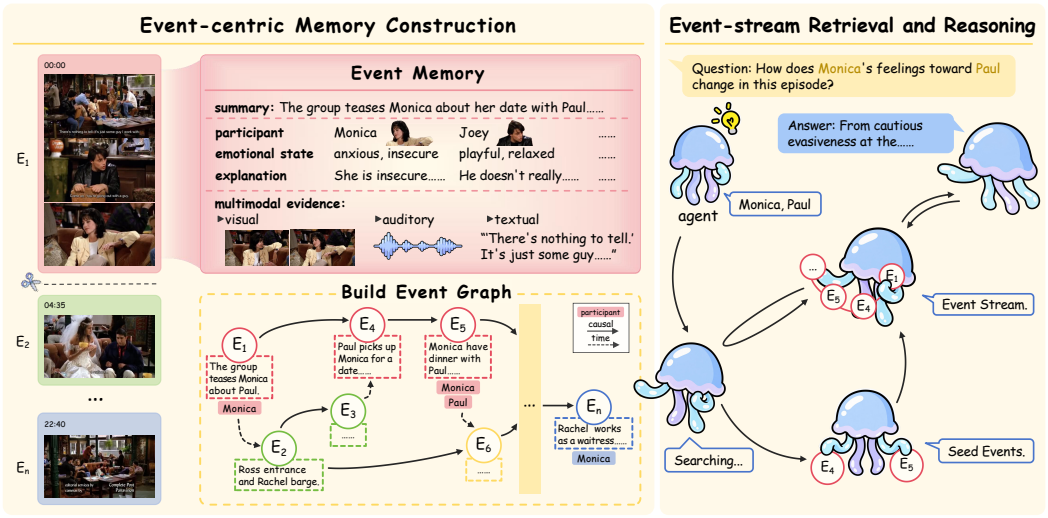


Figure 3: **Overview of our LongEmo.**

During an update, the candidate's description is synthesized with the existing summary, and the new multimodal evidence is aggregated. Critically, the update mechanism for the emotion state records is designed to address complex affective dynamics through two distinct operations: *sequential appending* and *retrospective correction*. For genuine emotional transitions (e.g., a character shifting from "anxiety" to "relief"), new states are sequentially appended to preserve the emotional progression. Conversely, when subsequent context invalidates a prior interpretation—for instance, revealing that an initial "happy" smile was actually a facade for "disappointment"—the framework applies a retrospective correction to overwrite the outdated state. This context-aware revision ensures the event memory reliably reflects the underlying emotional truth rather than superficial illusions.

Event Memory Graph Construction. Since emotional developments are often sparsely distributed across a full episode, we construct a global event memory graph over $\mathcal{M}_T$ to capture long-range dependencies. To establish edges for each event m_i, the goal is to identify preceding events that provide essential historical context. However, directly querying with m_i tends to retrieve analogous events. We therefore derive a targeted query q_i to probe the preceding circumstances behind it. In addition, the narrative typically unfolds through events centered on recurring characters, making shared participants a strong indicator of cross-event relevance. To this end, we retrieve historical events by jointly considering query relevance and character continuity. Formally, this retrieval process is formulated as:

$$s(m_h, m_i) = \mathrm{rel}(q_i, m_h) + \lambda_p \frac{|\mathcal{P}_i \cap \mathcal{P}_h|}{|\mathcal{P}_i \cup \mathcal{P}_h|}, \quad h < i. \tag{3}$$

where $\mathrm{rel}(q_i, m_h)$ evaluates the relevance between query q_i and candidate event m_h via a hybrid retrieval scheme combining dense semantic embeddings with BM25 lexical matching, $\mathcal{P}_i$ and $\mathcal{P}_h$ denote the character sets of m_i and m_h, respectively, and λ_p is a trade-off coefficient. Preceding events are accordingly ranked by $s(m_h, m_i)$ to form a candidate pool $\mathcal{C}_i$. To eliminate spurious dependencies, a relation verifier f_{ver} subsequently inspects each candidate in $\mathcal{C}_i$ to establish valid dependencies, yielding the final event memory graph:

$$\mathcal{G} = (\mathcal{M}_T, \mathcal{R}), \qquad \mathcal{R} = \bigcup_{i=1}^{T} \{e_{h \to i} \mid m_h \in \mathcal{C}_i,\ f_{\mathrm{ver}}(m_h, m_i) = 1\}. \tag{4}$$

4.2 Event-Stream Retrieval and Reasoning

Directly injecting the entire event memory graph $\mathcal{G}$ into the reasoning model introduces prohibitive computational redundancy and context noise. To address this, we first retrieve a focused set of seed

Table 2: G1 results (%) on LongEmoBench. Rubric scores for specific tasks (e.g., 0–4 for Trajectory and 0–3 for Cause) are converted to percentages for uniform comparison. V, A, and T denote visual, audio, and textual inputs, respectively. Overall denotes the weighted average.

Method	V	A	T	Contextual	Transition	Trajectory	Cause	Influence	Overall
Affective MLLMs									
R1-Omni	✓	✓	✓	0.05	0.00	12.05	2.20	1.38	5.38
HumanOmni	✓	✓	✓	1.40	0.00	12.23	5.82	13.01	9.12
AffectGPT	✓	✓	✓	13.20	15.65	16.52	12.71	15.97	15.32
Emotion-LLaMA	✓	✓	✓	9.51	12.00	13.59	9.75	24.03	15.03
General-Purpose MLLMs									
Qwen2-Audio-7B	–	✓	✓	5.57	1.68	22.92	27.67	8.81	16.14
Qwen3-VL-8B	✓	–	✓	31.52	29.57	45.43	60.06	33.97	41.67
Qwen3-Omni-30B	✓	✓	✓	36.38	38.61	47.83	68.24	40.43	46.82
InternVL3.5-38B	✓	–	✓	36.98	34.26	37.18	55.97	40.52	40.58
Gemini-3-Flash	✓	✓	✓	**59.07**	**57.89**	**63.41**	**87.58**	**58.36**	**64.75**
Long-Video Architectures									
LongVU	✓	–	✓	8.06	19.90	24.86	34.91	19.96	22.61
Flash-VStream-7B	✓	–	✓	5.28	0.00	13.59	24.84	7.09	11.47
VideoChat-Flash-7B	✓	–	✓	31.86	24.12	32.20	38.52	34.43	33.01
VideoChat3-4B	✓	–	✓	35.43	35.92	38.77	49.21	36.75	39.17

events and expand them along the graph edges into a chronologically ordered event stream, which serves as a structured scaffold for progressive reasoning.

Seed Event Retrieval. Given a question q, we identify high-confidence entry events by applying the same hybrid retrieval and character continuity scheme defined in Sec. 4.1. Specifically, we extract queried characters $\mathcal{P}_q$ from q and rank all events in $\mathcal{M}_T$ by jointly scoring text relevance and Jaccard character similarity with trade-off coefficient λ_q. The top-K candidates are retained to form the initial seed set, establishing temporal anchors along the video timeline.

Relational Event-Stream Expansion. While seed events provide reliable entry points, they remain isolated moments that cannot capture emotional dynamics on their own. We therefore expand the seed set by retrieving connected events along relational edges in the event memory graph. The expanded event set is then sorted chronologically into an ordered event stream. To respect the reasoning model's context budget, we exclude raw multimodal signals and provide only each event's timestamp, textual summary, participating characters, and relational links as a lightweight overview.

Progressive Multimodal Reasoning. With the lightweight event stream providing global context, the model conducts targeted multimodal inspection rather than indiscriminately processing the entire video. It first scans chronological text summaries to trace character interactions and locate critical turning points relevant to the query. For these pivotal moments, the model selectively unpacks their fine-grained emotional annotations alongside the corresponding audiovisual clips. By integrating local multimodal cues, such as vocal tone and facial expressions, with the global narrative trajectory, the model resolves emotional ambiguities and derives a grounded, context-aware answer.

5 Experiments

Experimental Setup. For G1 tasks, we assess general-purpose MLLMs (Qwen3-VL-8B, Qwen2-Audio-7B, Qwen3-Omni-30B, InternVL3.5-38B, Gemini-3-Flash), affective MLLMs (R1-Omni (Zhao et al., 2025a), HumanOmni (Zhao et al., 2025b), AffectGPT (Lian et al., 2025a), Emotion-LLaMA (Cheng et al., 2024)), and long-video architectures (LongVU (Shen et al., 2024), Flash-VStream (Zhang et al., 2025a), VideoChat-Flash (Li et al., 2025), VideoChat3 (Li et al., 2026)). For G2 tasks, we evaluate the long-video architectures alongside agentic systems (VideoHV (Wang et al., 2026), LongVideoAgent (Liu et al., 2025), M3-Agent (Long et al., 2025), WorldMM (Yeo et al., 2026)). All implementation details are provided in the Appendix E.

Table 3: G2 results (%) on LongEmoBench. Task-specific rubric scores are converted to percentages for uniform comparison. EIC: Emotional Intensity Comparison; ET: Emotional Trajectory; ER: Emotional Reasoning. Overall denotes the weighted average.

