EMOFORECASTER: ADVANCING MULTIMODAL EMOTION FORECASTING VIA STATE-TRANSITION MODELING AND EVIDENCE-GROUNDED REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion is fundamental to human experience and social interaction, making the anticipation of future emotional responses important for proactive interactive systems. However, existing multimodal emotion understanding methods largely focus on emotions that have already been expressed. Although several preliminary studies have explored emotion forecasting, they primarily predict emotion categories, providing limited explicit reasoning about how multimodal evidence and interaction dynamics support their predictions. To address these challenges, we propose EmoForecaster, an evidence-grounded multimodal emotion forecasting framework that integrates state-transition modeling with progressive multimodal reasoning. We first construct EmoEvolve with structured interaction reasoning annotations comprising five components that jointly capture conversational context, multimodal affective evidence, and interaction-aware emotional transitions, and establish a clip-disjoint evaluation protocol covering both emotion forecasting and rationale generation. Furthermore, we develop a complementary architecture in which the state-transition path uses historical reference and preceding interlocutor states to construct context-conditioned forecasts, while the evidence-reasoning path progressively interprets multimodal cues to explain how forthcoming emotional responses may develop, with the two paths coupled at the decision level and optimized through supervised fine-tuning followed by path-specific Group Relative Policy Optimization (GRPO). Extensive experiments demonstrate the superiority of EmoForecaster in multimodal emotion forecasting, with strong performance in both predictive accuracy and rationale quality.

1 INTRODUCTION

Emotion is fundamental to human experience and social interaction, playing a key role in communicating intentions, coordinating behavior, and maintaining interpersonal relationships Keltner & Haidt (1999). Rather than being conveyed through a single channel, human emotions are expressed through interrelated signals, including linguistic content, vocal prosody, and facial expressions Scherer (2009); Poria et al. (2017). Accordingly, integrating visual, acoustic, and textual information to perceive and understand emotional states has become an important research direction in affective computing and a foundation for emotionally aware artificial intelligence systems D'mello & Kory (2015). Such advances can help systems better understand users' emotional needs, facilitating more natural and empathetic interactions with potential applications in human–computer interaction Cheng et al. (2024a), educational assistance Yadegaridehkordi et al. (2019), and mental health support Schlicher et al. (2025).

Specifically, earlier research on multimodal affect recognition largely followed a discriminative paradigm, fusing representations from different modalities (Liu et al., 2018) to map inputs to sentiment polarities (Zadeh et al., 2017) or predefined emotion categories (Hu et al., 2021b). Despite their progress, these approaches typically output labels without explicitly explaining which observable cues and cross-modal relationships support their predictions (Lian et al., 2023). In recent years, multimodal large language models (MLLMs) have developed rapidly, demonstrating strong performance in video understanding (Lin et al., 2024), audiovisual question answering (Cheng et al., 2024b), and multimodal reasoning (Liu et al., 2023), thereby opening up new possibilities for emo-

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.

tion understanding beyond label prediction. Emerging emotion-capable MLLMs are often developed using constructed emotion-specific datasets and instruction tuning (Cheng et al., 2024a; Lian et al., 2025), enabling them to generate emotion descriptions and explanations grounded in audiovisual and textual information and advancing multimodal affective computing toward more fine-grained and interpretable emotion understanding. However, these advances primarily concern emotions that have already been expressed, whereas proactive interactive systems also need to anticipate how users' emotions may evolve to inform their responses. Developing models that use the available context to forecast emotional states that have yet to emerge (Wang et al., 2026) is therefore a promising research direction, with the potential to support more timely responses and proactive assistance.

Indeed, prior studies have already made initial attempts in this direction. Early representative approaches primarily focused on text-based conversations, modeling emotion propagation across multiple turns to infer a target participant's forthcoming emotion without observing their next response (Li et al., 2021). However, text alone does not directly capture nonverbal cues such as vocal prosody and facial expressions, potentially omitting important evidence for understanding interactions and anticipating emotional changes (Poria et al., 2017). Subsequent research has extended this line of work to multimodal settings. As a recent notable effort, Hi-EF introduced forecasting samples comprising four consecutive interaction clips, using the video, audio, and text from the first three clips to predict the interaction partner's emotion in the fourth (Wang et al., 2026). Nevertheless, these forecasting approaches primarily focus on predicting future emotion categories, rather than offering explicit reasoning that connects historical evidence, interaction dynamics, and future emotions. Therefore, multimodal emotion forecasting should move beyond isolated label prediction to explain how future emotions may develop from the preceding interaction, making the supporting multimodal evidence and inferred emotional transitions open to inspection.

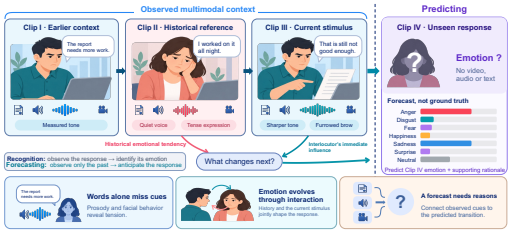


Figure 1: Overview of multimodal emotion forecasting, where the goal is to anticipate an unseen future emotional response from preceding multimodal interaction context and provide a supporting rationale.

Based on the above observations, we propose EmoForecaster, a new multimodal emotion forecasting framework that explicitly models interaction dynamics and progressively reasons over multimodal interaction history to anticipate future emotional responses. Specifically, we first construct EmoEvolve, a dataset enriched with structured interaction reasoning annotations consisting of five components, namely *Observed Dialogue*, *Observed Multimodal Evidence*, *Historical Evidence*, *Cross Evidence*, and *Context-response Comparison*, which collectively capture conversational context, multimodal affective cues, historical emotional evidence, current-source information, and the inferred transition toward the forthcoming emotional response, and establish a clip-disjoint evaluation protocol covering both emotion prediction and rationale generation. Furthermore, we develop a complementary architecture with state-transition and evidence-reasoning paths, where the former models the historical reference state and the preceding interlocutor state to construct a context-conditioned future emotion prediction, while the latter progressively interprets multimodal cues to infer how the forthcoming response may evolve. The two paths are coupled at the decision level and optimized with a two-stage strategy comprising supervised fine-tuning (SFT) followed by path-specific Group Relative Policy Optimization (GRPO). Extensive experiments demonstrate that the proposed EmoForecaster consistently outperforms both general-purpose MLLMs and emotion-capable MLLMs on multimodal emotion forecasting and the corresponding rationale generation. Our contributions can be summarized as follows:

- We propose EmoForecaster, an evidence-grounded multimodal emotion forecasting framework that anticipates future emotional responses through progressive reasoning.
- We construct EmoEvolve with structured interaction reasoning annotations and establish a clip-disjoint evaluation protocol for evaluating emotion forecasting and rationale generation.
- We develop a complementary architecture with state-transition and evidence-reasoning paths coupled at the decision level, and design a two-stage optimization strategy comprising SFT followed by path-specific GRPO.
- Extensive experiments demonstrate that EmoForecaster outperforms representative general-purpose and emotion-capable MLLMs in both predictive accuracy and rationale quality.

