EQ-AGENT: FROM MULTIMODAL EMOTION RECOGNITION TO EMOTIONALLY INTELLIGENT AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent foundation models have predominantly emphasized "IQ-like" capabilities—reasoning, mathematics, coding—while emotional intelligence (EQ) remains comparatively underexplored in the pursuit of artificial general intelligence (AGI). Multimodal emotion recognition (MER) has advanced affect perception but stays perception-centric: recognizing an emotion alone does not let an agent understand context, plan support, or respond effectively. We propose **EQ-Agent**, a framework that extends affective AI from the MER paradigm to a complete EQ loop spanning perception, understanding, planning, and response. EQ-Agent first establishes **EQ-MER**, a sub-1B perception foundation built on a Perception-to-Cognition (P2C) Bridge architecture and a two-stage training recipe. We then build **EQ-Dataset**, a 30,664-record multitask resource integrating social reasoning, supportive response, and structured emotional-reasoning supervision to bridge emotion labels and emotionally informed decisions. Initialized from EQ-MER, **EQ-Agentic** couples multitask supervised fine-tuning with single-reward, EQ-Judge-guided agentic policy optimization to learn compact agentic UNDERSTAND–PLAN–RESPONSE decisions, while **EQ-Harness** supplies the agentic runtime and deployment support for the model. On nine-dataset MER-UniBench, EQ-MER reaches an **82.22**% macro score with state-of-the-art parameter efficiency; on the agentic EQ-AppBench, EQ-Agentic scores **47.25** under a unified judging protocol applied identically to all baselines. Together, these components form a unified pathway from multimodal emotion recognition toward emotionally intelligent agents; **we will release the full project upon acceptance.**

1 INTRODUCTION

Emotions do not merely accompany human decisions: they shape how people appraise situations and risks, influencing subsequent judgments and choices (Lerner & Keltner, 2001; Lerner et al., 2015). For agents that interact with people, accurate affect perception is therefore an important input for interpreting user states and needs and selecting appropriate behavior. Yet affective communication in face-to-face interaction is inherently multimodal: lexical content, vocal prosody, and visual behavior may complement or conflict with one another (Bagher Zadeh et al., 2018; Poria et al., 2019; Castro et al., 2019). *Multimodal Emotion Recognition* (MER) provides a perceptual foundation for machine EQ by jointly modeling linguistic, acoustic, and visual cues; recent work further extends the task from predefined labels toward open-vocabulary emotion description and fine-grained understanding (Lian et al., 2025b;a). However, recognizing or describing emotions does not entail understanding context, selecting strategies, or generating appropriate responses; MER is therefore a perceptual foundation of machine EQ, but not its endpoint.

This technical limitation should be understood within a broader account of intelligent capability. Psychometric research places reasoning, planning, abstract thinking, and problem solving among important components of general intelligence, while recent foundation models have primarily developed around mathematics and code, tasks that exercise "IQ-like" capabilities (Neisser et al., 1996; Gottfredson, 1997; Carroll, 1993; OpenAI, 2024; DeepSeek-AI, 2025). By contrast, evaluation of emotional intelligence in large language models remains limited and largely recognition-centric (Sabour et al., 2024). We argue that emotional intelligence is an important manifestation of general intelligence in social and affective contexts; progress toward AGI therefore requires not only

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

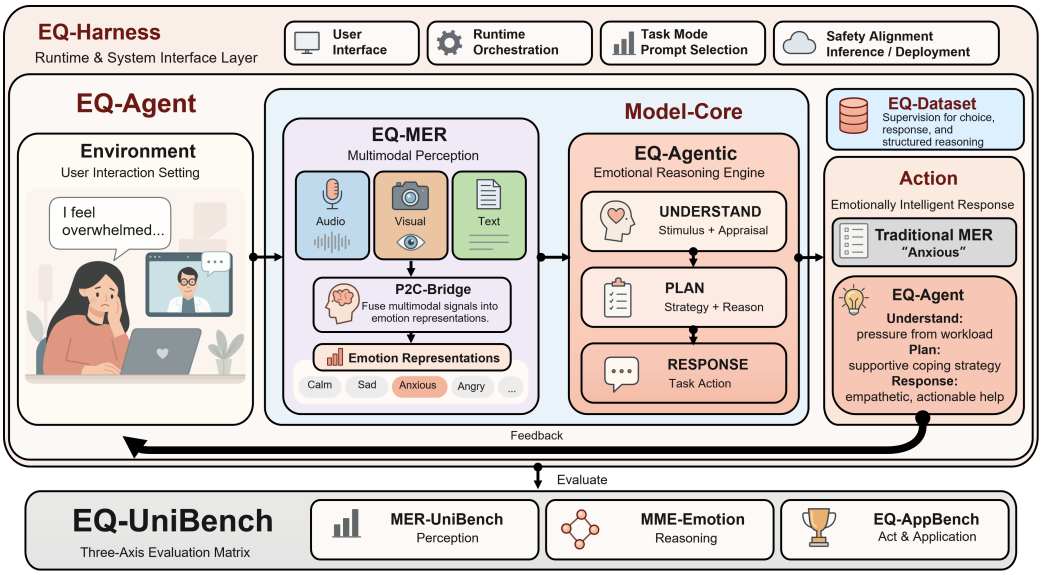

Figure 1: **System overview of EQ-Agent**, comprising EQ-Harness, EQ-MER, EQ-Agentic, and EQ-Dataset, together with EQ-UniBench: MER-UniBench, MME-Emotion, and EQ-AppBench.

cognitive problem solving but also the ability to understand human affect and social context and adapt decisions and responses accordingly. The ability model of Mayer, Caruso, and Salovey characterizes emotional intelligence through perceiving, using, understanding, and managing emotions; human and human–agent studies associate these abilities with interaction quality, trust, and support under specific conditions (Mayer et al., 2016; Lopes et al., 2004; Farh et al., 2012; Keskin et al., 2025; Brave et al., 2005; Fan et al., 2017). We therefore operationalize machine EQ as a complete loop of *emotion perception, contextual understanding, context-aware planning, and appropriate response*, where the response constitutes the task-specific *Task Action*.

More specifically, emotional intelligence extends beyond correctly recognizing emotions: accurate affect perception is only its foundation. Ability models of EQ also emphasize understanding, using, and managing emotions, while research on emotional-support conversations shows that effective support additionally requires selecting communication strategies appropriate to the situation (Mayer et al., 2016; Liu et al., 2021). We argue that a complete EQ loop should build on accurate affect perception and understanding, then use this information to make appropriate choices and responses, allowing emotional cognition to guide context-appropriate behavior. Advances in language agents enable AI to combine reasoning and acting while interacting with an environment, and to adjust subsequent judgments and behavior based on the resulting observations (Yao et al., 2023). This reasoning-and-acting paradigm offers technical inspiration for an EQ loop: grounded in multimodal emotion perception, an agent can further understand emotions and their context, select appropriate strategies, and respond to affective needs in real-world interactions. **To this end, we introduce EQ-Agent, connecting multimodal emotion perception, contextual understanding, strategy planning, and response generation to advance AI from emotion recognition toward emotionally intelligent agents and extend the paradigm from multimodal emotion recognition to multimodal emotional intelligence** (Figure 1).

EQ-Agent comprises EQ-MER, EQ-Agentic, and EQ-Harness, with EQ-Dataset providing multitask supervision for the second learning stage. EQ-MER establishes a lightweight multimodal emotion perception foundation; initialized from EQ-MER, EQ-Agentic learns compact agentic UNDERSTAND–PLAN–RESPONSE (UPR) decisions through Agentic-SFT on EQ-Dataset, where RESPONSE produces the task-specific *Task Action*. Agentic-RL further optimizes emotional decision states from the ESD-CoT portion of EQ-Support with a single EQ-Judge reward. EQ-Harness provides the agentic runtime that unifies task modes, input/output contracts, parsers, and evaluation workflows, allowing different tasks to share consistent training, evaluation, and deployment inter-

faces. Together, these components connect multimodal emotion perception to contextual understanding, strategy planning, and emotionally informed task actions.