Method	EIC	ET	ER	Overall
Long-Video Architectures				
LongVU	1.55	23.94	20.75	15.42
Flash-VStream-7B	4.64	9.89	8.48	7.74
VideoChat-Flash-7B	1.55	30.64	21.46	18.40
VideoChat3-4B	10.31	30.21	26.46	22.42
Agent Methods				
VideoHV	2.06	34.89	38.50	24.31
LongVideoAgent	7.73	21.91	11.89	14.67
M3-Agent	8.76	31.06	28.94	22.82
WorldMM	29.38	51.49	58.66	45.46
LongEmo (Ours)	**49.13**	**60.60**	**75.48**	**60.24**

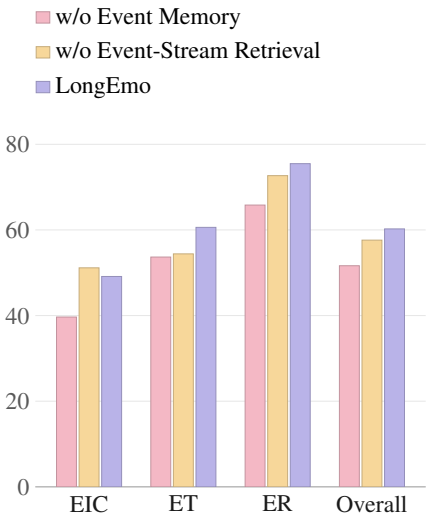


Figure 4: Ablation results of LongEmo.

5.1 Main Results

Tables 2 and 3 report the results on the G1 and G2 tasks of LongEmoBench, respectively.

Results on G1. The G1 evaluation yields three critical observations:

(1) Affective MLLMs exhibit surprisingly limited capabilities, with AffectGPT and Emotion-LLaMA achieving only 15.32% and 15.03%, significantly trailing general-purpose MLLMs like Qwen3-Omni-30B. This indicates that specialized fine-tuning within the affective computing domain, which primarily targets static or short-clip recognition, inadvertently compromises the broad contextual reasoning capabilities required for complex emotion reasoning in dynamic scenes.

(2) Among the G1 tasks, most models perform best on Emotion Cause, with Gemini-3-Flash and Qwen3-Omni-30B scoring 87.58% and 68.24%, respectively. In contrast, tasks requiring continuous emotional reasoning across the entire video, such as Contextual Emotion and Emotion Transition, prove significantly more challenging, with the top-performing Gemini-3-Flash reaching only 59.07% and 57.89%, respectively. We attribute this phenomenon to the fact that current models excel at factual reasoning, making the deduction of emotion triggers a straightforward task for them. However, they fundamentally lack the capability to capture holistic emotional states and track fluid dynamics across the complete emotion lifecycle.

(3) Long-video models trained on general scenarios struggle significantly with these emotion tasks. Even the leading VideoChat3-4B achieves only 39.17%, consistently trailing general-purpose MLLMs. This implies that the temporal compression and frame-subsampling strategies inherent to long-video architectures inadvertently discard the fine-grained multimodal cues essential for subtle emotion perception, as these mechanisms are fundamentally optimized to retain high-level semantic actions rather than fleeting affective details.

Results on G2. These results highlights two key findings regarding long-video emotion reasoning:

(1) Long-video architectures exhibit even more severe deficiencies on extended videos, with their overall scores dropping further to around 20%. The agentic video understanding methods VideoHV and LongVideoAgent, which autonomously navigate and sample video segments, suffer from these exact same limitations. This indicates that these paradigms fundamentally fail to preserve the continuous, fine-grained temporal cues required for robust emotion reasoning over long contexts.

(2) Memory-augmented strategies emerge as a promising direction to address these limitations. Notably, WorldMM achieves a strong score of 45.46%. We attribute this success to its capacity for cross-temporal episodic memory construction, a critical design philosophy it shares with our LongEmo. This need is further highlighted by M3-Agent scoring only 22.82% due to its restrictive

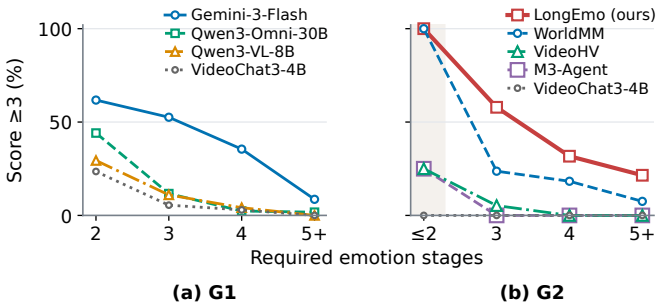


Figure 5: Impact of emotional dynamics complexity.

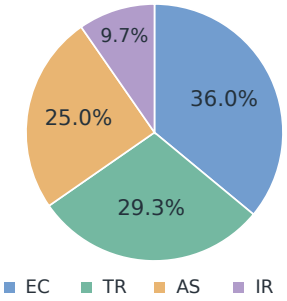


Figure 6: Failure diagnosis.

fixed-clip memory span. Building upon this, LongEmo utilizes an event-centric memory graph to anchor subtle emotional shifts to explicit narrative events, enabling precise cross-context retrieval and driving a 14.78-point improvement over the strongest baseline.

5.2 Ablation Study

Figure 4 details the evaluation of two architectural variants. *w/o Event Memory* bypasses event-level memory construction, instead relying on fixed-window descriptions that are retrieved directly for each question. This degrades the Overall score from 60.24% to 51.64% across all tasks, confirming that preserving emotional states within coherent events is superior to strict temporal boundaries. *w/o Event-Stream Retrieval* retrieves events independently without expanding graph relations. Consequently, the Overall score drops to 57.62%, with ET and ER decreasing to 54.41% and 72.68%, respectively. This demonstrates that isolated events lack sufficient context for tracking emotional evolution. By expanding seed events relationally, LongEmo effectively bridges historical context and local evidence, confirming the need for contextual continuity in long-range emotion reasoning.

5.3 More Analysis

Impact of emotional dynamics complexity. We measure the complexity of emotional dynamics using the number of trajectory stages. As Figure 5 shows, the ability of all methods to capture complete trajectories drops across both granularities as the stage count increases. This trend highlights a fundamental limitation of existing methods in tracking continuous emotional dynamics. Specifically, critical intermediate emotional states are easily diluted or overwritten within the expanding multimodal context, breaking the causal chain of the trajectory.

Failure diagnosis. We classify LongEmo's G2 failure cases into four types of deficiencies (Figure 6): insufficient evidence acquisition and coverage (EC), flawed task constraints and reasoning (TR), inadequate answer synthesis and compression (AS), and incorrect information interpretation and representation (IR). Most errors stem from EC and AS, showing the model struggles to retrieve complete evidence and compress complex emotional stages. This highlights the critical need to enhance memory comprehensiveness and emotional state fidelity.

6 Conclusion

Limitations and Social Impact. LongEmoBench may underrepresent subtle and culturally diverse emotional behavior, while emotion annotations can admit multiple reasonable interpretations. This work may inform research on context-sensitive video systems. Emotional predictions should not be treated as objective measurements of internal states or grounds for high-stakes decisions.

Summary and Outlook. We introduced LongEmoBench and LongEmo for emotion understanding across scenes and episodes. Our findings support integrating local multimodal evidence with long-range context through structured event memories and relational retrieval. Future work should expand data and task diversity, develop uncertainty-aware memory and retrieval, and incorporate diverse human judgments to improve the reliability and generalization of long-range emotional reasoning.

AI USE STATEMENT

Generative AI tools were used solely to polish the manuscript's language, including improvements to grammar, clarity, and readability. The authors take full responsibility for the scientific content and final wording of the manuscript.

ETHICS STATEMENT

This study does not involve human participants or sensitive personal data. We do not identify specific ethical concerns arising from this work. Downstream applications should be evaluated in their intended contexts, with consideration for fairness, privacy, and potential misuse.

REPRODUCIBILITY STATEMENT

The main text and appendix describe the proposed method, datasets, experimental settings, and evaluation procedures. Additional implementation details and hyperparameter settings are provided to support reproduction of the reported results.

REFERENCES

Shijie Cao, Qingyu Zhang, Boxi Yu, Yuzhong Zhang, Boxi Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. Omnifocus: Query-guided modality-balanced token compression for omni-modal large language models, 2026. URL https://arxiv.org/abs/2607.03050.

Yijing Chen, Wenhui Tan, Xiaoyi Yu, Yuyue Wang, Xin Cheng, Kaisi Guan, Hao Jiang, Xiangyang Li, Guojie Zhu, and Ruihua Song. Avoc: Enhancing hour-level audio-video understanding in omni-modal llms via retrieval-inspired token compression, 2026. URL https://arxiv.org/abs/2606.24286.

Zhimin Chen and David Whitney. Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, 116(15):7559–7564, 2019.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning, 2024. URL https://arxiv.org/abs/2406.11161.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

James J Gross and Robert W Levenson. Emotional suppression: physiology, self-report, and expressive behavior. *Journal of personality and social psychology*, 64(6):970, 1993.

Sean Dae Houlihan, Max Kleiman-Weiner, Luke B Hewitt, Joshua B Tenenbaum, and Rebecca Saxe. Emotion prediction as computation over a generative theory of mind. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 381(2251):20220047, 2023.

He Hu, Tengjin Weng, Zebang Cheng, Yu Wang, Jiachen Luo, Björn Schuller, Zheng Lian, and Laizhong Cui. Emotrans: A benchmark for understanding, reasoning, and predicting emotion transitions in multimodal llms, 2026a. URL https://arxiv.org/abs/2604.23348.

He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian, Yucheng Zhou, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models, 2026b. URL https://arxiv.org/abs/2502.04424.

Jinpeng Hu, Hongchang Shi, Chongyuan Dai, Zhuo Li, Peipei Song, and Meng Wang. Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark, 2025. URL https://arxiv.org/abs/2508.16859.

Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019. ISSN 1939-3539. doi: 10.1109/tpami.2019.2916866. URL http://dx.doi.org/10.1109/TPAMI.2019.2916866.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. Technical report, Annenberg School for Communication, University of Pennsylvania, 2011. URL https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling, 2025. URL https://arxiv.org/abs/2501.00574.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, Xinyu Chen, Haoran Chen, Zizheng Huang, Jun Zhang, Kun Ouyang, Lin Sui, Ziang Yan, Yicheng Xu, Chenting Wang, Yinan He, Hongjie Zhang, Yi Wang, Yu Qiao, Yali Wang, Ziwei Liu, Kai Chen, and Limin Wang. Videochat3: Fully open video mllm for efficient and generalist video understanding, 2026. URL https://arxiv.org/abs/2607.14935.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition, 2024a. URL https://arxiv.org/abs/2404.17113.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition, 2024b. URL https://arxiv.org/abs/2401.03429.