2 METHODOLOGY

To support evidence-grounded multimodal emotion forecasting, we first construct EmoEvolve, a dataset enriched with structured interaction reasoning annotations that capture both emotional states and multimodal evidence throughout the observed interaction. On this basis, we develop EmoForecaster, a complementary architecture with a state-transition path and an evidence-reasoning path to jointly model emotional dynamics within interactions and their supporting multimodal evidence, coupled at the decision level and optimized through supervised fine-tuning (SFT) followed by path-specific Group Relative Policy Optimization (GRPO).

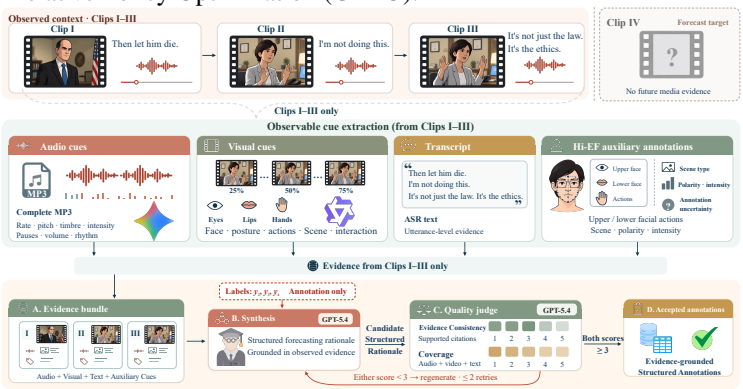


Figure 2: Construction pipeline of EmoEvolve, including multimodal cue extraction, structured rationale synthesis, quality assessment, and annotation acceptance.

2.1 EMOEVOLVE

EmoEvolve is grounded in Hi-EF (Wang et al., 2026), a recent multimodal dataset for emotion forecasting in human interaction. Specifically, each instance follows a four-clip Multilayered-Contextual Interaction Sample (MCIS) format, with Clips I and II providing interaction history, Clip III presenting Party A's current state, and Clip IV containing Party B's forthcoming emotional response; all clips provide video, audio, and transcripts, with some further supplemented with affective states, uncertainty, and brief facial and scene descriptions, and the first three clips serving as input for predicting Clip IV's emotion from Ekman's six basic emotion categories (Ekman, 1992) plus neutral. Here, we use the 2,830 publicly released samples available to us as the foundation for EmoEvolve.

To provide structured, evidence-grounded, and transition-aware supervision for emotion forecasting, we first compile an evidence bundle for each preceding clip in Clips I-III. Specifically, Gemini 3.1 Flash-Lite processes the audio of each clip to describe speaking rate, pitch, timbre, intensity, pauses, and rhythm, capturing vocal cues that potentially convey emotion beyond lexical content. In parallel, Qwen3.5-397B-A17B describes facial and bodily behavior, scenes, and interactions from video, capturing nonverbal affective cues and the surrounding situational context. Both models are prompted to prioritize objective, observable descriptions rather than directly assign emotion labels, with detailed prompts provided in the Appendix. Subsequently, these descriptions are combined with the corresponding transcripts and available auxiliary annotations on affect, uncertainty, and context to form clip-level evidence bundles.

We then provide the above bundles, together with the observed-state labels of Clips II and III and the target emotion label of Clip IV, to GPT-5.4 to generate structured forecasting rationales, each containing `<history>`, `<cross>`, `<think>`, and `<answer>` fields that record the historical reference state, preceding interlocutor state, evidence-grounded reasoning, and target emotion, respectively. The `<think>` span comprises five components: *Observed Dialogue*, *Observed Multimodal Evidence*, *Historical Evidence*, *Cross Evidence*, and *Context-response Comparison*, progressively organizing observable multimodal cues, relating them to the historical and preceding-interlocutor states, and comparing their implications for the forthcoming emotional response. Thereafter, a separate judging call evaluates evidence consistency and multimodal coverage on five-point scales, retaining candidates that score at least three on both criteria and triggering up to two regeneration attempts otherwise, after which experienced domain experts inspect and revise the outputs to produce the final structured annotations. Throughout annotation, Clip-IV video, audio, transcript, and descriptive annotations are excluded from the evidence bundles; only its emotion label is used

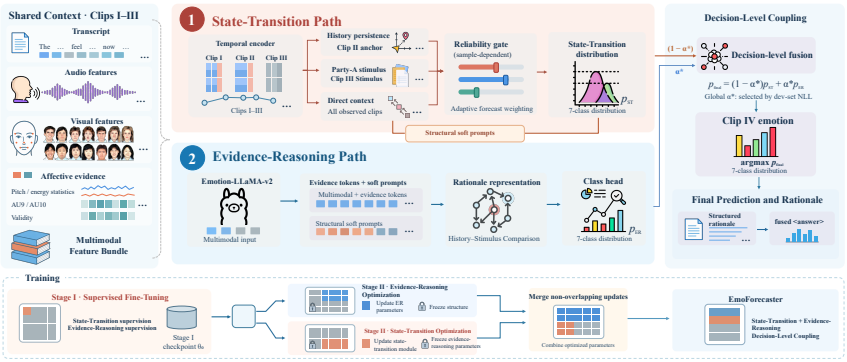


Figure 3: Overview of EmoForecaster, comprising complementary state-transition and evidence-reasoning paths with decision-level coupling and a two-stage optimization strategy.

to guide rationale construction, and none of the auxiliary or target emotion labels are supplied at inference. The construction pipeline is illustrated in Figure 2.