Our primary contributions are summarized as follows:

- **EQ-Agent: From Multimodal Emotion Recognition to Emotionally Intelligent Agents.** We introduce a two-stage framework that treats multimodal emotion perception as the foundation of an agentic EQ loop and connects it to contextual understanding, strategy planning, and responses that produce task-specific actions. Through EQ-Agentic learning and the unified EQ-Harness, EQ-Agent extends affective AI from recognizing emotions to making emotionally informed judgments and responses.
- **EQ-Dataset and EQ-UniBench: Beyond Emotion Recognition.** We construct EQ-Dataset, a multitask resource comprising **30,664 data units**, including **1,300 complete dialogues (18,376 supporter turns)**, to supervise social judgment, supportive response, and structured emotional reasoning. We also introduce EQ-UniBench, covering perception with MER-UniBench, reasoning with MME-Emotion, and application with the **6,360-instance EQ-AppBench**.
- **EQ-MER and EQ-Agentic: From Perception SOTA to Emotionally Intelligent Task Actions.** With approximately 0.9B total parameters, EQ-MER achieves an 82.22 macro score across the nine datasets of MER-UniBench, outperforming all existing external baselines, and reaches 91.36 points/B, establishing a new state of the art in both MER performance and score-to-parameter efficiency. Building on EQ-MER, EQ-Agentic extends the same model backbone from emotion perception to UPR-driven task actions and achieves **47.25** on EQ-AppBench under a unified protocol applied identically to all baselines, slightly exceeding the strongest reproduced 8B-scale baseline at roughly one-ninth the parameter count.

2 Related Work

Multimodal Emotion Recognition and Affective Language Models. Multimodal emotion recognition (MER) integrates linguistic, acoustic, and visual cues to provide a perceptual foundation for machine emotion understanding (Poria et al., 2017; Bagher Zadeh et al., 2018). Recent multimodal language models extend MER from fixed-label prediction toward open-vocabulary emotion description and understanding (Lian et al., 2025a); representative affective models, including Emotion-LLaMA, AffectGPT, Emotion-LLaMAv2, and Light-MER, further explore emotion-specific supervision and multimodal fusion (Cheng et al., 2024a; Lian et al., 2025a; Peng et al., 2026; Zheng et al., 2026). Classical fusion and bridging methods cover tensor or low-rank fusion, cross-modal attention, and lightweight language-model bridging, with related approaches also applied to emotion modeling (Zadeh et al., 2017; Liu et al., 2018; Tsai et al., 2019; Li et al., 2023; Lian et al., 2025a; Peng et al., 2026). Existing methods remain primarily perception- or description-oriented and do not directly cover contextual understanding, strategy selection, or response generation; for lightweight models, preserving modality-specific cues, allocating modality contributions according to the input, and adapting fused representations to a limited-capacity language model remain coupled challenges. EQ-MER provides the perceptual foundation, while our proposed P2C Bridge architecture performs source-preserving interaction, dynamic modality routing, and capacity matching; EQ-Agentic then learns contextual understanding and response decisions on top of these components.

Emotional Intelligence: Foundations, Datasets, and Evaluation. Emotional intelligence involves not only recognizing emotions but also understanding and using them, motivating learning and evaluation that cover situational judgment, support strategies, and actual responses. The ability model of Mayer, Caruso, and Salovey treats emotional intelligence as a mental ability, providing a theoretical foundation for research beyond emotion recognition (Mayer et al., 2016). Existing datasets and benchmarks cover complementary capabilities: SocialIQA targets social-situation judgment, CounselChat and ESConv provide supportive-response supervision, ESD-CoT provides structured emotional reasoning, and EmoBench and CounselBench evaluate emotional understanding, supportive-response quality, and adversarial question answering (Sap et al., 2019; Bertagnolli, 2020; Liu et al., 2021; Zhang et al., 2024; Sabour et al., 2024; Li et al., 2025b). Although these resources differ in task objectives, sample granularity, and output formats, together they cover key capabilities from emotional understanding to strategy selection and response generation. Following ESD-CoT and ESConv, we organize these resources into unified Understand–Plan–Response supervision in EQ-Dataset, while EQ-UniBench evaluates their application.

3 METHODOLOGY

3.1 ARCHITECTURAL OVERVIEW: THE TRIPARTITE SYSTEM

EQ-Agent is an emotionally intelligent agent system that connects multimodal perception of emotional environments, contextual understanding, and response decisions through sequential capability learning and unified runtime support, translating emotional information into contextually appropriate choices and responses. As illustrated in Figure 1, the system comprises EQ-MER, EQ-Agentic, and EQ-Harness, with EQ-Dataset providing multitask supervision for the second learning stage. EQ-MER generates emotion answers from multimodal environmental evidence, providing a perceptual foundation for subsequent decisions.

3.2 EQ-MER: MULTIMODAL EMOTION PERCEPTION

EQ-MER formulates multimodal emotion recognition as open-vocabulary answer generation conditioned on audiovisual evidence and textual context. We design a Perceptual-to-Cognitive (P2C) Bridge that turns features from frozen encoders into inputs for a small language model while retaining paths for the original modality features and adjusting fusion weights per input. Comprising three Prefusion layers and a final linear projection, the bridge operates in two functional stages (Figure 2): alignment and fusion apply source-preserving cross-modal interaction and input-dependent modality routing to produce $H = \mathcal{F}_{\text{align-fuse}}(G_v, G_a)$; capacity adaptation and projection then yield LLM-compatible tokens $Z = \text{Proj}(\mathcal{M}_{\text{matcher}}(H))$ via the residual matcher. Visual representations G_v and acoustic representations G_a enter the bridge, while text tokens T_t bypass it and join the LoRA-adapted Qwen3-0.6B alongside Z.

Alignment and fusion. Frozen-encoder representations G_v and G_a are projected and normalized into equal-length sequences $V, A \in \mathbb{R}^{N\times d}$; our implementation uses $N = 8$ and $d = 1024$. Let $X_v = V$ and $X_a = A$ denote the pre-interaction sequences, and let $P_m = \text{LN}_m^{\text{att}}(X_m)$ be the attention-normalized input of modality $m \in \{v, a\}$, with $\bar{m}$ the other modality.

Both directions have the shared form

$$\begin{aligned} C_m &= \text{ResFFN}_m(X_m + U_{m\leftarrow\bar{m}}), \\ U_{m\leftarrow\bar{m}} &= \text{MHA}(P_m, P_{\bar{m}}, P_{\bar{m}}), \quad m \in \{v, a\}. \end{aligned} \tag{1}$$

When m reads $\bar{m}$, P_m is the query and $P_{\bar{m}}$ supplies the key and value. ResFFN is a residual two-layer GELU feed-forward network, and both directions use the same pair X_v, X_a. A source-preserving gate, with modality-specific W_m, b_m, then interpolates between X_m and C_m at each token and channel:

$$g_m = \sigma(W_m[X_m; C_m] + b_m), \qquad \widetilde{X}_m = X_m + g_m \odot (C_m - X_m), \quad m \in \{v, a\}. \tag{2}$$

As g_m approaches zero, $\widetilde{X}_m$ stays near X_m; as it approaches one, $\widetilde{X}_m$ approaches C_m. The router receives $\widetilde{V} = \widetilde{X}_v$ and $\widetilde{A} = \widetilde{X}_a$.

Within the same alignment-and-fusion stage, the modality router applies masked mean pooling to the gated sequences $\widetilde{V}$ and $\widetilde{A}$ to obtain visual and acoustic summaries $\bar{v}$ and $\bar{a}$ for the current sample. Let $M_{m,t} \in \{0, 1\}$ indicate whether modality m is valid at token position t, and let $a_m = \mathbf{1}[\sum_t M_{m,t} > 0]$ indicate its availability in the sample. From the summary vector $q = [\bar{v}; \bar{a}; |\bar{v}-\bar{a}|; \bar{v}\odot\bar{a}]$, the router predicts sample-level weights and renormalizes them over valid modalities at each token position:

$$\begin{aligned} (w_v, w_a) &= \text{MaskedSoftmax}(f_{\text{router}}(q)/\tau;\, a_v, a_a), \\ \alpha_{m,t} &= \frac{w_m M_{m,t}}{\max\Big(\sum_{j\in\{v,a\}} w_j M_{j,t},\, \varepsilon\Big)}, \quad H_t = \alpha_{v,t}\widetilde{V}_t + \alpha_{a,t}\widetilde{A}_t. \end{aligned} \tag{3}$$

f_{router} maps q through normalization and a feed-forward network to modality scores, and the temperature $\tau > 0$ controls their concentration. Masked softmax enforces $w_v + w_a = 1$ when both modalities are available and assigns weight one to the only available modality. $H = (H_t)_{t=1}^N$ retains the token count shown in Figure 2 and enters the capacity adaptation layer.