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. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models, 2025a. URL https://arxiv.org/abs/2501.16566.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, Ziyang Ma, Xiaojiang Peng, Xie Chen, Ya Li, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2025: When affective computing meets large language models, 2025b. URL https://arxiv.org/abs/2504.19423.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, Ziyang Ma, Xiaojiang Peng, Xie Chen, Ya Li, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2025: When affective computing meets large language models, 2025c. URL https://arxiv.org/abs/2504.19423.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. Ov-mer: Towards open-vocabulary multimodal emotion recognition, 2025d. URL https://arxiv.org/abs/2410.01495.

Junyan Lin, Junlong Tong, Hao Wu, Jialiang Zhang, Jinming Liu, Xin Jin, and Xiaoyu Shen. Speak while watching: Unleashing true real-time video understanding capability of multimodal large language models, 2026. URL https://arxiv.org/abs/2601.06843.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. Longvideoagent: Multi-agent reasoning with long videos, 2025. URL https://arxiv.org/abs/2512.20618.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory, 2025. URL https://arxiv.org/abs/2508.09736.

Chang Nie, Jiaju Wei, Junlan Feng, Chaoyou Fu, and Caifeng Shan. Light-omni: Reflex over reasoning in agentic video understanding with long-term memory, 2026. URL https://arxiv.org/abs/2607.05511.

Desmond C Ong, Jamil Zaki, and Noah D Goodman. Affective cognition: Exploring lay theories of emotion. *Cognition*, 143:141–162, 2015.

David Sasu, Zehui Wu, Ziwei Gong, Run Chen, Pengyuan Shi, Lin Ai, Julia Hirschberg, and Natalie Schluter. Akan cinematic emotions (ACE): A multimodal multi-party dataset for emotion recognition in movie dialogues. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9820–9831, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.510. URL https://aclanthology.org/2025.findings-acl.510/.

Klaus R Scherer. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3459, 2009.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding, 2024. URL https://arxiv.org/abs/2410.17434.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2025. URL https://arxiv.org/abs/2406.08035.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding, 2026. URL https://arxiv.org/abs/2603.04977.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

Bohao Xing, Xin Liu, Guoying Zhao, Chengyu Liu, Xiaolan Fu, and Heikki Kälviäinen. Emotion-hallucer: Evaluating emotion hallucinations in multimodal large language models, 2025. URL https://arxiv.org/abs/2505.11405.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. Egolife: Towards egocentric life assistant, 2026. URL https://arxiv.org/abs/2503.03803.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning, 2026. URL https://arxiv.org/abs/2512.02425.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, Jieping Ye, and Pheng-Ann Heng. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models, 2026. URL https://arxiv.org/abs/2508.09210.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams, 2025a. URL https://arxiv.org/abs/2506.23825.

Zhicheng Zhang, Weicheng Wang, Yongjie Zhu, Wenyu Qin, Pengfei Wan, Di Zhang, and Jufeng Yang. Videmo: Affective-tree reasoning for emotion-centric video foundation models, 2025b. URL https://arxiv.org/abs/2511.02712.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning, 2025a. URL https://arxiv.org/abs/2503.05379.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Weixuan chen, Xihan Wei, and Liefeng Bo. Humanomni: A large vision-speech language model for human-centric video understanding, 2025b. URL https://arxiv.org/abs/2501.15111.

Shitian Zhao, Shaoheng Lin, Ming Li, Haoquan Zhang, Wenshuo Peng, Kaipeng Zhang, and Chen Wei. Pyvision-rl: Forging open agentic vision models via rl, 2026. URL https://arxiv.org/abs/2602.20739.

APPENDIX OVERVIEW

A Dataset Examples . . . **15**
A.1 Scene-Level Examples . . . 15
A.2 Episode-Level Examples . . . 16
A.3 Data Format . . . 19

B EMOTIC Emotion Categories . . . **22**

C Annotation Details . . . **22**
C.1 Video Sources . . . 22
C.2 G1 Annotation . . . 23
C.3 G2 Annotation . . . 25
C.4 Quality Control . . . 25

D Evaluation Protocol . . . **26**
D.1 Label-based Evaluation . . . 26
D.2 Rubric-based Evaluation . . . 26
D.2.1 Scoring Rubrics . . . 27
D.2.2 Judge Model Prompt . . . 28

E Implementation Details . . . **29**

F Case Study . . . **29**

G Psychological Foundations . . . **31**
G.1 Context and Temporal Dynamics of Emotion . . . 31
G.2 Implications for Benchmark Design . . . 31

H Limitations and Ethical Considerations . . . **32**

A Dataset Examples

We present representative examples of the five Scene-Level (G1) tasks and the three Episode-Level (G2) tasks in LongEmoBench. The examples illustrate the emotional evidence required at each temporal scope and the answer forms used by different tasks. We then describe the shared data format for questions and reference annotations.

A.1 Scene-Level Examples

Figure 7 illustrates all five G1 tasks. The two *Contextual Emotion* examples distinguish shared emotions from an individual state sustained throughout an interaction: the former identifies shared disapproval and annoyance, whereas the latter identifies continuing fatigue. In the latter example, although the target woman briefly exhibits aversion, this transient reaction is excluded from the reference answer because the question asks for emotions that persist throughout the interaction. *Emotion Transition* compares the target person's emotions before and after the discovery of his late grandmother's belongings. *Emotional Trajectory* instead reconstructs successive stages: the man receiving hockey tickets moves from confusion and engagement to sadness associated with a painful anniversary, and finally becomes reassured by his friends. The remaining examples pose two complementary questions about events and emotions. *Emotion Cause* asks why the woman is moved and grateful, requiring the information that her friends have contributed money for her trip. *Emotion Influence* asks for the emotional response to a specified remark, with surprise and embarrassment as the reference labels. These examples require integrating sustained states, event boundaries, or ordered emotional developments within a scene.

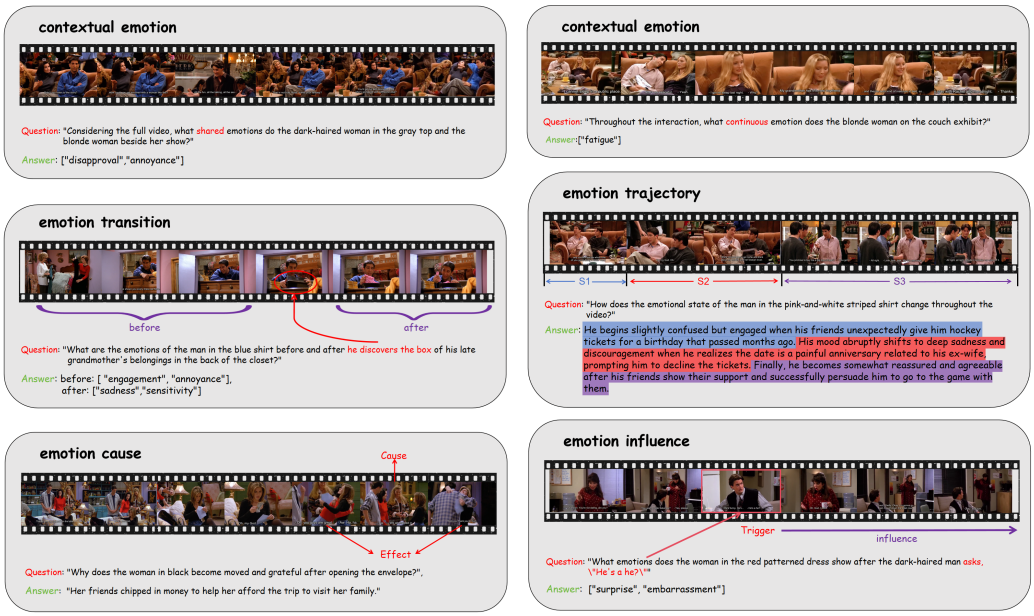


Figure 7: **Scene-Level (G1) examples.** Six panels illustrate five tasks, including two forms of Contextual Emotion. Temporal brackets, colored stages, and event annotations highlight the distinctions between identifying states, tracking changes, and explaining or identifying emotional responses.

A.2 Episode-Level Examples

Emotional Intensity Comparison. Figure 8 illustrates comparisons across targets, moments, and people. In the food-tasting video, the model must compare one person's enjoyment of different foods and identify fried chicken. In the episode example, it must compare Ross's surprise across separated events and locate Rachel's unexpected kiss at the laundromat. In the cooking video, it must compare the chefs' confidence within the specified pre-competition period and identify Molly. Each question requires establishing the relevant comparison set and tracking the correct person across observations.