Leveraging the constructed EmoEvolve, we establish a clip-disjoint protocol for evaluating both emotion forecasting and rationale generation by partitioning the 2,830 samples into a training pool of 1,980 samples and a held-out test set of 850 samples, while assigning samples that share any source clip, regardless of its position within an MCIS, to the same partition to prevent shared-clip leakage. We further reserve 299 samples from the training pool for development, leaving 1,681 samples for model fitting under the same clip-disjoint constraint. For emotion forecasting, we report weighted average recall (WAR), unweighted average recall (UAR), and Macro-F1. For rationale generation, we compare each generated `<think>` span with its corresponding reference using BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee & Lavie, 2005), and BERTScore-F1 (Zhang et al., 2019) to measure lexical and semantic similarity. Higher values indicate better performance for all metrics.

2.2 EmoForecaster

Given three observed clips $X_{1:3} = (X_1, X_2, X_3)$, where $X_t = (T_t, A_t, V_t)$ contains the transcript, audio, and video, EmoForecaster aims to predict Party B's emotion $y_4 \in \mathcal{Y}$ in the unseen Clip IV and generate a corresponding rationale, with $|\mathcal{Y}| = 7$. To connect emotional transition modeling with multimodal evidence, we introduce complementary state-transition and evidence-reasoning paths coupled at the decision level, where both paths build on Emotion-LLaMA-v2 (Peng et al., 2026).

To retain explicit affective cues alongside learned representations, we additionally extract a shared 14-dimensional evidence vector $\mathbf{e}_t$ for each observed clip. It contains six pitch and energy statistics to characterize vocal prosody and intensity variation, the mean, 90th percentile, and maximum activations of each of AU9 and AU10 to capture salient local facial movements, and two modality-validity indicators to account for unreliable or unavailable cues. Continuous evidence features are normalized using training-set statistics. The state-transition path concatenates $\mathbf{e}_t$ with the pooled textual, audio, and visual features, while the evidence-reasoning path projects it into an additional evidence token for language decoding.

State-Transition Path. To model emotional dynamics across the observed interaction, we concatenate the mean-pooled audio and visual features $\mathbf{a}_t, \mathbf{v}_t$ of each clip with the evidence vector $\mathbf{e}_t$, together with the textual feature $\mathbf{t}_t$. A learned projection and a two-layer temporal Transformer encoder are then used to contextualize these evidence-enhanced clip representations:

$$\mathbf{x}_t = \phi([\mathbf{t}_t; \mathbf{a}_t; \mathbf{v}_t; \mathbf{e}_t]) + \mathbf{r}_t, \quad \mathbf{h}_{1:3} = \mathrm{TrEnc}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3), \quad \mathbf{c} = \frac{1}{3}\sum_{t=1}^{3} \mathbf{h}_t, \tag{1}$$

where $[\cdot;\cdot]$ denotes concatenation, ϕ is the projection, $\mathbf{r}_t$ encodes temporal position rather than speaker identity, and the pooled context $\mathbf{c}$ summarizes all three clips for subsequent state-transition modeling.

To distinguish the predictive implications of the historical reference in Clip II and the preceding interlocutor in Clip III, denoted by h and s, we use the contextualized states $\mathbf{h}_2$ and $\mathbf{h}_3$ to estimate

their emotional states and map them to future-emotion distributions:

$$\mathbf{q}_h = \mathrm{softmax}(W_h \mathbf{h}_2), \quad \mathbf{p}_h = \mathbf{q}_h M_h(\mathbf{c}), \quad \mathbf{q}_s = \mathrm{softmax}(W_s \mathbf{h}_3), \quad \mathbf{p}_s = \mathbf{q}_s M_s(\mathbf{c}). \tag{2}$$

Here W_h, W_s are auxiliary state heads, where $\mathbf{q}_h$ and $\mathbf{q}_s$ estimate the emotional states of the historical reference and the preceding interlocutor, respectively. The context-conditioned transition matrices $M_h(\mathbf{c}), M_s(\mathbf{c}) \in \mathbb{R}^{7\times 7}$ are nonnegative and row-normalized, mapping these estimated states to the corresponding forecasts $\mathbf{p}_h$ and $\mathbf{p}_s$ of the forthcoming response. Additionally, a direct contextual forecast $\mathbf{p}_d = \mathrm{softmax}(W_d \mathbf{c})$ provides a route without intermediate state prediction.

Since the relevance of these forecasts varies across interactions, a learned gate maps $[\mathbf{h}_2; \mathbf{h}_3; \mathbf{c}]$ to nonnegative weights (w_d, w_h, w_s) summing to one:

$$\mathbf{p}_{\mathrm{ST}} = w_d \mathbf{p}_d + w_h \mathbf{p}_h + w_s \mathbf{p}_s = w_d \mathbf{p}_d + (1 - w_d)\mathbf{p}_h + g \underbrace{(\mathbf{p}_s - \mathbf{p}_h)}_{\Delta_{\mathrm{ST}}}, \tag{3}$$

where $g = w_s$ and $0 \le g \le 1 - w_d$. The residual form expresses an interlocutor-conditioned adjustment relative to the historical reference. Because both forecasts use the full observed context, Δ_{ST} is a conditional predictive contrast, not an identified causal effect.

Evidence-Reasoning Path. To provide an explicit interpretation of the evidence supporting the forecast, this path receives the transcripts and original multimodal tokens of Clips I–III, together with three evidence tokens projected from $\mathbf{e}_{1:3}$ and three structural soft prompt tokens derived from $\mathbf{q}_h$, $\mathbf{q}_s$, and $\mathbf{p}_{\mathrm{ST}}$. In this way, the decoder jointly accesses the observed multimodal evidence, the estimated states of the historical reference and preceding interlocutor, and the state-transition forecast in the language space.

The LoRA-adapted decoder (Hu et al., 2021a) follows the structured format defined in EmoEvolve, consisting of `<history>`, `<cross>`, `<think>`, and `<answer>`. The first two fields verbalize the inferred historical-reference and preceding-interlocutor states, while the five-part `<think>` span progressively organizes the observed multimodal evidence and compares its implications for the forthcoming response through *Context-response Comparison*.

To further allow the reasoning representation to contribute directly to emotion forecasting, we extract the hidden state $\mathbf{z}$ at the end of the rationale, before generating `<answer>`, and obtain

$$\mathbf{p}_{\mathrm{ER}} = \mathrm{softmax}(W_{\mathrm{ER}}\, \mathrm{LN}(\mathbf{z}))\,, \tag{4}$$

where LN denotes layer normalization and W_{ER} is a seven-class projection. The evidence-reasoning path therefore produces both an explicit textual rationale and a reasoning-conditioned emotion distribution.