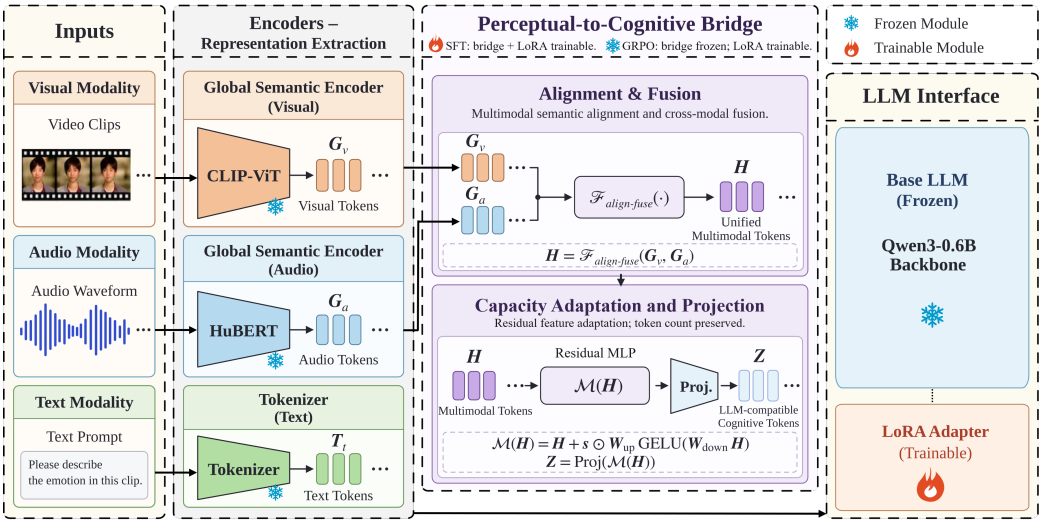

Figure 2: **EQ-MER architecture and the P2C Bridge.** P2C transforms audiovisual features G_v, G_a from frozen encoders into the fused representation H and then projects it into LLM inputs Z; text tokens T_t bypass the bridge. SFT updates P2C and LoRA, whereas answer-level GRPO updates only LoRA.

Capacity adaptation and projection. After audiovisual fusion, the third Prefusion layer further transforms the fused representation H through a residual bottleneck network, followed by a linear projection into the language model's input embedding space, i.e., $Z = \mathrm{Proj}(H + s \odot W_{\mathrm{up}}\,\mathrm{GELU}(W_{\mathrm{down}}H))$. Here, W_{down} and W_{up} map each token's features into a bottleneck space and back to the original dimension, respectively, with GELU providing a nonlinear transformation; biases are omitted for brevity. The learnable scaling vector s shares a coefficient within each channel group and is reused across all token positions to modulate the residual branch. The additive path for H retains the input representation and adds a learned feature update. The resulting multimodal tokens Z and text tokens T_t jointly condition emotion-answer generation.

Supervised fine-tuning and answer-level reinforcement learning. EQ-MER first undergoes SFT on MER-Caption+ using answer-token negative log-likelihood, training the P2C Bridge and LoRA in stages while freezing the audiovisual encoders and base language model. We then initialize answer-level GRPO from the SFT model using 1,000 MER2025-OV training examples, freeze the complete bridge, and update only LoRA, sampling $K = 4$ candidate answers per input. The reward is

$$R(\hat{y}, y) = 2\sqrt{P(\hat{y}, y)}\;\mathrm{EW}_{\mathrm{legacy}}(\hat{y}, y) + R_{\mathrm{fmt}}(\hat{y}), \qquad P(\hat{y}, y) = \begin{cases} 1, & n_{\mathrm{pred}} \le n_{\mathrm{ref}}, \\ \dfrac{n_{\mathrm{ref}}}{n_{\mathrm{pred}}}, & n_{\mathrm{pred}} > n_{\mathrm{ref}},\ n_{\mathrm{ref}} \le 1, \\ \dfrac{\log n_{\mathrm{ref}}}{\log n_{\mathrm{pred}}}, & n_{\mathrm{pred}} > n_{\mathrm{ref}},\ n_{\mathrm{ref}} \ge 2. \end{cases} \tag{4}$$

Here, $\hat{y}, y$ are the generated and reference answers, with $n_{\mathrm{pred}}, n_{\mathrm{ref}}$ extracted emotion labels, respectively. $\mathrm{EW}_{\mathrm{legacy}}$ averages F1 over level-one mappings from five emotion wheels, and R_{fmt} is one for the prescribed answer format and zero otherwise. We introduce P to discourage listing excessive emotion labels to inflate hit rate, reducing the risk of this form of reward hacking; the square root softens the penalty while retaining its suppression of excess labels. GRPO uses group-standardized rewards as relative advantages with KL regularization, referencing the same frozen backbone with LoRA disabled. The SFT objective, advantage calculation, and optimization settings are detailed in Appendix C.

3.3 EQ-DATASET: RESOURCES FOR EQ LEARNING

EQ-Dataset provides complementary supervision for contextual understanding and response: social-situation judgment from SocialIQA (Sap et al., 2019), supportive replies from CounselChat (Bertagnolli, 2020), and emotional understanding, strategy planning, and response generation from EQ-Support. EQ-Support combines ESD-CoT (Zhang et al., 2024) with augmented ESConv (Liu et al., 2021). Construction combines source-specific curation, structured emotional annotation, and human sample review, preserving all original ESConv conversations and human supervision. Table 1 summarizes source sizes before and after curation; detailed construction, splits, and quality control are provided in Appendix A.

Table 1: EQ-Dataset before and after curation.

Source	Raw	Retained
SocialIQA	37,588	26,156
CounselChat	2,775	1,616
ESD-CoT	1,708	1,592
ESConv	1,300	1,300
Total	**43,371**	**30,664**

3.4 EQ-AGENTIC: LEARNING EMOTIONAL UNDERSTANDING AND RESPONSE

EQ-Agentic learns from two complementary types of supervision in EQ-Dataset—verifiable judgment over closed candidate sets and complete responses in open-ended supportive situations—under a unified UNDERSTAND–PLAN–RESPONSE structure: UNDERSTAND summarizes emotional triggers and the individual's appraisal; PLAN selects a support strategy and explains its rationale; RESPONSE realizes the strategy as the task-specific output—a verifiable final answer in judgment tasks and a context-appropriate reply in response tasks. Training and evaluation thereby share the same capability axes, Judgment and Response (Section 4).

Agentic-SFT. Initialized from EQ-MER, Agentic-SFT trains a single set of LoRA adapters on the language model with multitask supervision from EQ-Dataset, while keeping the audiovisual encoders, P2C Bridge, and base language model frozen in this and all subsequent stages. The training view combines three pools selected for complementary strengths: although all carry UPR annotations, 23,213 SocialIQA records emphasize verifiable judgment with a single correct option; 1,297 CounselChat records provide human-authored ground-truth replies, stressing accurate final answers; and 1,112 ESD-CoT records supply high-quality reasoning chains, stressing the UPR reasoning process; complete ESConv conversations are held out for evaluation. The objective is the token-mean negative log-likelihood of each pool's assistant output tokens—supervising the UPR tags, segment bodies, terminator, and trailing newline while masking every other position—with the three pool losses equally weighted so SocialIQA's record count does not dominate the gradient:

$$\mathcal{L}_{\mathrm{SFT}} = \frac{1}{3}(\mathcal{L}_{\mathrm{SIQA}} + \mathcal{L}_{\mathrm{CC}} + \mathcal{L}_{\mathrm{ESDCoT}})\,. \tag{5}$$

Here $\mathcal{L}_{\mathrm{SIQA}}, \mathcal{L}_{\mathrm{CC}}, \mathcal{L}_{\mathrm{ESDCoT}}$ are the token-mean NLLs of the SocialIQA, CounselChat, and ESD-CoT pools; the expanded per-pool form is given in Appendix D.