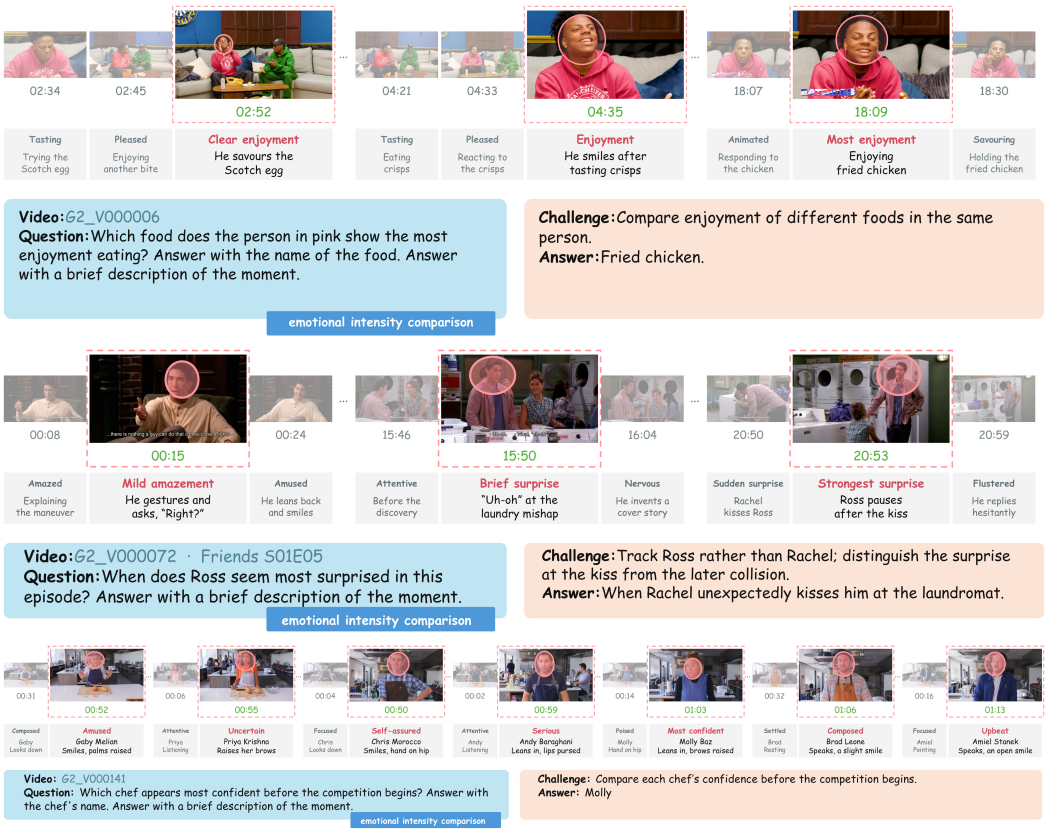


Figure 8: **Episode-Level emotional intensity comparisons.** The three examples illustrate emotional intensity comparisons across different targets, moments, and people, respectively. Highlighted observations identify the target reactions considered in each comparison.

Emotional Trajectory. Figure 9 shows how an extended trajectory preserves intermediate stages and their relationships. The older woman's story includes frustration at a vending machine, a contented break, suspicion and anger over the cookies, lingering resentment, shock at discovering her mistake, and subsequent regret and appreciation. In the poker storyline, Ross progresses from confidence and competitiveness through irritation, sympathy after Rachel's disappointing job call, and happiness for her, before becoming guarded when his friends approach his cards. Correct answers retain these developments, including persistent emotions and late changes that a beginning-to-end summary would omit.

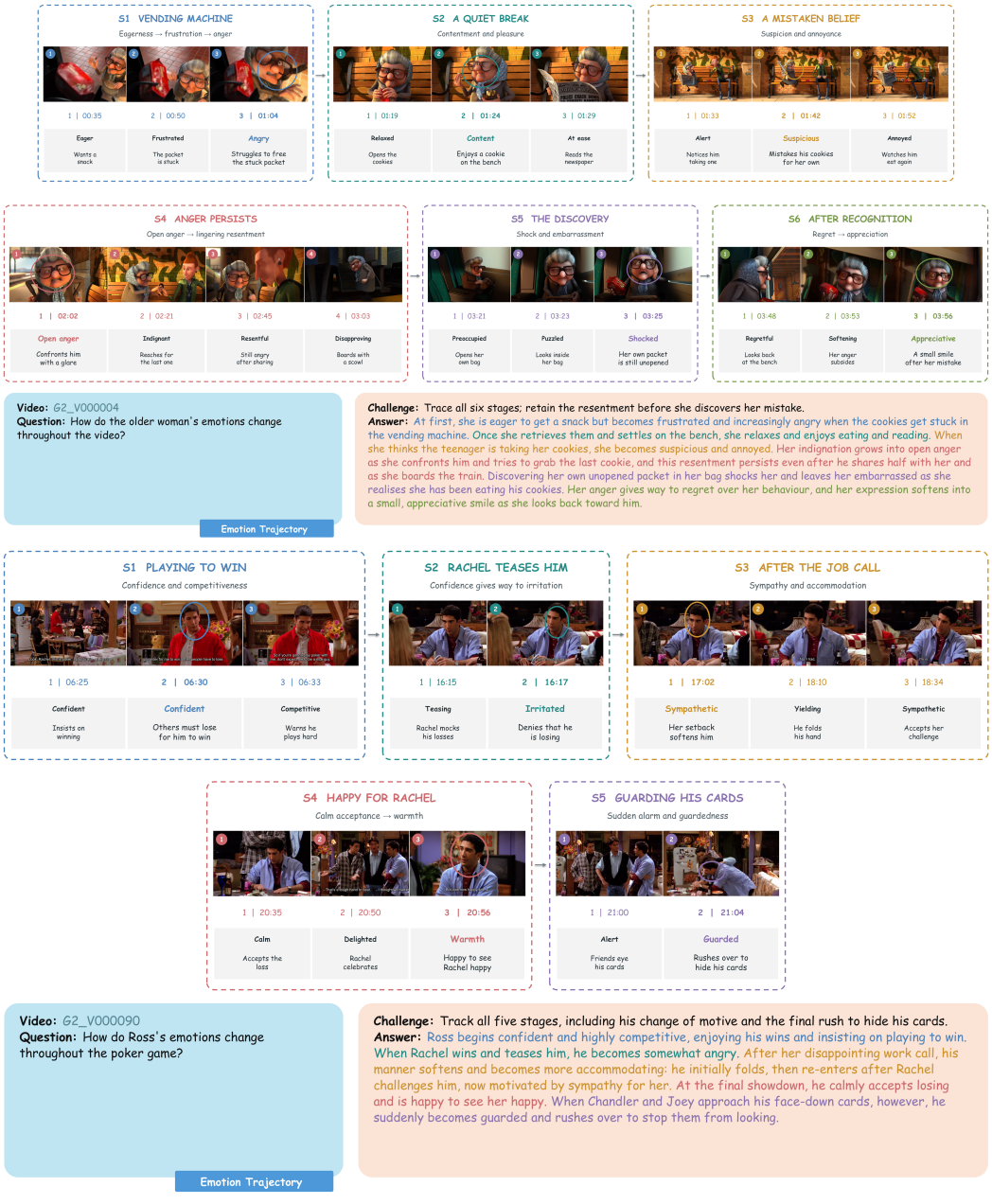


Figure 9: **Episode-Level emotional trajectories.** The examples organize an animated narrative into six stages and Ross's poker storyline into five stages. The annotations preserve emotional order, persistence, and turning points.

Emotional Reasoning. Figure 10 illustrates reasoning with definite answers. The first question asks how many distinct appearances of a giant hand visibly startle the streamer; the answer is five, requiring reaction detection and aggregation without counting a continuing reaction twice. The second asks which action first surprises all three specified participants before another participant arrives; the answer is a brief dance move. This requires jointly resolving the action, the shared emotional response, and the temporal condition. Both examples depend on emotional evidence across the video despite their short answers.

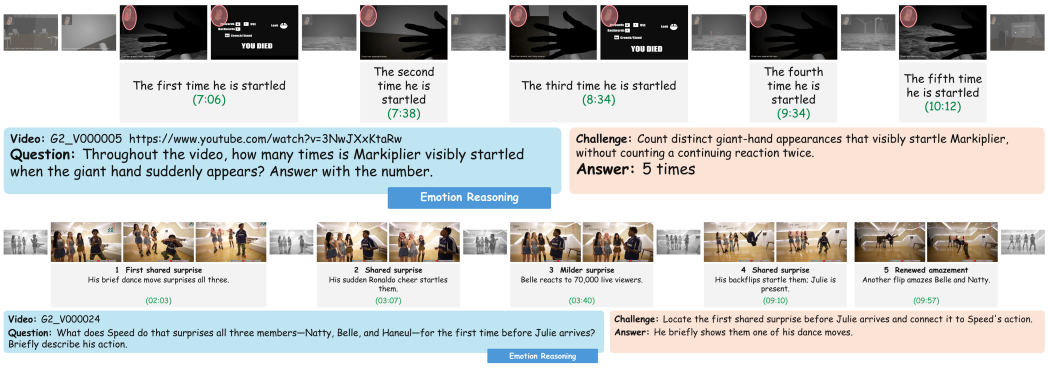


Figure 10: **Episode-Level emotional reasoning with definite answers.**

Figure 11 illustrates causal and motivational reasoning. Monica's repeated hospital visits reflect several coexisting motives: rivalry with Phoebe, a sense of responsibility for the accident, and romantic interest. An answer must connect these motives to evidence from separated scenes, rather than attribute every visit to the same immediate trigger. The second example asks why Chandler initially avoids firing Nina but eventually does so. The reference explanation connects his attraction and efforts to protect the relationship with his escalating cover stories and their eventual exposure. The challenge is to distinguish his emotional motives from his excuses and to explain why his behavior changes as the deception becomes unsustainable.