Decision-Level Coupling. To integrate the complementary predictions from the two paths, we fuse their output distributions:

$$\mathbf{p}_{\mathrm{final}} = (1 - \alpha)\mathbf{p}_{\mathrm{ST}} + \alpha \mathbf{p}_{\mathrm{ER}}, \quad \widehat{y}_4 = \arg\max_{y \in \mathcal{Y}} \mathbf{p}_{\mathrm{final}}(y). \tag{5}$$

Here, the global fusion coefficient $\alpha \in [0, 1]$ is selected on the development set. The generated rationale is retained, and $\widehat{y}_4$ is serialized as the final `<answer>`. At inference, the model uses only Clips I–III, without access to gold emotion labels or reference rationales.

2.3 Training Strategy

We optimize EmoForecaster through supervised fine-tuning (SFT) for state estimation, emotion forecasting, and rationale generation, followed by path-specific policy optimization with group-relative rewards and supervised anchors.

Stage I: Supervised Fine-Tuning. To establish the state-transition predictor before introducing rationale supervision, we train it for three epochs and then jointly optimize both paths for two additional epochs. The state-transition objective combines future-emotion prediction with auxiliary supervision of the historical reference and preceding interlocutor states:

$$\mathcal{L}_{\mathrm{ST}} = -\log \mathbf{p}_{\mathrm{ST}}(y_4) + \lambda_s \Big[\omega_2\, \mathrm{CE}(\mathbf{q}_h, y_2) + \omega_3\, \mathrm{CE}(\mathbf{q}_s, y_3)\Big], \tag{6}$$

where y_2, y_3 are the observed-state labels, y_4 is the Clip-IV target, and CE denotes cross-entropy. The coefficient λ_s controls auxiliary supervision, while ω_2, ω_3 weight the two state objectives.

To couple rationale generation with predictive learning, the evidence-reasoning objective combines weighted token supervision, pre-answer classification, and distributional alignment:

$$\begin{aligned} \mathcal{L}_{\text{tok}} &= -\frac{1}{\sum_{\ell=1}^{L} \eta_\ell} \sum_{\ell=1}^{L} \eta_\ell \log \pi_\theta(u_\ell \mid X_{1:3}, u_{<\ell}), \\ \mathcal{L}_{\text{ER}} &= \mathcal{L}_{\text{tok}} + \text{CE}(\mathbf{p}_{\text{ER}}, y_4) + \lambda_a D_{\text{KL}}(\mathbf{p}_{\text{ER}} \parallel \text{sg}[\mathbf{p}_{\text{ST}}]) \,. \end{aligned} \tag{7}$$

Here $u_{1:L}$ is the complete structured reference from EmoEvolve, π_θ is the decoder policy, and η_ℓ assigns greater weight to answer tokens. The coefficient λ_a controls alignment, with $\text{sg}[\cdot]$ stopping gradients through its target. During teacher forcing, $\mathbf{p}_{\text{ER}}$ is computed from $\mathbf{z}$ before any reference answer token is read.

To reduce interference between forecasting and rationale supervision, we apply asymmetric gradient projection (Yu et al., 2020) to parameters θ_{sh} receiving gradients from both objectives. With $\mathbf{g}_{\text{ST}} = \nabla_{\theta_{\text{sh}}} \mathcal{L}_{\text{ST}}$ and $\mathbf{g}_{\text{ER}} = \nabla_{\theta_{\text{sh}}} \mathcal{L}_{\text{ER}}$, we compute

$$\widetilde{\mathbf{g}}_{\text{ER}} = \mathbf{g}_{\text{ER}} - \frac{\min\{0, \langle \mathbf{g}_{\text{ER}}, \mathbf{g}_{\text{ST}} \rangle\}}{\|\mathbf{g}_{\text{ST}}\|_2^2 + \epsilon} \mathbf{g}_{\text{ST}}, \quad \mathbf{g} = \mathbf{g}_{\text{ST}} + \lambda_{\text{ER}} \widetilde{\mathbf{g}}_{\text{ER}}. \tag{8}$$

This reduces conflicting updates to the state-transition objective when gradients conflict. The constant ϵ stabilizes the denominator, and λ_{ER} controls the reasoning update; parameters receiving gradients from only one objective are updated without projection.

Stage II: Path-Specific Policy Optimization. To refine each path through task-specific feedback, we initialize two separate runs from the same Stage-I checkpoint. One updates only the state-transition module, while the other updates the decoder LoRA parameters and reasoning-state classifier, with all remaining parameters fixed. Both use group-relative advantage estimation inspired by GRPO (Shao et al., 2024), path-specific rewards, and supervised anchors.

Evidence-reasoning reward. To favor rationales that support accurate forecasts and observed-state estimates, a frozen rollout policy generates candidate structured outputs $\{o_j\}_{j=1}^{G}$ for each input. Their pre-answer states yield $\mathbf{p}_{\text{ER},j}$, which are fused with the frozen state-transition distribution:

$$\mathbf{p}_j^R = (1 - \alpha_R)\mathbf{p}_{\text{ST}} + \alpha_R \mathbf{p}_{\text{ER},j}, \quad \widehat{y}_j = \arg\max_{y \in \mathcal{Y}} \mathbf{p}_j^R(y), \quad \Delta_j = \log \frac{\mathbf{p}_j^R(y_4)}{\mathbf{p}_{\text{ST}}(y_4)}. \tag{9}$$

Here $\alpha_R = 0.5$ is a training-only coefficient, distinct from the development-selected α. The candidate reward is

$$R_{\text{ER},j} = R_4(\widehat{y}_j, y_4) + \lambda_p \,\text{clip}_{[-1,1]}(\Delta_j) + \lambda_h R_h(o_j, y_2) + \lambda_s^R R_s(o_j, y_3) + \lambda_f R_f(o_j). \tag{10}$$

The term R_4 rewards correct forecasts with capped inverse-frequency class weights, and Δ_j measures the ground-truth log-probability gain over the structural forecast. The rewards R_h, R_s evaluate the generated `<history>` and `<cross>` states against y_2, y_3, while R_f checks format compliance. Their coefficients balance the respective contributions.