Single-reward EQ-Judge Agentic-RL. Agentic-RL performs GRPO policy optimization on top of Agentic-SFT, updating the same LoRA adapters. For each emotional-decision state x drawn from the ESD-CoT split of EQ-Support (1,112 states, once per epoch for two epochs), the old policy samples $K{=}4$ complete UPR trajectories $o_1, \ldots, o_K$, and a single EQ-Judge reward scores each whole trajectory rather than the final reply alone:

$$r_k = g_{\mathrm{EQ}}(x, o_k) \in [0, 10], \qquad A_k = \frac{r_k - \bar{r}}{\mathrm{Std}(r_1, \ldots, r_K) + \varepsilon}. \tag{6}$$

where g_{EQ} is a frozen Qwen-based judge returning a single `Overall EQ Score`, $\bar{r}$/Std are the group mean/standard deviation, ε a stabilizer, and unusable trajectories contribute $r_k{=}0$. The policy is then updated with the standard GRPO clipped surrogate (Shao et al., 2024) under token-level aggregation: token-level likelihood ratios $\rho_{k,t} = \pi_\theta(o_{k,t} \mid x, o_{k,<t}) / \pi_{\mathrm{old}}(o_{k,t} \mid x, o_{k,<t})$ are clipped at $\epsilon{=}0.2$ and weighted by A_k, together with a $\beta{=}0.02$ KL penalty toward the frozen SFT reference (full objective in Appendix D). The trajectory-level reward makes planning itself optimizable; equal-reward groups contribute no gradient.

3.5 EQ-HARNESS: UNIFIED INFRASTRUCTURE, RUNTIME, AND INTERFACE

EQ-Harness is EQ-Agent's unified infrastructure, runtime, and interface layer, connecting data, models, and applications through consistent task protocols for training, evaluation, and deployment. The caller specifies one of two task modes, *judgment* or *response*; all tasks share the UNDERSTAND–PLAN–RESPONSE output contract while retaining their own input/output requirements. EQ-Harness further supplies prompt-based safety guidelines, batched execution, and EQ Trace logging of outputs, parse status, and protocol versions; a streaming deployment and interactive-interface framework is validated on a Jetson Orin Nano Super 8 GB with the same Qwen3-0.6B backbone and Large audiovisual encoders, and will be open-sourced upon acceptance. Appendix E details protocols, prompts, safety alignment, and deployment.

4 EXPERIMENTS AND RESULTS

EQ-UniBench: Three-axis evaluation of emotional intelligence. The design of EQ-UniBench draws on the ability model of emotional intelligence (Mayer et al., 2016), which treats emotional intelligence as a mental ability to reason about emotions—perceiving, using, understanding, and managing them—assessed through performance on concrete problems. The theoretical basis of each axis is given in Table 11. EQ-UniBench reports three axes separately, without a cross-axis aggregate: **perception on MER-UniBench, reasoning on MME-Emotion, and application on EQ-AppBench.**

EQ-AppBench is designed in this paper specifically to evaluate EQ-related agent capabilities. Its four subtests are grouped into two modes by the structure of their answer space. *Judgment* covers closed-set decisions with verifiable answers and is scored deterministically without an LLM judge; *Response* covers open-ended supportive situations, in which a judge scores the candidate's complete generated output rather than an extracted reply alone, because the quality of a response cannot be separated from the understanding and planning that produced it. The two mode scores are equally averaged.

4.1 EVALUATION PROTOCOL AND BASELINES

We evaluate EQ-Agent along the perception, reasoning, and application axes of EQ-UniBench and report each axis separately. Baselines use public implementations and author-released checkpoints, with test splits, I/O requirements, generation configs, and scoring protocols specified per axis; we distinguish reproduced results from literature-reported scores and disclose deviations from official settings. For MER baselines under both official and unified greedy decoding, we report the complete higher-scoring run without per-dataset maxima. Detailed protocols, baseline configurations, and complete results appear in Appendix B; task interfaces and prompts in Appendix E.

4.2 RESULTS AND ANALYSIS

Perception: MER-UniBench. As shown in Table 2, EQ-MER achieves a MER-UniBench score of **82.22** with only **0.9B parameters**, exceeding the similarly sized Light-MER by **7.18 points** and the strongest reproduced baseline Emotion-LLaMAv2 by **6.24 points**. Unlike Light-MER, which uses teacher–student distillation (Zheng et al., 2026), EQ-MER achieves this performance through its architecture and training strategy. Its parameter efficiency, **91.36 points/B**, is the table's highest. The margin is consistent: EQ-MER also leads Hits@1 on MER24 (96.02) and weighted-F1 on SIMS (88.18) and SIMSv2 (85.39).

Reasoning and application. On the 6,500-question MME-Emotion suite (unresolved cases score zero), EQ-MER's recognition score (35.70) matches the emotion-finetuned baselines while its reasoning-step score is near zero (0.62)—strong perception alone does not yield verifiable emotional reasoning, motivating EQ-Agentic's UPR training; it is not evaluated on this axis (– in Table 2). On EQ-AppBench, Judgment deterministically parses SocialIQA and EmoBench; a judge scores each candidate's complete generated output on CounselBench-Eval and EQ-Support-Test (one 'Overall EQ Score' in $[0, 10]$), since baselines' UPR compliance is unstable; empty, failed, or over-length items contribute zero on fixed denominators (Appendix B).

Table 2: **EQ-UniBench.** Lightly shaded rows are our models; bold marks the highest score in each column. **(A)** MER-UniBench on nine datasets. S is the overall score; $E_P = S/P_B$ is parameter efficiency, with P_B in billions. **(B)** MME-Emotion and EQ-AppBench. Judg. averages unrounded SIQA and EmoB, Resp. averages unrounded CB-Eval and ES-T., and Mean averages those two unrounded scores; each is shown to two decimals. A "–" marks a score that is not evaluated on that axis.

(A)

MER-UniBench	Basic emotion (HIT)				Sentiment (weighted F1)				OV-MERD+	Overall	
Model	MER23	MER24	MELD	IEMO	MOSI	MOSEI	SIMS	SIMSv2	EW-F1	S	E_P
General-purpose MLLMs											
VideoLLaMA2.1-AV	43.58	49.87	29.41	25.32	54.52	50.34	78.01	75.27	11.35	46.41	6.11
VITA-1.5	48.66	65.98	38.42	38.27	49.66	26.94	61.07	54.09	46.68	47.75	5.97
HumanOmni-7B	65.30	77.42	24.12	47.14	50.18	28.87	75.52	71.61	22.75	51.43	5.36
Qwen2.5-Omni-7B	55.18	44.62	37.09	35.01	63.90	37.53	79.73	79.92	47.49	53.38	4.99
video-SALMONN 2	41.09	53.86	15.16	25.09	77.94	58.49	81.91	82.80	48.52	53.87	6.65
Baichuan-Omni-1.5	50.16	51.66	29.36	44.22	73.83	70.23	82.55	83.32	43.84	58.80	5.35
MiniCPM-o-4.5	60.67	68.18	33.96	50.18	75.45	62.41	83.62	81.76	46.51	62.53	6.72
Emotion-finetuned MLLMs											
Emotion-LLaMA	67.26	76.49	31.21	59.16	76.15	78.79	77.32	76.57	**62.79**	67.30	8.21
AffectGPT	69.48	65.00	54.84	49.82	83.37	76.70	80.47	79.76	55.67	68.35	8.33
Nano-EmoX	76.21	78.13	53.98	57.22	75.06	80.33	86.41	84.54	61.02	72.55	29.02
Light-MER	76.53	85.20	61.20	67.86	75.91	76.06	86.90	84.53	61.15	75.04	83.38
Emotion-LLaMAv2	72.20	81.41	43.77	**83.87**	**86.87**	**84.86**	84.54	84.12	62.18	75.98	8.17
Our MLLM											
EQ-MER	**93.25**	**96.02**	**80.61**	80.70	75.84	78.95	**88.18**	**85.39**	61.08	**82.22**	**91.36**

(B)

	MME-Emotion				EQ-AppBench						
Model	**Rec**	**Rea**	**CoT**	**Mean**	**SIQA**	**EmoB**	**Judg.**	**CB-Eval**	**ES-T.**	**Resp.**	**Mean**
Nano-EmoX	20.54	28.53	24.53	24.53	5.35	0.25	2.80	8.80	8.21	8.51	5.65
Light-MER	23.93	8.10	16.02	16.02	12.90	29.38	21.14	21.10	20.37	20.73	20.94
Emotion-LLaMA 2	**40.19**	**44.67**	**42.43**	**42.43**	34.35	21.38	27.86	18.80	9.69	14.25	21.05
Emotion-LLaMA v1	28.07	8.52	18.29	18.29	36.74	20.88	28.81	28.90	22.32	25.61	27.21
EQ-MER	35.70	0.62	18.16	18.16	47.03	27.38	37.20	20.20	34.89	27.54	32.37
AffectGPT	26.22	24.01	25.11	25.11	**68.08**	**43.63**	**55.85**	19.40	**57.25**	38.32	47.09
EQ-Agentic	–	–	–	–	62.86	29.75	46.30	**45.40**	50.97	**48.19**	**47.25**