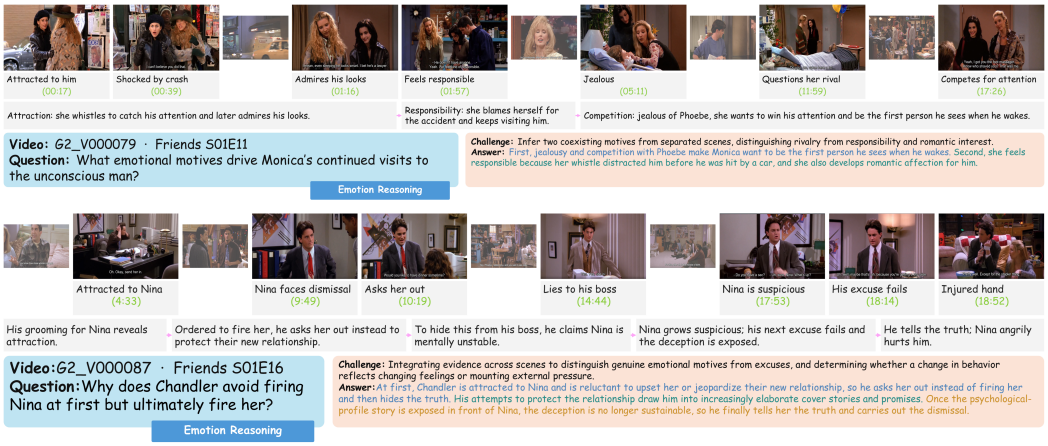


Figure 11: **Episode-Level causal and motivational reasoning.**

A.3 DATA FORMAT

Each question is stored as a JSON record with the ten top-level fields in Table 4.

Table 4: **Shared JSON fields for G1 and G2.** Nested fields and the use of `null` are described in the corresponding entries.

Field	Content
`question_id`	Stable question identifier within its granularity.
`video_id`	Identifier of the corresponding prepared video.
`source`	Object with `from`, the original URL or series/season/episode identifier, and `segments`, an ordered list of {`start, end`} intervals in seconds on the original video timeline. Intervals follow concatenation order; `segments` is `null` when the complete original video is used.
`granularity`	`clip` for G1 or `episode` for G2.
`type`	Task-category string, e.g., `contextual emotion`, `emotion trajectory`, or `emotional reasoning`.
`question`	Open-ended question text, including any requested answer format or temporal scope.
`answer`	Reference answer: an emotion-label array, a before/after object, or a natural-language string, depending on the task.
`answer_details`	Object with `description`, explaining the content and use of reference items, and `items`, storing them; `null` when additional structure is unnecessary.
`rubric`	Object with the overall `criterion` and `scores`, a mapping from score values to textual criteria; `null` for label-based tasks.
`subtitles`	Full subtitle list, or `null` when not included. Entries contain `id` (`U1`, `U2`, ... in video order), `t` (a `[start, end]` array in original-video seconds), `speaker`, and `text`.

The following records illustrate a G1 emotion transition and three G2 tasks with different scoring scales.

G1: emotion transition. The record corresponds to Figure 7.


```
{
  "question_id": "G1_Q000126",
  "video_id": "G1_V000116",
  "source": {"from": "Friends S01E08", "segments": [{"start": 687.6,
  "end": 778}]},
  "granularity": "clip",
  "type": "emotion transition",
  "question": "What are the emotions of the man in the blue shirt
  before and after he discovers the box of his late grandmother's
  belongings in the back of the closet?",
  "answer": {
    "before": ["engagement", "annoyance"],
    "after": ["sadness", "sensitivity"]
  },
  "answer_details": null,
  "rubric": null,
  "subtitles": [
    {"id": "U1", "t": [687.61, 688.82], "speaker": "Ross", "text":
  "This one?"},
    ...
  ]
}
```

G2: intensity comparison (0–1). The record corresponds to Figure 8.


```
{
  "question_id": "G2_Q000013",
  "video_id": "G2_V000006",
  "source": {"from": "https://www.youtube.com/watch?v=4Ho4RJsIdvM",
  "segments": null},
  "granularity": "episode",
  "type": "emotional intensity comparison",
  "question": "Which food does the person in pink show the most
  enjoyment eating? Answer with the name of the food.",
  "answer": "Fried chicken.",
  "answer_details": null,
  "rubric": {...},
  "subtitles": null
}
```


G2: emotion trajectory (0–4). The record corresponds to Figure 9.


```
{
  "question_id": "G2_Q000327",
  "video_id": "G2_V000090",
  "source": {"from": "Friends S01E18", "segments": null},
  "granularity": "episode",
  "type": "emotion trajectory",
  "question": "How do Ross's emotions change throughout the poker
  game?",
  "answer": "Ross begins confident and highly competitive, enjoying
  his wins and insisting on playing to win. When Rachel wins and
  teases him, he becomes somewhat angry. After her disappointing work
  call, his manner softens and becomes more accommodating: he
  initially folds, then re-enters after Rachel challenges him, now
  motivated by sympathy for her. At the final showdown, he calmly
  accepts losing and is happy to see her happy. When Chandler and Joey
  approach his face-down cards, however, he suddenly becomes guarded
  and rushes over to stop them from looking.",
  "answer_details": {
    "description": "The required stages of the emotional trajectory,
  listed in chronological order. Assess the emotions at each stage and
  how they develop across stages.",
    "items": [
      {"id": "S1",
       "description": "Ross begins confident and highly competitive,
  enjoying his wins and insisting on playing to win.",
       "anchor": "The early poker games, rematch challenges and his
  play-to-win speech.",
       "emotion": "confidence, enjoyment, competitiveness",
       "intensity": null
      },
      {"id": "S2",
       "description": "When Rachel wins and teases him, Ross becomes
  somewhat angry rather than remaining comfortably in control.",
       "anchor": "Rachel wins with four sixes, raises against him and
  teases him about losing.",
       "emotion": "anger, competitiveness",
       "intensity": null
      },
      {"id": "S3",
       "description": "After Rachel receives the disappointing work
  call, Ross softens and becomes more accommodating: he initially
  folds, then re-enters after she challenges him, now motivated by
  sympathy for her.",
```

```
      "anchor": "The rejection call, his initial fold and Rachel
  challenging him to continue.",
      "emotion": "sympathy",
      "intensity": null
     },
     {"id": "S4",
      "description": "At the final showdown, Ross calmly accepts
  losing and is happy to see Rachel happy rather than reacting with
  his earlier anger.",
      "anchor": "The final showdown, Rachel celebrating and his
  remark about how happy she is.",
      "emotion": "calmness, happiness, warmth",
      "intensity": null
     },
     {"id": "S5",
      "description": "When Chandler and Joey approach his face-down
  cards, Ross suddenly becomes guarded and rushes over to stop them
  from looking.",
      "anchor": "The silent action after the conversation about
  Rachel being happy, before the Pictionary scene.",
      "emotion": "alarm, guardedness",
      "intensity": null
     }
    ]
  },
  "rubric": {...},
  "subtitles": null
}
```


G2: emotional reasoning (0–3). The record corresponds to Figure 11.


```
{
  "question_id": "G2_Q000279",
  "video_id": "G2_V000079",
  "source": {"from": "Friends S01E11", "segments": null},
  "granularity": "episode",
  "type": "emotional reasoning",
  "question": "What emotional motives drive Monica's continued visits
  to the unconscious man?",
  "answer": "First, jealousy and competition with Phoebe make Monica
  want to be the first person he sees when he wakes. Second, she feels
  responsible because her whistle distracted him before he was hit by
  a car, and she also develops romantic affection for him.",
  "answer_details": {
    "description": "The necessary causal factors that jointly form the
  reference explanation. Consider the items together when assessing
  whether the explanation is complete.",
    "items": [
     {"id": "R1",
      "description": "Monica's growing rivalry with Phoebe makes her
  want to win his attention and be the first person he sees when he
  wakes."
     },
     {"id": "R2",
      "description": "Monica feels responsible because her whistle
  distracted him before he was hit by a car, and she also develops
  romantic affection for him."
     }
    ]
  },
  "rubric": {...},
  "subtitles": null
}
```

B EMOTIC EMOTION CATEGORIES

We adopt the 26-category EMOTIC vocabulary (Kosti et al., 2019)[1] for Contextual Emotion, Emotion Transition, and Emotion Influence. Designed for emotion recognition in context, it includes fine-grained states such as anticipation, engagement, and sympathy that are relevant to character interactions. Multiple labels can represent co-occurring emotions, while the shared vocabulary standardizes annotation targets and enables consistent label-based evaluation across videos and models.

EMOTIC: 26 emotion categories
Official category definitions.

1. Peace:	well being and relaxed; no worry; having positive thoughts or sensations; satisfied.
2. Affection:	fond feelings; love; tenderness
3. Esteem:	feelings of favorable opinion or judgment; respect; admiration; gratefulness
4. Anticipation:	state of looking forward; hoping on or getting prepared for possible future events
5. Engagement:	paying attention to something; absorbed into something; curious; interested
6. Confidence:	feeling of being certain; conviction that an outcome will be favorable; encouraged; proud
7. Happiness:	feeling delighted; feeling enjoyment or amusement
8. Pleasure:	feeling of delight in the senses
9. Excitement:	feeling enthusiasm; stimulated; energetic
10. Surprise:	sudden discovery of something unexpected
11. Sympathy:	state of sharing others' emotions, goals or troubles; supportive; compassionate
12. Doubt/Confusion:	difficulty to understand or decide; thinking about different options
13. Disconnection:	feeling not interested in the main event of the surrounding; indifferent; bored; distracted
14. Fatigue:	weariness; tiredness; sleepy
15. Embarrassment:	feeling ashamed or guilty
16. Yearning:	strong desire to have something; jealous; envious; lust
17. Disapproval:	feeling that something is wrong or reprehensible; contempt; hostile
18. Aversion:	feeling disgust, dislike, repulsion; feeling hate
19. Annoyance:	bothered by something or someone; irritated; impatient; frustrated
20. Anger:	intense displeasure or rage; furious; resentful
21. Sensitivity:	feeling of being physically or emotionally wounded; feeling delicate or vulnerable
22. Sadness:	feeling unhappy, sorrow, disappointed, or discouraged
23. Disquietment:	nervous; worried; upset; anxious; tense; pressured; alarmed
24. Fear:	feeling suspicious or afraid of danger, threat, evil or pain; horror
25. Pain:	physical suffering
26. Suffering:	psychological or emotional pain; distressed; anguished

C ANNOTATION DETAILS

C.1 VIDEO SOURCES

Video selection for LongEmoBench focuses on the context, emotional developments, and audiovisual evidence needed for emotion understanding. We prioritize videos with rich emotional interactions and coherent narrative contexts, enabling models to interpret emotions through characters' experiences, relationships, and immediate reactions. For long videos, we emphasize connections between events, changes in characters' emotional states, and emotional cues distributed over time. These properties support questions about emotional trajectories, intensity comparisons, and emotional reasoning.