State-transition rewards. To improve forecasting while preserving intermediate state estimates, we optimize three categorical policies: $\pi_\theta^{(4)} = \mathbf{p}_{\text{ST}}$, $\pi_\theta^{(h)} = \mathbf{q}_h$, and $\pi_\theta^{(s)} = \mathbf{q}_s$, all conditioned on $X_{1:3}$. Each forms a separate seven-action candidate group rewarded against y_4, y_2, or y_3, respectively, with independently computed group advantages and weighted auxiliary objectives.

Group-relative objective. To compare candidates for the same input, we standardize group rewards and clip the resulting advantages:

$$\widehat{A}_j = \text{clip}_{[-2,2]} \left(\frac{R_j - \overline{R}}{\sqrt{\text{Var}(R) + \epsilon}} \right), \tag{11}$$

where $\overline{R}$ and $\text{Var}(R)$ are the group reward mean and variance. The rollout policy π_{old}, rewards, and advantages remain fixed while optimizing each group. For variable-length rationale outputs, we

compute the policy ratio using the mean token log-likelihood over each candidate sequence:

$$\rho_j^{\mathrm{ER}} = \exp\left(\frac{1}{L_j}\sum_{\ell=1}^{L_j}\log\frac{\pi_\theta(o_{j,\ell} \mid X_{1:3}, o_{j,<\ell})}{\pi_{\mathrm{old}}(o_{j,\ell} \mid X_{1:3}, o_{j,<\ell})}\right), \tag{12}$$

where L_j is the candidate output length. For categorical actions in state-transition group $k \in \{4, h, s\}$, the ratio is

$$\rho_j^{(k)} = \frac{\pi_\theta^{(k)}(o_j \mid X_{1:3})}{\pi_{\mathrm{old}}^{(k)}(o_j \mid X_{1:3})}. \tag{13}$$

Using the corresponding ratio, we discourage large policy changes through a clipped surrogate:

$$\mathcal{J} = -\frac{1}{G}\sum_{j=1}^{G}\min\left\{\rho_j\widehat{A}_j,\ \mathrm{clip}(\rho_j, 1-\varepsilon, 1+\varepsilon)\widehat{A}_j\right\}, \tag{14}$$

where ε defines clipping interval. Supervised anchors preserve the capabilities learned in Stage I:

$$\begin{aligned}\mathcal{L}_{\mathrm{ER}}^{\mathrm{II}} &= \mathcal{J}_{\mathrm{ER}} + \lambda_{\mathrm{sup}}\mathcal{L}_{\mathrm{ER}},\\ \mathcal{L}_{\mathrm{ST}}^{\mathrm{II}} &= \mathcal{J}_4 + \lambda_{\mathrm{state}}^{\mathrm{RL}}(\mathcal{J}_h + \mathcal{J}_s) + \lambda_{\mathrm{sup}}\mathcal{L}_{\mathrm{ST}}.\end{aligned} \tag{15}$$

Here $\mathcal{J}_{\mathrm{ER}}$ is the rationale surrogate, while $\mathcal{J}_4, \mathcal{J}_h, \mathcal{J}_s$ correspond to the three categorical policies. The coefficients $\lambda_{\mathrm{state}}^{\mathrm{RL}}$ and λ_{sup} weight auxiliary state optimization and supervised anchoring. With rewards and advantages detached, the rationale surrogate updates decoder LoRA, while the reasoning-state classifier learns through its supervised classification and alignment terms.

Model selection and refitting. We merge the non-overlapping updates from both runs and generate fresh rationales using the updated structural soft prompts. For each merged configuration, the global fusion coefficient minimizes development-set negative log-likelihood:

$$\alpha^* = \arg\min_{\alpha\in[0,1]} -\frac{1}{|\mathcal{D}_{\mathrm{dev}}|}\sum_{i\in\mathcal{D}_{\mathrm{dev}}}\log \mathbf{p}_{\mathrm{final}}^{(i)}(y_4^{(i)}; \alpha), \tag{16}$$

where $\mathbf{p}_{\mathrm{final}}^{(i)}(\cdot; \alpha)$ is defined in Equation 5. Selection uses a clip-disjoint development subset and freely generated rationales without teacher forcing. After fixing the checkpoint configuration, training schedule, and α^*, we refit on the complete training pool and evaluate the held-out test set.

3 Experiments

3.1 Implementation Details

We implement EmoForecaster using Emotion-LLaMA-v2 (Peng et al., 2026) as the multimodal backbone on one NVIDIA RTX A6000 (48GB) GPU. The language decoder is adapted through LoRA (Hu et al., 2021a) with rank $r = 64$ and scaling parameter $\alpha_{\mathrm{LoRA}} = 16$. We use the AdamW optimizer with a learning rate of 3×10^{-7}, a batch size of 1, and 8 gradient accumulation steps. Stage I comprises three epochs of state-transition training followed by two epochs of joint supervision with asymmetric gradient projection. Stage II initializes two separate path-specific optimization runs from the same Stage-I checkpoint, using 256 training steps and a GRPO group size of $G = 7$ while retaining supervised anchors. The training-only reward fusion coefficient is fixed at $\alpha_R = 0.5$. After merging the non-overlapping parameter updates, we generate fresh rationales with the resulting model. We reserve 299 of the 1,980 training-pool samples for clip-disjoint development, leaving 1,681 for model fitting. The development set is used for model selection and hyperparameter tuning, after which we refit the selected configuration on the complete training pool and evaluate the 850 held-out test samples.

3.2 Quantitative Comparison

We compare EmoForecaster with representative general-purpose MLLMs, including Qwen3-VL-8B-Thinking, MiniCPM-o-2.6-8B, InternVL3.5-8B, and Qwen2.5-Omni-7B, as well as emotion-capable MLLMs, including AffectGPT and Emotion-LLaMA-v2, on the same 850-example test set

Table 1: Comparison with representative general-purpose and emotion-capable MLLMs on EmoEvolve. Emotion forecasting scores are reported as percentages, while rationale-similarity scores are multiplied by 100 for display. Higher values indicate better performance.