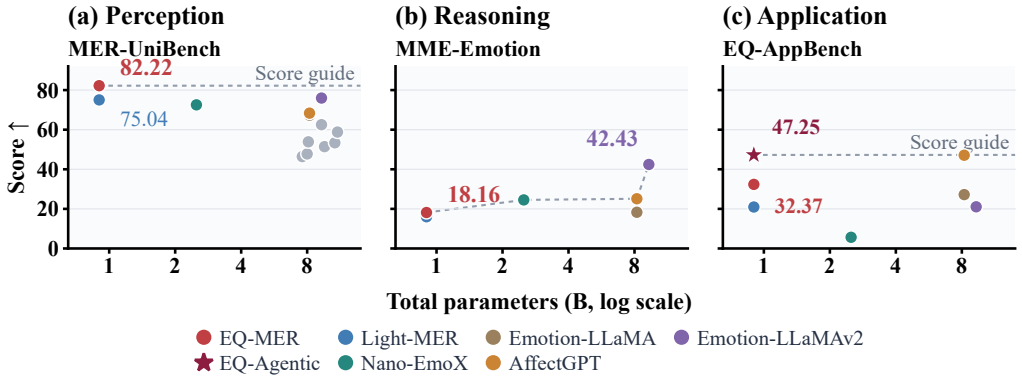

Figure 3: Parameter–score comparison on EQ-UniBench: (a) perception, (b) reasoning, and (c) application (the mean of Judgment and Response). Star: EQ-Agentic; gray: general-purpose MLLMs; dashed: Pareto frontier (a lone frontier point extends into a score guide). EQ-Agentic is omitted from (b); parameter counts use a logarithmic axis.

Table 3: **Deployment efficiency** on identical H100 hardware. Trainable lists per-stage update counts; †marks Light-MER's config-derived stage-1 student count (its Qwen3-8B teacher excluded). S is the MER-UniBench overall score; E_P uses rounded parameter counts. Latency: per-sample end-to-end; memory: batch-8 peak; Nano-EmoX runs fp32.

Model	Params (M)	Trainable (M)	VRAM (MiB)	Latency (s)	FLOPs (G)	S	E_P
EQ-MER	888.0	44.14 + 10.09	6,156	1.801	659.0	**82.22**	**91.36**
Light-MER	844.8	15.94† + 14.89	5,970	1.678	649.8	75.04	83.38
Nano-EmoX	2,452.0	77.70	13,338	1.877	2,429.0	72.55	29.02
AffectGPT	8,415.8	57.17	22,714	3.228	3,686.1	68.35	8.33

Table 4: P2C ablations on MER-UniBench.

P2C Variant	Macro	Δ
EQ-MER	**82.22**	0.00
w/o Alignment & Fusion		
Entire	76.47	−5.75
Cross-Attention	80.48	−1.74
Source-Preserving Gate	80.14	−2.08
Uniform Router	79.76	−2.46
w/o Capacity Adaptation & Projector		
Capacity Matcher	79.95	−2.27
Projector-Only	78.45	−3.77

Table 5: Training stages. S is the MER-UniBench overall score; the remaining columns are EQ-AppBench scores under the unified profile.

Stage	S	B	SIQA	EmoB	CB	ES
EQ-MER SFT	75.44	–	–	–	–	–
EQ-MER GRPO	**82.22**	32.37	47.03	27.38	20.20	34.89
EQ-Agentic SFT	–	39.51	62.59	31.25	27.00	37.18
EQ-Agentic RL	–	**47.25**	62.86	29.75	45.40	50.97

Table 6: Multi-Seed Experiments.

Model	Greedy	α	β	γ	Mean ± SD
EQ-MER	82.22	79.74	80.14	79.89	80.50 ± 1.16
EQ-Agentic	47.25	46.08	45.40	46.23	46.24 ± 0.76

Ablations and efficiency. Table 4 ablates six P2C Bridge paths under the full SFT–GRPO schedule (SFT trains bridge and LoRA; answer-level GRPO updates only LoRA): dropping the entire alignment-and-fusion stage costs 5.75 points, while every single-path ablation stays within 3.8. Table 5 separates training stages: answer-level GRPO raises MER-UniBench 75.44→**82.22**; on EQ-AppBench, Agentic-SFT reaches 39.51 and Agentic-RL reaches **47.25**, with the larger gain on the Response axis. Agentic-SFT establishes the decision contract, whereas Agentic-RL turns plans into substantially better replies. Per Table 3, the gains are cheap: EQ-MER tracks Light-MER within 5% on every efficiency axis while scoring 7.18 higher, and needs substantially fewer parameters and FLOPs than the 8B-scale baseline. Table 6 further probes decoding robustness: all reported scores use greedy decoding, while $\alpha/\beta/\gamma$ are three randomly drawn sampling seeds (temperature 0.7); sampled runs stay within 2.5 points of greedy (SD $\leq$1.16), confirming that our results are not a lucky decoding draw. Under this unified protocol, EQ-Agentic's **47.25** edges past AffectGPT's 47.09 through the Response axis (48.19 vs. 38.32), trading Judgment (46.30 vs. 55.85) for stronger situated replies.

5 Conclusion and Limitations

We present **EQ-Agent**, an end-to-end framework advancing multimodal emotional intelligence from passive classification to active, agentic communication. **EQ-MER**, a sub-1B perception foundation built on the Perception-to-Cognition Bridge, sets a new state of the art at **82.22** on MER-UniBench; initialized from it, **EQ-Agentic** reaches **47.25** on EQ-AppBench, edging past the strongest reproduced 8B-scale baseline AffectGPT (47.09). **Limitations.** This margin is narrow—within run-to-run variance—but is achieved at roughly $9\times$ the parameter efficiency; resource constraints limited hyperparameter tuning across training stages, so although baselines ran under their official best configurations, our results likely understate the design's ceiling while still attaining parameter-efficiency SOTA; and the P2C Bridge's cross-attention adds computation, yet total FLOPs remain $5.6\times$ below the 8B-scale baseline. Lightweight models, fortified by structured agentic deliberation, can thus approach much larger models in emotional intelligence—a step toward empathetic AGI.

AI USE STATEMENT

We used generative AI in the following capacities. **Research and design:** discussing and refining research hypotheses, conceptual framing, experimental methodology, literature search and summarization, and brainstorming. **Implementation and data:** assisting with training and evaluation code, cleaning, reformatting, and screening data, and augmenting ESConv with structured emotional explanations under human-defined rules (Appendix A). **Writing:** experimental-parameter suggestions, interpreting experimental results, manuscript drafting, translation, language revision, and the preparation of figures, tables, and references. Formal theorem and proof tasks are not applicable to this work.