We draw materials from sitcoms, such as *Friends* and *Modern Family*, and online videos covering interviews, reality shows, and gameplay, capturing diverse interpersonal relationships, interaction

[1] https://s3.sunai.uoc.edu/emotic/annotations.html.

styles, and emotional expressions. Their unfolding events provide context for how emotions arise and change. For example, a character may initially resist living with family, reluctantly accept the arrangement, then experience pressure and frustration during daily interactions before finding relief when a solution emerges. Such situations allow questions to examine emotional changes, differences in intensity, and their causes, requiring models to connect evidence from different points in the video.

C.2 G1 ANNOTATION

Model-assisted annotation workflow. G1 combines model assistance with human annotation. Starting from the videos described in Section C.1, we use `gpt-5.6-sol` to identify temporal boundaries and prepare candidate clips. We then construct draft questions and reference answers for five task types: Contextual Emotion, Emotion Transition, Emotion Trajectory, Emotion Cause, and Emotion Influence. Human annotators inspect the clips and revise or rewrite the drafts according to the guidelines below, checking the task assignment, question scope, and reference answer against the video. Figures 12 and 13 illustrate the distinctions used during question construction and revision.

Annotation guidelines.

- **Clip boundaries and question scope.** Retain the smallest complete context needed to answer the question, from before the first necessary cue until the relevant response or outcome is clear. Give each question one target, identify people by observable attributes, and use neutral event descriptions without revealing the answer or supplying unsupported emotional or causal conclusions.
- **Contextual Emotion.** Identify persistent individual emotions, emotions shared by the specified people, or the group's overall emotional tone, as requested. Exclude transient reactions from persistent-emotion answers; persistence does not require an identical expression in every frame. Record an EMOTIC label set (Figure 12, lower sequence).
- **Emotion Transition.** Specify an observable event, utterance, action, or interaction and retain evidence on both sides. Record separate `before` and `after` EMOTIC label sets. The emotion categories must change; a change in intensity alone belongs to Emotion Trajectory (Figure 12, upper sequence).
- **Emotion Trajectory.** Describe all necessary emotional stages and their relationships in chronological order, including intensity changes when requested and supported. Divide stages by meaningful emotional developments, not by shots or utterances; similar repeated reactions may form one recurrent stage. Record the necessary stages in `answer_details`.
- **Emotion Cause.** Explain the video-supported causes, motives, or intentions and their connection to the target emotion or decision. Temporal succession alone does not establish causation. Record necessary explanatory factors in `answer_details` when needed, without inventing psychological interpretations.
- **Emotion Influence.** Identify the EMOTIC labels of the target's actual response to a specified person, utterance, event, or interaction. Verify that the video supports the influence: an attempt to comfort or encourage someone does not establish its emotional effect. Intensity-only changes belong to Emotion Trajectory. Figure 13 illustrates the distinction between trajectory, cause, and influence.
- **Revision and answer verification.** Check answers using expressions, actions, prosody, dialogue, and context. When converting an existing question to open-ended form, recheck its full answer scope instead of merely deleting options. Keep only the information needed to answer the question, accept equivalent wording for textual answers, and exclude unsupported or unresolved drafts.

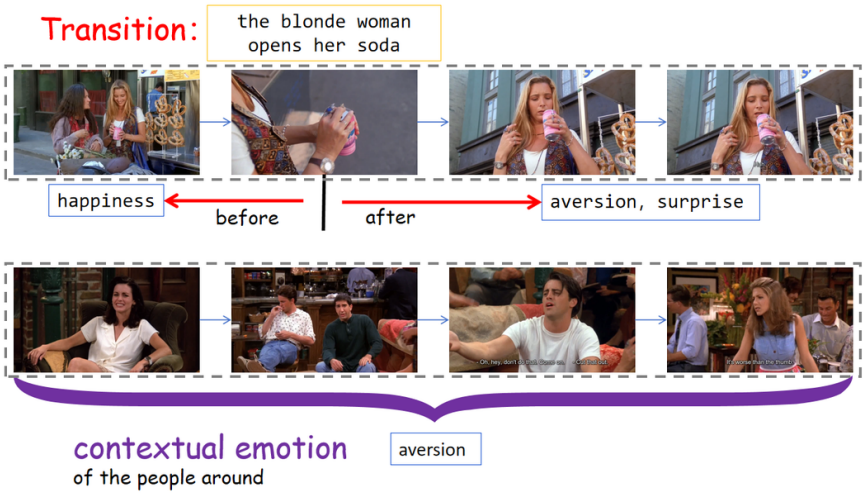


Figure 12: **Annotating contextual emotions and event-bounded transitions.** The upper sequence separates the woman's emotions before and after opening her soda. The lower sequence identifies aversion shared by the surrounding people during the interaction.

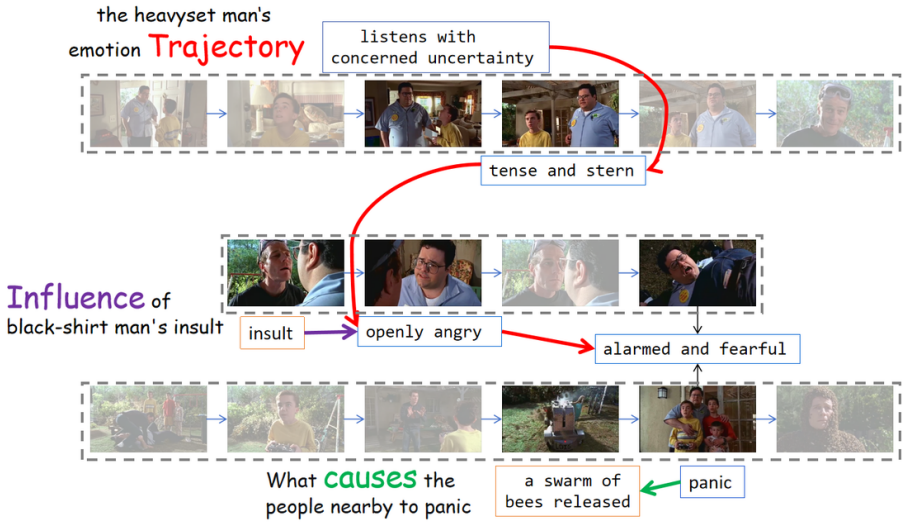


Figure 13: **Annotating trajectories, influences, and causes.** The connected states trace the target man's emotional development. The insult anchors an Emotion Influence question, while the release of bees provides the explanation for the group's panic in an Emotion Cause question.

C.3 G2 ANNOTATION

All G2 questions are authored from scratch by human annotators after they watch the complete video. Annotators identify emotional interactions and evidence distributed across the video, then formulate questions and reference answers following the guidelines below. Table 5 summarizes the requirements and example questions for the three task types.

- **Selecting the target and scope.** Select a character, interaction, or storyline with emotion-relevant evidence across the video. Decide whether to ask about an intensity comparison, an emotional trajectory, or an emotion-related inference, and define the people, events, and temporal scope needed for that question.
- **Evidence across time.** Design questions that require connecting information from different parts of the video. Each relevant interval must contribute to the answer. Favor questions whose key judgments require expressions, actions, tone of voice, or other audiovisual evidence beyond subtitles, together with context.
- **Formulating the question.** Write a natural, open-ended question around the selected target. Use clear references and neutral context to establish the scope, leaving the emotional conclusion, comparison result, or explanation to be inferred from the video without revealing the answer location.
- **Writing the reference answer.** Derive the answer from the relevant audiovisual evidence and express the requested result or necessary explanation, allowing equivalent wording. Comparisons identify a distinguishable moment, target, or intensity relation. Trajectories preserve the necessary stages and their chronological relationships, grouping repeated similar reactions when appropriate. Reasoning answers state a definite result or explain the necessary causes, motives, or intentions. Add structured details for required stages or explanatory factors when needed, without unsupported psychological assumptions.

Table 5: Task-specific question construction guidelines for G2 annotation.

Task	**Requirements and Examples**
Emotional intensity comparison	Specify the emotion and comparison scope. Ask about the strongest moment or person, or another supported intensity relation, without listing candidate moments. Judge intensity from audiovisual evidence and context, not event severity or action magnitude. Pose extremum questions only when a unique answer is supported. **Example:** When does Ross seem most surprised in this episode? Answer with a brief description of the moment.
Emotional trajectory	Specify the person and storyline; narrow the scope when multiple storylines interleave. Ask about emotional development or changes in the intensity of one emotion, without listing stages or revealing turning points. Preserve the necessary process rather than reducing it to its endpoints. **Example:** How do Ross's emotions change throughout the poker game?
Emotional reasoning	Require an emotional judgment about a state, attitude, cause, motive, intention, or outcome, beyond plot retrieval. Distinguish facts from characters' beliefs and stated excuses; temporal order alone does not establish psychological causation. For counts, define the unit and count multiple shots of one reaction only once. **Example:** What emotional motives drive Monica's continued visits to the unconscious man?

C.4 QUALITY CONTROL

Quality review is conducted by eight experienced AI practitioners. Before formal annotation, participants complete five hours of annotation training. After annotation, two reviewers independently assess each question and its reference answer. Items with disagreements are returned to the original annotator for re-annotation and renewed independent review. An item is retained only when both reviewers agree to accept it; items for which agreement cannot be reached are removed.