	Emotion forecasting (%)			Rationale similarity ($\times$100)			
Method	WAR ↑	UAR ↑	Macro-F1 ↑	BLEU-4 ↑	ROUGE-L ↑	METEOR ↑	BERTScore-F1 ↑
General-purpose MLLMs							
Qwen3-VL-8B-Thinking (Bai et al., 2025)	24.35	16.73	13.30	3.32	19.93	14.04	13.69
MiniCPM-o-2.6-8B (Yao et al., 2025)	23.06	16.48	11.10	3.80	21.54	16.45	12.51
InternVL3.5-8B (Wang et al., 2025)	17.53	14.31	12.28	4.64	20.75	19.51	10.31
Qwen2.5-Omni-7B Xu et al. (2025)	25.53	16.76	11.68	5.79	21.43	17.81	11.16
Emotion-capable MLLMs							
AffectGPT (Lian et al., 2025)	25.06	17.40	14.67	14.24	24.98	28.52	10.38
Emotion-LLaMA-v2 (Peng et al., 2026)	25.76	16.44	11.40	10.50	23.17	23.74	8.36
Emotion-LLaMA-v2 (1-shot)	25.06	15.73	10.21	9.50	21.35	24.57	10.90
Emotion-LLaMA-v2 (3-shot)	24.82	15.42	9.25	11.20	23.21	24.40	12.40
EmoForecaster (Ours)							
Supervised Fine-Tuning (Stage I)	39.06	25.57	22.08	17.91	32.87	31.79	24.33
Path-Specific Policy Optimization (Stage II)	**43.18**	**29.51**	**24.28**	**19.51**	**33.76**	**33.18**	**28.72**

Table 2: Ablation study of the state-transition and evidence-reasoning paths under supervised fine-tuning and path-specific policy optimization.

	Emotion forecasting (%)			Rationale similarity ($\times$100)			
Variants	WAR ↑	UAR ↑	Macro-F1 ↑	BLEU-4 ↑	ROUGE-L ↑	METEOR ↑	BERTScore-F1 ↑
Baseline	34.71	23.79	20.03	11.08	23.27	24.16	9.18
Stage I: Supervised Fine-Tuning							
State-Transition	37.18	24.24	20.92	12.50	23.30	24.04	8.56
Evidence-Reasoning	35.41	24.53	21.80	16.23	30.02	31.03	23.65
Complete	39.06	25.57	22.08	17.91	32.87	31.79	24.33
Stage II: Path-Specific Policy Optimization							
State-Transition	42.71	29.28	23.98	16.60	32.88	32.01	24.76
Evidence-Reasoning	43.06	29.36	24.12	20.10	31.43	33.41	25.30
Complete	43.18	29.51	24.28	19.51	33.76	33.18	28.72

of EmoEvolve, as summarized in Table 1. It can be observed that general-purpose MLLMs struggle with evidence-grounded multimodal emotion forecasting, potentially reflecting their limited affective understanding and insufficient sensitivity to emotion-specific patterns required for anticipating future responses. Emotion-capable MLLMs benefit from specialized optimization for multimodal emotion recognition and understanding, achieving better agreement with the reference rationales on several rationale-similarity metrics and demonstrating that they can capture some relevant emotional information. Yet, their forecasting performance remains limited, highlighting the gap between interpreting observed emotions and anticipating future emotional responses. Furthermore, we conduct few-shot prompting experiments (1-shot and 3-shot) on Emotion-LLaMA-v2 to examine whether in-context demonstrations can facilitate multimodal emotion forecasting, but neither setting improves performance, suggesting that limited in-context guidance is insufficient to adapt its recognition-oriented capabilities to future-emotion prediction.

In contrast, EmoForecaster outperforms the above models after Stage I in both emotion forecasting and rationale similarity, benefiting from structured supervision that jointly learns state-transition modeling and evidence-grounded rationale generation. Subsequent path-specific policy optimization further improves all reported metrics by separately refining the state-transition and evidence-reasoning paths with task-specific group-relative rewards while preserving their supervised capabilities, demonstrating the effectiveness of the complete framework in predicting forthcoming emotions and generating corresponding rationales.

3.3 Qualitative Analysis

We qualitatively compare the forecasts and rationales generated by different MLLMs using the representative example in Figure 4. It can be seen that the general-purpose Qwen2.5-Omni-7B recognizes the solution-oriented proposal in Clip III but interprets the dissatisfaction in Clip II as anger and overemphasizes its persistence, resulting in an incorrect angry prediction. The emotion-capable Emotion-LLaMA-v2 provides a more nuanced description of the observed emotional expression, citing the restrained tone, lowered brows, and guarded expression, yet still treats the earlier aversion as dominant and incorrectly predicts disgust. These responses illustrate that identifying emotion-related information does not necessarily enable a model to anticipate how the interaction may alter the forthcoming emotional response. In contrast, EmoForecaster progressively examines the his-

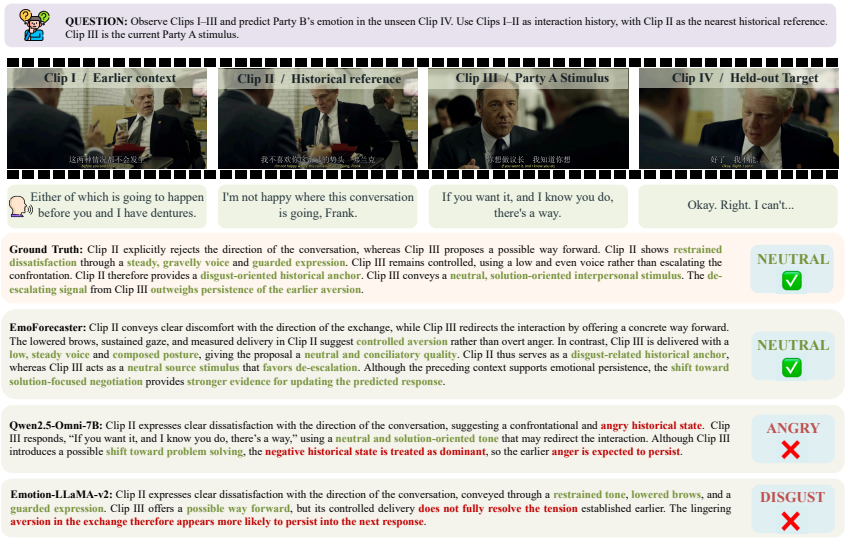


Figure 4: Qualitative comparison of EmoForecaster and representative MLLMs on a multimodal emotion forecasting example, highlighting differences in evidence interpretation, transition reasoning, and final prediction.

torical reference and preceding interlocutor state, distinguishing the controlled aversion in Clip II from the neutral, conciliatory proposal in Clip III. Its rationale connects prosodic and facial cues with the conversational context and explicitly weighs historical emotional continuity against the potential de-escalation introduced by the proposal, ultimately predicting neutral in agreement with the ground-truth label. This example illustrates how EmoForecaster connects multimodal evidence with an explicit interpretation of emotional change rather than simply carrying the preceding negative state forward.