External language-model judges are used for parts of the evaluation of supportive responses and structured emotional reasoning. The human authors reviewed all AI-assisted research designs, code, data artifacts, citations, and manuscript text; code and metrics were verified through tests, run logs, and recorded evaluation outputs; and citations were checked against their original sources. Human review of the data was sample-based and included batch-level acceptance rather than exhaustive certification of every data unit. The human authors take responsibility for the final text, claims, code, data, figures, and evaluation procedures, including all AI-assisted components.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 specifies the model architecture, staged training procedure, reward definitions, and optimization settings, while Section 4 records the data splits, evaluation metrics, decoding protocols, checkpoint-selection rules, and provenance of the reported results. Appendices A and B.2 further document EQ-Dataset construction, parameter accounting, and differences between original-paper and in-environment evaluations. Every main result is produced by one complete checkpoint; we do not assemble a row by selecting different checkpoints for different datasets. We commit to open-sourcing all project-generated artifacts that upstream licenses permit us to redistribute, including the EQ-Agent implementation, training and evaluation code, configuration files, EQ-Harness, benchmark adapters and parsers, data-construction pipelines, prompts, scoring protocols, experiment records, and redistributable model weights and annotations. For third-party raw data that cannot legally be redistributed, we will provide source identifiers, processing scripts, split manifests, and provenance records so that researchers with lawful access can reconstruct the corresponding resources. **We will provide anonymous supplementary code during review and release the complete public repository upon acceptance to facilitate reproducible research on Agentic EQ, emotionally intelligent agents, and multimodal affective reasoning.**

ETHICS STATEMENT

This work uses public research datasets under their documented licenses and research-use restrictions. Resources with non-commercial or no-derivatives restrictions are used only for academic research, and we do not redistribute modified versions of restricted data. Supportive-dialogue and help-seeking data may contain sensitive content, cultural biases, stereotypes, noisy annotations, and subjective emotion judgments. We do not attempt to identify individuals represented in the data, and human sample review serves as quality control rather than clinical validation. EQ-Agent is intended for research on emotional understanding and supportive communication; it is not intended for clinical diagnosis, psychotherapy, crisis intervention, or emergency services. The model may still misidentify emotions or generate inappropriate, biased, or unsafe advice, so any high-stakes deployment would require additional safeguards and qualified human oversight. External language-model judges may also introduce model-specific preferences and scoring biases; their results should therefore be interpreted together with the stated prompts, scoring protocols, and human checks.

REFERENCES

Artificial Analysis. Gemini 3.8 Flash (high): Intelligence, performance & price analysis, September 2026a. URL https://artificialanalysis.ai/models/gemini-3-8-flash.

Artificial Analysis. Benchmarking GPT-6 Astra, September 2026b. URL https://artificialanalysis.ai/articles/benchmarking-gpt-6-astra.

Artificial Analysis. Qwen3.8 27B: Release intelligence, performance & price, 2026c. URL https://artificialanalysis.ai/models/releases/qwen3-8-27b.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Nicolas Bertagnolli. Counsel Chat: Bootstrapping high-quality therapy data. Author's dataset and project repository, 2020. URL https://github.com/nbertagnolli/counsel-chat.

Scott Brave, Clifford Nass, and Kevin Hutchinson. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62(2):161–178, 2005. doi: 10.1016/j.ijhcs.2004.11.002.

John B. Carroll. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press, Cambridge, 1993. doi: 10.1017/CBO9780511571312.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an Obviously perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1455. URL https://aclanthology.org/P19-1455/.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c7f43ada17acc234f568dc66da527418-Abstract-Conference.html.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024b. URL https://arxiv.org/abs/2406.07476.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Lisa Fan, Matthias Scheutz, Monika Lohani, Marissa McCoy, and Charlene Stokes. Do we need emotionally intelligent artificial agents? first results of human perceptions of emotional intelligence in humans compared to robots. In *Intelligent Virtual Agents*, volume 10498 of *Lecture Notes in Computer Science*, pp. 129–141. Springer, 2017. doi: 10.1007/978-3-319-67401-8_15.

Crystal I. C. Chien Farh, Myeong-Gu Seo, and Paul E. Tesluk. Emotional intelligence, teamwork effectiveness, and job performance: The moderating role of job context. *Journal of Applied Psychology*, 97(4):890–900, 2012. doi: 10.1037/a0027377.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. URL https://arxiv.org/abs/2501.01957.

Linda S. Gottfredson. Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24(1):13–23, 1997. doi: 10.1016/S0160-2896(97)90011-8.

Jiahao Huang, Fengyan Lin, Xuechao Yang, Chen Feng, Kexin Zhu, Xu Yang, and Zhide Chen. Nano-EmoX: Unifying multimodal emotional intelligence from perception to empathy. *arXiv preprint arXiv:2603.02123*, 2026. URL https://arxiv.org/abs/2603.02123.

Mehmet Onur Keskin, Umut Çakan, and Reyhan Aydoğan. An adaptive emotion-aware strategy for human-agent negotiation: Insights from real-world human-robot experiments. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, 2025. doi: 10.1145/3717511.3747087.

Jennifer S. Lerner and Dacher Keltner. Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81(1):146–159, 2001. doi: 10.1037/0022-3514.81.1.146.

Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. Emotion and decision making. *Annual Review of Psychology*, 66:799–823, 2015. doi: 10.1146/annurev-psych-010213-115043.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Yadong Li, Jun Liu, Tao Zhang, et al. Baichuan-Omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025a. URL https://arxiv.org/abs/2501.15368.

Yahan Li, Jifan Yao, John Bosco S. Bunyi, Adam C. Frank, Angel Hsing-Chi Hwang, and Ruishan Liu. CounselBench: A large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. *arXiv preprint arXiv:2506.08584*, 2025b. URL https://arxiv.org/abs/2506.08584.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014. PMLR, 2025a. URL https://proceedings.mlr.press/v267/lian25a.html.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37015–37050. PMLR, 2025b. URL https://proceedings.mlr.press/v267/lian25b.html.

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, and Xiaobai Li. AffectGPT-R1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*, 2025c. URL https://arxiv.org/abs/2508.01318v2.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *ACL*, pp. 3469–3483, 2021.

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. doi: 10.18653/v1/P18-1209. URL https://aclanthology.org/P18-1209/.

Paulo N. Lopes, Marc A. Brackett, John B. Nezlek, Astrid Schütz, Ina Sellin, and Peter Salovey. Emotional intelligence and social interaction. *Personality and Social Psychology Bulletin*, 30(8): 1018–1034, 2004. doi: 10.1177/0146167204264762.

Carolyn MacCann and Richard D. Roberts. New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, 8(4):540–551, 2008. doi: 10.1037/a0012746.

John D. Mayer, David R. Caruso, and Peter Salovey. The ability model of emotional intelligence: Principles and updates. *Emotion Review*, 8(4):290–300, 2016. doi: 10.1177/1754073916639667.

Ulric Neisser, Gwyneth Boodoo, Thomas J. Bouchard, A. Wade Boykin, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, John C. Loehlin, Robert Perloff, Robert J. Sternberg, and Susana Urbina. Intelligence: Knowns and unknowns. *American Psychologist*, 51(2):77–101, 1996. doi: 10.1037/0003-066X.51.2.77.

Yuki Nozaki and James J. Gross. Bridging supportive communication and interpersonal emotion regulation: An integrative review. *Journal of Social and Personal Relationships*, 42(8):2231–2262, 2025. doi: 10.1177/02654075251335816.

OpenAI. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. URL https://arxiv.org/abs/2412.16720.

OpenBMB. MiniCPM-o 4.5: Official model release and documentation. Official project repository, 2026. URL https://github.com/OpenBMB/MiniCPM-V.

Xiaojiang Peng, Jingyi Chen, Zebang Cheng, Bao Peng, Fengyi Wu, Yifei Dong, Shuyuan Tu, Qiyu Hu, Huiting Huang, Yuxiang Lin, Jun-Yan He, Kai Wang, Zheng Lian, and Zhi-Qi Cheng. Emotion-LLaMAv2 and MMEVerse: A new framework and benchmark for multimodal emotion understanding. *arXiv preprint arXiv:2601.16449*, 2026. URL https://arxiv.org/abs/2601.16449.

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.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5986–6004, 2024. doi: 10.18653/v1/2024.acl-long.326. URL https://aclanthology.org/2024.acl-long.326/.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Klaus R. Scherer. Emotions are emergent processes: They require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3459–3474, 2009. doi: 10.1098/rstb.2009.0141.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Zhang, Y Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. URL https://arxiv.org/abs/2506.15220.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pp. 6558–6569, 2019.

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. URL https://arxiv.org/abs/2503.20215.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pp. 1103–1114, 2017.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, YiFan 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. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/50d277e84b2bcbaadcd84548a87e8cc4-Abstract-Conference.html.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13395–13412, 2024. doi: 10.18653/v1/2024.acl-long.723. URL https://aclanthology.org/2024.acl-long.723/.

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. *arXiv preprint arXiv:2501.15111*, 2025. URL https://arxiv.org/abs/2501.15111.

Kaiwen Zheng, Junchen Fu, Wenhao Deng, Hu Han, Joemon M. Jose, and Xuri Ge. Do we really need multimodal emotion language models larger than 1B parameters? *arXiv preprint arXiv:2607.12787*, 2026. URL https://arxiv.org/abs/2607.12787.