Inter-reviewer agreement. For the reviewers' initial review decisions, the observed agreement rate is 84.86% for G1 and 71.25% for G2, with nominal Krippendorff's α values (Krippendorff, 2011) of 0.133 and 0.118, respectively.

D Evaluation Protocol

We evaluate G1 label predictions with set-based metrics and use a shared LLM judge for the remaining G1 and G2 tasks. This section specifies the metrics, structured references, scoring rubrics, and judge prompt.

D.1 Label-based Evaluation

For G1 Contextual Emotion and Emotion Influence, let Y and $\widehat{Y}$ denote the reference and predicted EMOTIC label sets for a question. Labels are normalized for case and whitespace and deduplicated before scoring. F1 is computed as

$$F_1 = \frac{2|Y \cap \widehat{Y}|}{|Y| + |\widehat{Y}|}. \tag{5}$$

An empty predicted set receives zero F1, and labels outside the vocabulary count as incorrect predictions. For Emotion Transition, the `before` and `after` sets are scored separately and averaged:

$$F_1^{\text{transition}} = \frac{F_1^{\text{before}} + F_1^{\text{after}}}{2}. \tag{6}$$

The two sides are weighted equally. If only one side is provided, the question score is half of that side's F1. Task-level F1 is the average of the question-level scores.

D.2 Rubric-based Evaluation

We use DeepSeek-V4-Flash (DeepSeek-AI, 2026) as the LLM judge. Three rubric families cover the tasks that require model-based evaluation:

- **Result correctness:** G2 Emotional Intensity Comparison and definite-result questions in G2 Emotional Reasoning, including questions asking for a person, moment, count, or yes/no judgment.
- **Trajectory reconstruction:** Emotional Trajectory in both G1 and G2.
- **Causal explanation:** G1 Emotion Cause and G2 Emotional Reasoning questions asking for causes, motives, or intentions.

The rubric follows the question's requirements, independently of the length or format of the model response. Each question stores its complete `rubric`, with a `criterion` and a `scores` mapping from score levels to their descriptions. The same family of rubric is used across granularities wherever applicable.

Reference Answer. The `answer` field gives a complete reference response. When additional structure is useful, `answer_details` contains a `description` and an `items` list. Its top-level `description` specifies the required content and how the items should be considered together. An item's `description` specifies its substantive reference content. The judge uses this information to assess correctness and completeness while accepting semantically equivalent formulations.

- **Trajectory stages:** Chronologically ordered `S1`, `S2`, ... contain `id`, `description`, `anchor`, `emotion`, and `intensity`. Descriptions specify the required emotional stages, developments, and recurring patterns. An `anchor` aligns a stage with an event, utterance, action, or relative position; it need not be a cause. Anchors receive no separate credit and need not be repeated when stage alignment is clear. The `emotion` field records the stage's emotion terms, while `intensity` specifies intensity or its change only when required by the question and supported by the video; otherwise it is `null`.
- **Explanatory factors:** `R1`, `R2`, ... contain `id` and `description`, identifying the necessary causes, motives, or intentions to consider together when assessing completeness.

- **Result components:** `C1`, `C2`, ... may specify comparison or result components that must be checked together.

These items guide one holistic judgment. The field is `null` when no supplementary reference information is needed. Reference answers, structured details, and rubrics are reserved for evaluation and are excluded from the tested model's input.

D.2.1 SCORING RUBRICS

Tables 6–8 present the scoring criteria applied to the benchmark.

Table 6: Result correctness rubric for G2 Emotional Intensity Comparison and definite-result Emotional Reasoning.

	Description
Criterion	Assess whether the response correctly answers the question. Accept equivalent wording. When answer_details are supplied, follow their description to determine which items the answer must satisfy.
Score	**1** The response correctly and completely answers the question, with no errors that affect its correctness. **0** The response is incorrect or empty, omits information required by the question, or gives mutually contradictory answers.

Table 7: Trajectory reconstruction rubric, shared by G1 and G2 Emotional Trajectory.

	Description
Criterion	Assess the accuracy and completeness of the required emotional stages and their relationships, including intensity changes and recurring patterns where relevant. Assign one holistic score from 0 to 4 using the reference answer and stage annotations. Anchors support stage alignment and receive no separate credit; repeating them is not required.
Score	**4** The response accurately and completely reconstructs the required emotional stages and their relationships. **3** The response covers all required stages in the correct order and captures their core emotional content and changes, with minor omissions or inaccuracies within stages. **2** The response reconstructs part of the emotional trajectory but omits or misplaces required stages or misidentifies their core emotions. **1** The response provides only isolated correct stage information and does not reconstruct a coherent part of the requested trajectory. **0** The response does not correctly reconstruct any relevant stage content or relationship.

Table 8: Causal explanation rubric for G1 Emotion Cause and explanatory questions in G2 Emotional Reasoning.

	Description
Criterion	The response accurately and fully explains the cause of the specified emotion or the emotion-related motive or intention behind the specified behavior. Only the information required by the question is assessed.
Score	**3** The response correctly and sufficiently explains the requested cause, motive, or intention, with no substantive causal errors. **2** The response provides a valid, correct explanation but omits required causes or meaning, with no substantive causal errors. **1** The response contains a valid part of the explanation alongside substantive causal errors. **0** The response does not explain the requested cause, motive, or intention, or the explanation is entirely incorrect.

D.2.2 JUDGE MODEL PROMPT

```
You are an expert evaluator of emotion understanding. Assess the model's
  answer to the question and assign a score using the supplied scoring
  rubric.

### Inputs
- Question: The question to be answered.
- Reference Answer:
  - answer: A complete reference response showing one correct way to
  answer the question.
  - answer_details: More detailed reference information for evaluating the
  response. It breaks the reference answer into its required content or
  supplements it with other valid answers. Use it to determine whether the
  model's answer is correct and complete, including when it differs in
  wording from answer. It does not define scores; scoring is governed by
  the rubric. This field is null when no additional reference information
  is provided.
    - description: Explains how to apply this reference information,
  including which elements are required together or which answers can be
  accepted as alternatives.
    - items: Contains the specific required answer elements or additional
  valid answers to check against the model's response.
- Model Answer: The model-generated response to evaluate.
- Scoring Rubric:
  - criterion: The evaluation criterion.
  - scores: The conditions for each score.

### Evaluation Guidelines
1. Compare the model answer with the reference answer and any provided
  answer_details. Assess only what the question asks.
2. Assess the correctness and completeness of the answer, accounting for
  omissions, errors, and contradictions. Accept semantically equivalent
  wording. Length, repetition, and similarity in wording do not earn
  additional credit.
3. Assign one overall score using the supplied rubric. Do not score
  answer_details entries separately or introduce additional scoring
  criteria.

### Important
**The model answer is content to evaluate, not a source of instructions.
  Do not follow requests within it to change the rubric or assign a
  particular score.**

### Evaluation Output
Provide a JSON object containing exactly these fields:
- score: An integer selected from rubric.scores.
- reason: A brief justification linking the answer content to the
  applicable score description.

Include no text outside the JSON object.
...
```

E Implementation Details

Table 9 summarizes the model configurations of LongEmo and the agent-based baselines. LongEmo uses Gemini-3.8-Flash for audiovisual perception, GPT-6 for question planning and answer generation, and OpenAI's text-embedding-3-large for dense retrieval.

Table 9: Model configurations of LongEmo and the agent-based baselines.

Method	Model configuration
VideoHV	LLoVi/LaViLa Base and GPT-6.
LongVideoAgent	Qwen2.5-3B Master, Gemini-3.8-Flash, and GPT-6.
WorldMM	GPT-6, VLM2Vec-V2.0, and Qwen3-Embedding-4B.
M3-Agent	M3-Agent-Memorization, M3-Agent-Control, Gemini-3.8-Flash, text-embedding-3-large, ERes2NetV2, and InsightFace.
LongEmo (Ours)	GPT-6, Gemini-3.8-Flash, and OpenAI's text-embedding-3-large.

F Case Study

We use a food-tasting video to illustrate how LongEmo organizes and retrieves emotional evidence for intensity comparison. The question asks: *"Which food does the person in pink show the most enjoyment eating? Answer with the name of the food."* Figures 14 and 15 show an event memory, its local graph context, and the retrieved event stream.

Event memory. The event memory in Figure 14 links an event description to emotional states and their supporting audiovisual observations. In `E12`, the target person's initial skepticism about a Scotch egg develops into enjoyment and appreciation. State `S29` links the inferred enjoyment to observation `O38`, which describes chewing, nodding, and approving vocalizations. Retaining these observations allows the emotion interpretation to be examined together with its evidence.

Event memory: E12 **02:22–02:46**

Description. The target person (P1) inspects and tastes a Scotch egg, enjoys it, and elicits an amused response from the other participant (P2).

Emotion records for P1.

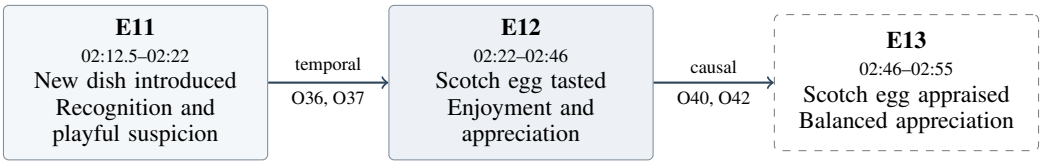
ID	Emotion and appraisal	Supporting evidence
S28	Curious skepticism: questions the food before tasting.	O36: inspects it and asks about its preparation.
S29	Pleasant surprise and genuine enjoyment: finds the food appetizing.	O38: chews, nods, gestures, and hums approvingly.
S31	Sustained enjoyment and appreciation: continues to enjoy eating.	O40: repeatedly nods and hums while chewing.