3.4 Ablation Study

We conduct ablation studies of the proposed components using direct supervised fine-tuning of Emotion-LLaMA-v2 as the baseline, as shown in Table 2. In Stage I, the state-transition path brings relatively larger gains in overall prediction accuracy, whereas the evidence-reasoning path more effectively supports reference-consistent explanation generation while also improving emotion forecasting. Combining both paths further strengthens these capabilities overall, suggesting complementary benefits from modeling emotional transitions and interpreting multimodal evidence. In Stage II, separately optimizing either path benefits both emotion forecasting and several aspects of rationale generation, and merging the two path-specific updates yields additional predictive gains and synergistic improvements in the generated explanations. Together, these findings demonstrate the effectiveness of jointly modeling the two paths and integrating their separately optimized updates to enhance both future-emotion prediction and corresponding rationale generation.

4 Conclusion and Future Work

In this paper, we propose EmoForecaster, an evidence-grounded multimodal emotion forecasting framework that anticipates future emotional responses through emotional transition modeling and progressive multimodal reasoning. We construct a dataset, EmoEvolve, with structured interaction reasoning annotations and establish a clip-disjoint evaluation protocol covering both emotion forecasting and rationale generation. Building on this, EmoForecaster integrates complementary state-transition and evidence-reasoning paths through decision-level coupling, with supervised fine-tuning followed by path-specific policy optimization. Extensive experiments demonstrate that EmoForecaster consistently outperforms representative general-purpose and emotion-capable MLLMs in both emotion forecasting and rationale generation, while ablation studies support the complementary contributions of the two paths and their separately optimized updates. Future work will explore longer and more diverse interactions to advance the practical applicability of multimodal emotion forecasting in real-world interactive scenarios.

AI USE STATEMENT

AI tools were used to assist the construction of EmoEvolve and the polishing of this manuscript. Specifically, Gemini 3.1 Flash-Lite and Qwen3.5-397B-A17B supported audio and visual evidence extraction, respectively, while GPT-5.4 generated structured forecasting rationales. AI assistance was also used to improve the language and clarity of the manuscript.

REPRODUCIBILITY STATEMENT

The main paper and Appendix provide sufficient details on dataset construction, annotation prompts, model architecture, training objectives, and experimental settings to support reproducibility, and we also plan to publicly release the EmoEvolve dataset and implementation code to facilitate independent evaluation and future research.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024*, pp. 13590–13618, 2024.

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

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024b.

Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.

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

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1122–1131, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021a.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021b.

Dacher Keltner and Jonathan Haidt. Social functions of emotions at four levels of analysis. *Cognition & emotion*, 13(5):505–521, 1999.

Dayu Li, Yang Li, and Suge Wang. Interactive double states emotion cell model for textual dialogue emotion prediction. *Knowledge-Based Systems*, 189:105084, 2020.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. Emotion inference in multi-turn conversations with addressee-aware module and ensemble strategy. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3935–3941, 2021.

Yuanchao Li, Koji Inoue, Leimin Tian, Changzeng Fu, Carlos Toshinori Ishi, Hiroshi Ishiguro, Tatsuya Kawahara, and Catherine Lai. I know your feelings before you do: Predicting future affective reactions in human-computer dialogue. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2023.

Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, et al. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*, 2023.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2247–2256, 2018.

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

Xiaojiang Peng, Jingyi Chen, Zebang Cheng, Bao Peng, Fengyi Wu, Yifei Dong, Shuyuan Tu, Qiyu Hu, Huiting Huang, Yuxiang Lin, et al. Emotion-llamav2 and mmeverse: A new framework and benchmark for multimodal emotion understanding. *arXiv preprint arXiv:2601.16449*, 2026.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.

Klaus R Scherer. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351, 2009.

Michelle Schlicher, Yupei Li, Sunil Munthumoduku Krishna Murthy, Qiyang Sun, and Björn W Schuller. Emotionally adaptive support: a narrative review of affective computing for mental health. *Frontiers in Digital Health*, 7:1657031, 2025.

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

Haoxiang Shi, Ziqi Liang, and Jun Yu. Emotional cues extraction and fusion for multi-modal emotion prediction and recognition in conversation. *arXiv preprint arXiv:2408.04547*, 2024.

Haoran Wang, Xinji Mai, Zeng Tao, Junxiong Lin, Xuan Tong, Ivy Pan, Shaoqi Yan, Yan Wang, and Shuyong Gao. Hi-ef: Benchmarking emotion forecasting in human-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2110–2118, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction. *Frontiers in Computer Science*, 3:767767, 2021.

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

Qinfu Xu, Liyuan Pan, Yiwei Wei, Shaozu Yuan, Jiaqi Chen, and Tianyu Liu. Emothinker: Advancing visual-acoustic emotion analysis via structural token selection and chain-of-thought reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1672–1682, 2026.

Elaheh Yadegaridehkordi, Nurul Fazmidar Binti Mohd Noor, Mohamad Nizam Bin Ayub, Hannyzzura Binti Affal, and Nornazlita Binti Hussin. Affective computing in education: A systematic review and future research. *Computers & education*, 142:103649, 2019.