A EQ-DATASET CONSTRUCTION

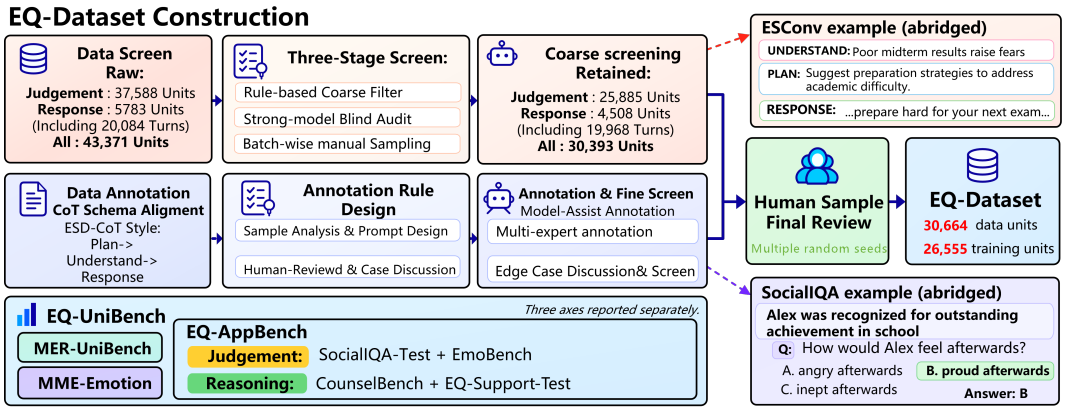

Figure 4: **Human-guided construction of EQ-Dataset.** Source-specific processing distinguishes the curation of existing annotations from the augmentation of ESConv with structured emotional explanations. Human sample review supports quality control. Counts cover all splits and use source-specific units; ESConv conversations and original supervision are fully retained. The lower-left panel summarizes the complementary evaluation axes of EQ-UniBench. Blind second review and human verification are sampled rather than exhaustive.

Three-stage data screening. Before annotation, we apply a three-stage screening procedure to all four sources. The first stage is a rule-based coarse filter for obvious surface-quality issues, including clear spelling errors, meaningless repetitions, and duplicate records. The second stage is a strong-model blind audit governed by manually specified review rules. We provide GPT-6 Astra, the flagship model available during data construction (Artificial Analysis, 2026b), with raw inputs without reference answers and require it to solve each task independently; for SocialIQA, for example, the model first selects an answer, after which its prediction is compared with the original annotation. Records showing clear and irreconcilable logical conflicts with the original annotation enter a candidate issue set, which human reviewers inspect individually to decide whether to retain, correct, or exclude them. The third stage is batch-wise manual sampling with multiple random seeds: we partition the data into batches according to fixed rules and draw samples from each batch using three independent random seeds for manual inspection. If a batch reveals a systematic issue, we return to the second stage for another strong-model blind audit and human disposition. All human review in this screening process is conducted by three qualified data annotators and reviewers under a shared training and calibration standard; strong-model judgments are used to identify risks and are not treated as final ground truth.

UPR annotation and validation. After screening, we annotate the retained records from all four sources using a source-adapted schema based on the structured emotional reasoning and supportive-dialogue paradigms of ESD-CoT and ESConv. We first use GPT-6 Astra to analyze sampled records from each source, design source-specific prompts, and discuss edge cases under a shared annotation rubric. We then use Gemini 3.8 Flash (high) to annotate independently sampled pilot records under the candidate rules; Artificial Analysis Intelligence Index v4.3.2 reports a score of 41 and eighth place in its displayed frontier comparison for this model (Artificial Analysis, 2026a). Three qualified annotators and reviewers inspect each sampled annotation, checking the contextual consistency, logical coherence, strategy appropriateness, and response quality of the UNDERSTAND, PLAN, and RESPONSE fields. We iteratively revise the rubric and prompts using human review findings and GPT-6 Astra's feedback on edge cases until all pilot samples pass the predefined criteria. After freezing the rules, we apply the final prompt to all admitted records. We then conduct another batch-wise manual audit with three random seeds; systematic issues trigger rule updates and re-annotation of affected batches, and the audit is repeated until it passes. We do not reproduce the full prompts in this paper; upon acceptance, we will release the final prompts and source-specific prompt configurations in the project GitHub repository.

Dataset size and counting units. The frozen release contains 30,664 records, including 26,826 training records. Table 7 reports the source-specific splits. SocialIQA, CounselChat, and ESD-CoT are counted as question instances, question–answer pairs, and contextual reasoning samples, respectively; each ESConv record is a complete conversation. The 1,300 ESConv conversations contain 18,376 supporter turns (12,759/2,722/2,895 in train/validation/test). Turn counts are nested within conversation counts and are not added to the record totals. ESD-CoT and ESConv together form EQ-Support. These counts describe EQ-Dataset; the canonical evaluation benchmarks retain their separately specified test sets.

Table 7: EQ-Dataset records by source and split. ESConv is counted in complete conversations; a dash denotes an absent split.

Source	Train	Validation	Test	Total
SocialIQA	23,484	1,265	1,407	26,156
CounselChat	1,320	296	—	1,616
ESD-CoT	1,112	160	320	1,592
ESConv	910	195	195	1,300
Total	**26,826**	**1,916**	**1,922**	**30,664**

Source licenses and release boundaries. Table 8 records the conservative release policy applied to the resources used for EQ-Dataset construction and EQ-AppBench evaluation. We do not assert rights beyond those granted by each upstream source.

Table 8: Source terms and planned release boundaries. "Source access" means that users obtain the raw resource from its upstream provider; our release supplies the corresponding adapters, identifiers, manifests, and processing code.

Resource	Documented upstream terms	Project release policy
SocialIQA	Source terms; no broader rights asserted here	Source access; release adapters, split identifiers, and processing manifests.
CounselChat	Upstream repository states the MIT License	Release project processing code and permitted metadata; preserve attribution to the source.
ESD-CoT / ESCoT	No standalone license located in the archived upstream repository	Source access; release mappings, identifiers, and independently separable project artifacts where permitted.
ESConv	CC BY-NC	Preserve attribution and non-commercial use; redistribute project artifacts only to the extent permitted by the upstream license.
CounselBench	CC BY-NC-ND 4.0; evaluation only	Source access; do not redistribute modified versions of the restricted evaluation data.

Commercial-model assistance and evaluator disclosure. GPT-6 Astra was used for source-specific screening, prompt design, and edge-case analysis. Gemini 3.8 Flash (high) was used for pilot annotation under candidate rules, while LLM-as-a-Judge scoring and the GRPO reward model use Qwen3.8 with fixed rubric prompts and deterministic decoding. These proprietary systems were selected for strong general reasoning, but their outputs were not treated as evidence of correctness without the stated human quality controls. Their external benchmark standings provide model-selection context only and do not establish annotation quality or benchmark validity.

A.1 LLM Selection Rationale

Table 9 summarizes the rationale behind each external-model role in data construction and evaluation.

Table 9: Selection rationale for the external LLMs used in this work.

Role	Model	Selection rationale
Data screening & rule authoring	GPT-6 Astra	100% of annotators agreed that screening decisions and rule design require the strongest available model to safeguard data quality.
Dataset construction	Gemini 3.8 Flash, generating conditioned on real human-annotated seed labels; GPT-6 Astra for spot-checking and prompt-rule authoring	66.7% of annotators rated Gemini-family generations as the most natural and human-like; following the minority view, GPT-6 Astra authors the construction prompt rules and audits generated samples.
LLM-as-a-Judge & GRPO reward	Qwen3.8	Most evaluated baselines and our own model share the Qwen backbone and <|im_start|>-style protocols; a same-family judge scores outputs in a format-native context. Judgment-axis scores use deterministic parsing with no LLM judge.

Flagship status and a controlled protocol comparison. Qwen3.8 is the current flagship of the Qwen family, and Gemini 3.8 Flash is the strongest model in the Gemini family: the Gemini Pro flagship line (Gemini 3.1 Pro) has not been updated for an extended period and now lags behind Gemini 3.8 Flash in capability. The judge choice is further supported by a controlled protocol comparison: on 220 supportive-response evaluation items under identical frozen EQ-MER weights, a mismatched legacy dialogue protocol produced empty responses in 73.6% of cases, versus 0% under the native Qwen-format protocol. Generation protocol therefore materially affects small-model outputs; accordingly, each baseline is evaluated under its official generation protocol (Appendix B.2), and the LLM judge only scores responses—it never generates answers on behalf of the models under test.