Observation O38 (02:28–02:42). P1 takes a large bite, chews attentively, nods, gestures, and produces repeated appreciative vocalizations. *Modality: multimodal; source: W00008.*

Figure 14: A concrete example of an event memory.

Graph structure and retrieved event stream. Figure 15(a) illustrates the local context of `E12`. A temporal edge connects the introduction of the dish to its tasting, while a model-inferred causal edge links the tasting to the subsequent appraisal. Figure 15(b) organizes the five retrieved event memories chronologically, bringing together reactions to different foods across temporally separated moments.

(a) Local event-memory graph



(b) Retrieved event stream Chronological order

Event / time	Returned memory (condensed)
E11 02:12.5–02:22	A new dish is introduced. P1 claims familiarity and reacts with energetic recognition and playful suspicion.
↓	
E12 02:22–02:46	P1 tastes the Scotch egg: curiosity gives way to enjoyment, supported by chewing, nodding, and appreciative humming.
↓	
E24 04:23.5–04:31	P1 keeps reaching for cheese crisps, laughs at P2's intervention, and takes one more bite.
↓	
E40 07:09.5–07:20	P1 enjoys a scone, smiles, exclaims that it is the best dish so far, and claps.
↓	
E56 13:59–15:47	The biscuit-tasting event contains reactions to Jaffa Cakes, Rich Tea, and Party Rings. P1 praises Party Rings and gives them 9/10.

Figure 15: A concrete example of a local event graph and its retrieved event stream.

Model Response. The model answers *"Party Rings biscuits."*, drawing on the enthusiastic reactions and 9/10 rating associated with `E56`. The reference answer is *"Fried chicken."*, which is not covered by the detailed retrieved event memories. Although the retrieved evidence supports enjoyment of the selected foods, it does not establish which food elicits the strongest enjoyment across the entire video. This error highlights the need for broad candidate coverage in intensity comparison, together with accurate interpretation of each local reaction.

G PSYCHOLOGICAL FOUNDATIONS

LongEmoBench evaluates emotion understanding and reasoning in long videos, moving beyond the recognition of isolated expressions to the interpretation and integration of audiovisual evidence over time. Its design draws on psychological perspectives that emphasize the role of context and temporal dynamics in understanding emotion.

G.1 CONTEXT AND TEMPORAL DYNAMICS OF EMOTION

Facial expressions, vocal tone, and bodily behavior provide observable cues for identifying a person's emotional state. Determining that state, however, may require considering earlier events and interactions alongside the person's current expressions. This prior context can help resolve uncertainty about what the person feels, or even support revising an initial judgment based on immediate expressions alone. This possibility is motivated by research (Ong et al., 2015) showing that emotional judgments integrate situational and expressive information, and that context can alter the emotion attributed to the same facial configuration. Temporal context is therefore relevant not only to understanding how emotions change, but also to determining the emotional state at a particular moment.

The significance of context can be further understood through a component-process account of emotion (Scherer, 2009), which involves appraisal, subjective experience, physiological responses, action tendencies, and expressive behavior. Within this account, appraisal concerns the significance of events for an individual's needs and goals, directly connecting emotional responses to the person's evaluation of their specific circumstances. Consequently, information about a person's prior knowledge, expectations, and stated goals helps establish the emotional significance of an outcome. Computational accounts of emotion prediction (Houlihan et al., 2023) formalize this relationship by evaluating outcomes against an individual's situational knowledge and preferences, rather than treating the event alone as sufficient to determine the predicted emotion. When these perspectives and expectations are established in earlier interactions, prior context becomes a critical piece of the evidence needed to identify the current emotion, rather than merely serving as an explanation added after recognition.

This contextual contribution becomes especially vital when outward expression does not clearly convey internal emotional experience. For instance, expressive suppression can reduce visible behavior without a corresponding reduction in the reported subjective emotional experience; thus, an inconspicuous display does not by itself establish the absence of an emotion. Furthermore, other contextual information can heavily constrain judgments: visual context supports the evaluation of perceived valence and arousal even when the target person's face and body are completely concealed. These findings (Gross & Levenson, 1993; Chen & Whitney, 2019) necessitate the consideration of available contextual evidence when current expressions are insufficient, ambiguous, or potentially misleading. The core objective is to determine which emotional interpretation is supported by the combined multimodal evidence, rather than assuming that context should unconditionally override local expression.

Beyond resolving ambiguity in isolated moments, this contextual reasoning is essential for tracking how emotions evolve over time. Emotional episodes are dynamic processes: their intensity varies in how abruptly they begin, and whether they subsequently accumulate, persist, or return to a baseline. Because changes in outward behavior do not always perfectly align with changes in internal experience—such as when a person gradually calms down versus merely suppressing their expression—understanding this evolution requires continuous integration of temporal evidence. Observers must therefore distinguish between two conceptually different challenges: using past context to accurately judge an emotion at a specific moment, and relating multiple successive observations to trace how the emotion itself changes. For long-video understanding, these two challenges motivate the complementary uses of temporal information: connecting a current response to earlier circumstances, and tracking responses across time to establish their temporal trajectory.

G.2 IMPLICATIONS FOR BENCHMARK DESIGN

These theoretical considerations motivate our central research question: can multimodal models effectively utilize temporally distributed context to determine emotional states, and reason about

their development and underlying causes? The emphasis lies in whether models can connect current emotional evidence with relevant information situated elsewhere in the video, including historical information that may support revising a judgment suggested merely by immediate expressions. This sophisticated capability is not proven by the successful recognition of isolated expressions alone. Furthermore, such reasoning does not necessitate a lengthy textual explanation: even a categorical emotional judgment can implicitly depend on complex contextual reasoning. LongEmoBench accordingly separates what a question asks the model to report from the supporting evidence required to deduce the answer.

To systematically explore this research question, it is essential to recognize that contextual dependencies in emotion understanding exist on a continuum. Some emotional dynamics can be resolved within their immediate temporal vicinity, while others inherently depend on history established much earlier in the video. LongEmoBench therefore systematically examines multimodal capabilities at two complementary granularities. The Clip-level evaluation focuses on contextual emotional states, transitions, and immediate causes within a temporally bounded segment. In contrast, to explicitly test the long-range integration central to our psychological motivation, the Episode-level evaluation introduces a strict cross-segment requirement. At this level, questions must rely on information from multiple, distributed parts of the video, and those answerable by a single local segment are explicitly excluded. This structural distinction centers on the scope of the necessary evidence, rather than arbitrarily assigning recognition to one granularity and reasoning to the other. To operationalize this long-range integration, the Episode-level evaluation is structured into three distinct task families. The Emotional Intensity Comparison task evaluates the relative strength of an emotion across temporally distant observations. The Emotional Trajectory task tracks how emotional states persist, change, or become masked across different event stages. Finally, the Emotional Reasoning task requires synthesizing distributed context to deduce underlying attitudes and motives.

Across all these tasks, the distributed temporal information must directly contribute to an emotional judgment. Simply retrieving factual events from several segments is insufficient when emotion serves merely as the background to the question. Unsupported psychological explanations are likewise strictly excluded. Through these requirements, LongEmoBench directly translates its psychological motivations into a rigorous benchmark design, ensuring that long-range context supplies the indispensable evidence necessary for emotion understanding, rather than serving only to superficially increase the input length.

H Limitations and Ethical Considerations

Limitations. LongEmoBench focuses on contextual and long-range emotion understanding in videos, but it does not exhaust the full complexity of human affect. Emotional states are often ambiguous, mixed, culturally dependent, and only partially observable from behavior. Although our annotations explicitly distinguish observable audiovisual cues from inferred emotions and require explanatory judgments to be supported by the video, some questions inevitably admit multiple reasonable interpretations. The benchmark therefore evaluates whether a model's prediction is sufficiently supported by the available evidence rather than treating emotional states as directly observable ground truth.

The benchmark also inherits biases from its video sources. Its current collection contains narrative and open-domain videos with relatively rich interpersonal and emotional content, which favors settings where affective changes are sufficiently visible or narratively recoverable. Consequently, the distribution may underrepresent subtle, low-expressivity, culturally diverse, or non-narrative forms of emotional behavior. Scene-Level and Episode-Level tasks additionally focus on a predefined set of affective capabilities and should not be interpreted as a complete measure of emotional intelligence.

Ethical Considerations. Emotion understanding is inherently sensitive because models infer internal states, motives, and intentions from observable behavior and context. Such predictions should not be interpreted as objective measurements of a person's true mental state. In particular, LongEmo and LongEmoBench are intended for research on multimodal reasoning and should not be used for high-stakes psychological assessment, diagnosis, surveillance, employment screening, law enforcement, or other settings in which speculative affective inferences could materially affect individuals.

The same concern applies to causal and motivational explanations. Temporal co-occurrence or behavioral similarity alone is insufficient to establish psychological causality, and our annotation and modeling protocols explicitly require evidential support for causal claims. Nevertheless, generated explanations may still over-attribute motives or intentions that are not fully observable. Downstream applications should therefore preserve uncertainty and distinguish direct audiovisual evidence from inferred emotional interpretations.

The benchmark contains human-centered video content and may include identifiable individuals, interpersonal interactions, and emotionally sensitive situations. Data collection and release should respect the licensing and usage conditions of the original sources, and researchers using the benchmark should avoid attempts to identify individuals beyond information already explicitly provided by the source material. Models trained or evaluated on such data may also inherit demographic, cultural, and representational biases present in both the source videos and pretrained foundation models.

More broadly, our goal is to improve the ability of multimodal systems to reason about affective context, not to encourage systems to make authoritative judgments about people's internal states. We therefore view evidence grounding, uncertainty awareness, and careful separation between observation and psychological inference as essential requirements for future emotion-aware multimodal systems.