Yuan Yao, Tianyu Yu, Chongyi Wang, Junbo Cui, Bokai Xu, Hongji Zhu, Tianchi Cai, Fuwei Huang, Tianran Wang, Wenshuo Ma, et al. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

A Related Work

Multimodal Emotion Recognition and Understanding. Human emotions are manifested through interrelated multimodal cues, including linguistic content, vocal prosody, and facial movements, making multimodal emotion recognition an important topic in affective computing (Scherer, 2009; Poria et al., 2017; D'mello & Kory, 2015). Early studies largely followed a discriminative paradigm that learned and fused features from different modalities for emotion or sentiment classification. For example, MISA (Hazarika et al., 2020) projects each modality into modality-invariant and modality-specific subspaces to capture commonalities across modalities and characteristic features of each modality before fusion. With the rapid development of multimodal large language models (MLLMs), which have demonstrated strong capabilities in multimodal understanding and reasoning (Caffagni et al., 2024; Wang et al., 2024), recent research has begun to move beyond categorical prediction toward more fine-grained and interpretable emotion understanding. Emotion-LLaMA (Cheng et al., 2024a) integrates audio, visual, and textual inputs through emotion-specific encoders, aligns their features into a shared space, and employs instruction tuning to enhance emotion recognition and reasoning, while AffectGPT (Lian et al., 2025) leverages large-scale descriptive emotion annotations and pre-fusion operations to enhance multimodal integration for complex emotion understanding. More recently, EmoThinker (Xu et al., 2026) introduces structural token selection and audio evidence extraction together with Chain-of-Emotion-Thought reasoning for more interpretable visual-acoustic emotion analysis. Despite these advances, existing methods predominantly focus on recognizing and explaining emotions that have already been expressed, whereas proactive multimodal affective systems should also anticipate future emotional states to better organize subsequent responses.

Emotion Forecasting. Unlike emotion recognition, which identifies affective states from observed expressions, emotion forecasting predicts a participant's forthcoming emotional state from the preceding interaction context before the response is observed, enabling interactive systems to prepare more timely and proactive responses (Li et al., 2023; 2021). Early studies largely focused on text-based conversations. Li et al. (2020) proposed an interactive double states emotion cell model that captures context dependence, emotional persistence, and contagiousness in textual dialogues to predict an interlocutor's future emotional state without observing the forthcoming utterance. Subsequent work incorporated acoustic information into conversational emotion forecasting. Shi et al. (2024) proposed an emotional cues extraction and fusion network that combines word-level relation modeling and prosody learning to forecast the emotion of forthcoming utterances from textual and acoustic cues. Beyond these settings, richer multimodal cues have further been explored to model affective dynamics. Wu et al. (2021) extended the Dynamic Ordinal Markov Model to multimodal inputs to jointly model emotion states and emotion changes from audio and visual cues, while Hi-EF (Wang et al., 2026) further leveraged video, audio, and text from preceding interaction clips to capture multimodal contextual information and forecast the emotion expressed in the subsequent interaction. Nevertheless, existing approaches mainly focus on forecasting emotion categories, with limited explicit reasoning about how interaction history and multimodal evidence support the predicted future state. To address the above limitation, EmoForecaster jointly models interaction-conditioned state transitions and evidence-grounded multimodal reasoning, enabling future emotion predictions to be accompanied by explicit reasoning grounded in the observed interaction.

B Detailed Prompts

We construct the released SFT data from the transcripts, multimodal observations, and annotations provided by Hi-EF Wang et al. (2026). The construction pipeline first extracts acoustic and visual observations from each observed clip. GPT-5.4 then integrates these observations with the dialogue and dataset annotations to produce the structured SFT target.

Audio Observation Prompt

```
You are a speech perception analyst. Describe only directly audible acoustic
cues: speaking rate, pitch level and contour, loudness and intensity, voice
quality, pauses or hesitations, rhythm, and non-speech vocalizations. Note
meaningful changes and uncertainty when cues are weak. Do not infer emotion,
intent, identity, or dialogue content.
```

Visual Observation Prompt

```
You are a visual analyst. The input contains three chronologically ordered,
uniformly sampled keyframes from one interaction clip. Describe only visible
cues: brows, eyes and gaze, mouth and lips, head pose, posture or gesture,
scene, and interaction. Separate stable appearance from changes supported
across frames; do not infer motion from one frame. Do not infer emotion,
intent, speaker identity, or unseen events.
```

GPT-5.4 receives the transcripts, acoustic observations, visual observations, and available Hi-EF annotations for Clips I–III. During offline SFT-data construction, it also receives the gold Clip-IV emotion label as the supervision target. Clip-IV video, audio, and transcript are not provided.

Reference Rationale Construction Prompt

```
You are constructing an evidence-grounded reference rationale for multimodal
emotion forecasting.

OBSERVED INPUT:
- Transcripts from Clips I--III
- Acoustic observations from Clips I--III
- Visual observations from Clips I--III
- Available Hi-EF annotations, including facial movement, scene context,
  interaction polarity, interaction intensity, and annotation confidence

CONSTRUCTION LABELS:
- y2: emotion of the Clip-II historical anchor
- y3: emotion of the Clip-III Party-A stimulus
- y4: gold emotion of the unseen Clip-IV Party-B response

TASK:
Integrate the observed dialogue, acoustic evidence, visual evidence, and
dataset annotations into a coherent English rationale.

Describe only evidence observed in Clips I--III. Explain the emotional
continuity supported by the Clip-II historical anchor, the response change
supported by the Clip-III stimulus, and the comparison between these two
signals. Do not describe Clip IV as observed. Avoid repeated evidence,
generic emotion definitions, and unsupported visual or acoustic details.

OUTPUT FORMAT:
- Write the Clip-II emotion y2 inside <history>...</history>.
- Write the Clip-III emotion y3 inside <cross>...</cross>.
- Write the evidence-grounded rationale inside <think>...</think>.
- Write the sole final Clip-IV emotion y4 inside <answer>...</answer>.

Inside <think>, use the following headings exactly once and in this order:

Observed dialogue:
Observed multimodal evidence:
Historical evidence:
Cross evidence:
Context-response comparison:

Return only <history>, <cross>, <think>, and <answer>.
```

The symbolic labels are replaced with sample-specific emotion labels during construction. The released SFT records contain concrete fields such as `<history>disgust</history>`, `<cross>neutral</cross>`, and `<answer>neutral</answer>`.

The released data use 12 semantically equivalent task instructions to reduce dependence on one question formulation. One representative instruction is shown below.

Multimodal Emotion Forecasting Instruction

```
Using only observed Clips I--III, forecast the emotion that Party B will
express in the unseen Clip IV. Treat Clips I and II as role-agnostic
interaction history, use Clip II as the nearest affective anchor, and treat
Clip III as the current Party-A stimulus.

Output exactly <history>, <cross>, <think>, and <answer>. Inside <think>,
use these headings exactly once and in this order:

Observed dialogue:
Observed multimodal evidence:
Historical evidence:
Cross evidence:
Context-response comparison:

The <answer> field must contain the sole final emotion class.
```