A.2 Representative Annotated Records

B Evaluation Protocols, Baselines, and Detailed Results

B.1 EQ-UniBench Axis Design

The three axes of EQ-UniBench are each anchored in a distinct scientific tradition. Table 11 summarizes, for every axis, the suite used, the evaluation focus, the corresponding theoretical foundation, and the supporting references.

B.2 MER-UniBench

Reproduction scope and scoring protocol. We evaluate public implementations and author-released checkpoints on the same nine MER-UniBench test sets, comprising 12,799 samples, using a common label-mapping and scoring pipeline. Basic emotion recognition uses HIT, sentiment polarity classification uses weighted F1, and open-vocabulary emotion recognition uses EW-F1; the overall score follows MER-UniBench's aggregation rule (Lian et al., 2025a). All scores are on a 0–100 scale, with higher values indicating better performance. Each model's nine scores come from

Table 10: **Representative annotated records from EQ-Dataset v2.1.** Each record fills the shared UNDERSTAND–PLAN–RESPONSE contract with task-specific semantics: SocialIQA (choice mode) grounds the predicted option in context evidence; CounselChat (response mode) produces strategy-level planning before the supportive reply. Provenance fields (statuses, origins, review markers) are omitted for readability.

SocialIQA (*choice* mode; id `socialiqa_v1.4_train_12637`)

Input. *Context:* After many years of struggling to understand and deal with her family issues, Sasha finally felt closure. *Question:* How would you describe Sasha? *Options:* A. very happy; B. was introspective; C. was a waitress.

<understand> Sasha finally felt closure after spending many years struggling to understand and deal with her family issues. The inference target is a description of Sasha's character or state, grounded on her long struggle and final sense of closure.

<plan> Evaluate which option best describes a personality trait supported by the character's long-term behavior. Spending years struggling to understand personal and family issues points toward being introspective, whereas feeling closure does not necessarily mean being very happy, and her occupation is entirely unmentioned.

<response> **B**

CounselChat (*response* mode; id `counselchat_20220401_00091`)

Input. My mother is combative with me when I say I don't want to talk with her about my depression. She hasn't been supportive of me in the past and she isn't someone that I feel comfortable opening up to. She constantly tries to instigate conversations where she asks me questions that I don't want to or can't answer. I tell her I don't want to talk and she starts arguments with me. How can I get her to understand?

<understand> The seeker's mother becomes combative and starts arguments whenever the seeker refuses to answer questions about their depression. The seeker views their mother as unsupportive and not someone they feel comfortable opening up to.

<plan> Inquire about the living situation, affirm personal autonomy in choosing what to disclose, and recommend establishing firm boundaries or limiting contact. The seeker is facing persistent boundary violations and arguments from an unsupportive mother, making strategies to control communication and protect their emotional health directly relevant.

<response> Do you live with your mom and have constant interaction with her? It is your choice who you share things with. You can set some clear boundaries. She can either talk about the interests you feel comfortable with or you can reduce or eliminate conversations with her. These arguments are most likely not good for your depression. *[truncated]*

a single checkpoint. Methods without obtainable implementations or compatible weights sufficient for reproduction are excluded from the main table; their published scores are listed separately as literature comparisons.

Baseline sources and checkpoints. General-purpose baselines include Qwen2.5-Omni, HumanOmni, VideoLLaMA2.1-AV, VITA-1.5, Baichuan-Omni-1.5, MiniCPM-o-4.5, and video-SALMONN 2 (Xu et al., 2025; Zhao et al., 2025; Cheng et al., 2024b; Fu et al., 2025; Li et al., 2025a; OpenBMB, 2026; Tang et al., 2025). Emotion-finetuned baselines include Emotion-LLaMA, AffectGPT, Emotion-LLaMAv2, Light-MER, and Nano-EmoX (Cheng et al., 2024a; Lian et al., 2025a; Peng et al., 2026; Zheng et al., 2026; Huang et al., 2026). We cite Emotion-LLaMAv2 in its arXiv preprint version (Peng et al., 2026). The seven general-purpose models and Emotion-LLaMA use the archived nine-dataset `fair_v1` results. AffectGPT uses its released `emercoarse` epoch-45 checkpoint, whose configuration differs from the paper's MER-Caption+ configuration. Emotion-LLaMAv2 uses the released stage-2 checkpoint, Light-MER the final released GRPO step-5000 checkpoint, and Nano-EmoX the final released epoch-60 checkpoint. EQ-MER SFT and GRPO use B1.5 epoch 5 and D0.9 epoch 1, respectively. No row combines scores from different checkpoints. In Table 12, asterisks indicate original-paper results, while † and ‡ indicate results reported second-hand by Emotion-LLaMAv2 and AffectGPT. These literature rows retain their source-specific protocols.

Table 11: Three-axis evaluation in EQ-UniBench and its theoretical associations.

Axis and suite	Evaluation focus	Theoretical foundation	References
Emotion perception MER-UniBench	Recognize emotions and sentiment from visual, acoustic, and linguistic cues.	**Multichannel perception:** identify others' emotions through facial expressions, voice, language, and behavior.	Mayer et al. (2016); Lian et al. (2025a).
Emotional reasoning MME-Emotion	Understand emotions and their contextual basis using multimodal evidence.	**Appraisal:** interpret emotions in relation to event relevance, implications, coping potential, and normative significance.	Scherer (2009); Zhang et al. (2026).
Emotional application EQ-AppBench	Make closed-set situational judgments (*Judgment*) and produce open-ended supportive responses (*Response*).	**(i) Emotion management:** judge appropriate responses to situations. **(ii) Supportive communication:** identify support needs, select strategies, and formulate responses.	(i) MacCann & Roberts (2008); (ii) Nozaki & Gross (2025).

Tasks and benchmarks for each axis are carefully selected based on the cited EQ theories.

Inference configurations and result selection. Our common evaluation uses the open-vocabulary emotion-recognition question and scoring procedure introduced by AffectGPT (ICML 2025 Oral), the original MER-UniBench paper (Lian et al., 2025a). The question is "Please recognize all possible emotional states of the character." Model-specific input wrappers and retained official prompts are recorded separately: Emotion-LLaMAv2 retains its official dataset-specific prompt pools and greedy decoding, while Nano-EmoX uses its supported frame-based input path. For Light-MER and Nano-EmoX, we evaluate both the authors' released sampling configurations and the greedy decoding used for EQ-MER, which removes token-sampling randomness. We report the complete run with the higher nine-dataset overall score, without combining per-dataset maxima across configurations. Light-MER scores **73.52** with official sampling and **75.04** with greedy decoding; Nano-EmoX scores **70.21** and **72.55**, respectively. Thus, both main-table rows use greedy results. Table 12 includes both runs' per-dataset scores.

The official sampling configurations use $(\texttt{temperature}, \texttt{top_p}) = (0.9, 0.9)$ for Light-MER and $(1.0, 0.9)$ for Nano-EmoX, with `max_new_tokens=1200`. Both greedy runs use `do_sample=false` and `max_new_tokens=160`. Thus, these are comparisons of complete generation configurations, including their output-length limits.

Comparison with published results. For transparency, we report the measured scores of both generation configurations alongside the corresponding published scores, whether reproduction improves or reduces performance, and retain the higher complete result for baselines evaluated under both configurations. Light-MER achieves **75.04**, exceeding its published **74.61**, whereas Nano-EmoX achieves **72.55**, below its published **74.01** (Zheng et al., 2026; Huang et al., 2026). **Even when published results are included, EQ-MER achieves the highest MER-UniBench overall score among the methods covered in this appendix, at 82.22**, exceeding AffectGPT-R1's reported **79.98** (Lian et al., 2025c). This lead therefore does not depend on lower reproduced scores for some baselines. We also include Emotion-LLaMAv2's higher published stage-1 result of 78.91 alongside its stage-2 result of 78.52 (Peng et al., 2026). Because we have not obtained an official implementation and compatible checkpoint sufficient to reproduce AffectGPT-R1, it is listed only as a literature result. Complete scores and provenance are provided in Table 12.

Parameter scope. The total-size column follows the project's component inventory, rounded to 0.1B. It includes the base language model, frozen modality encoders, and model-side fusion/adaptation components at this reporting precision; released speech-generation branches are included for omni models where present. It excludes external scoring models. Thus, it is a whole-model inventory rather than trainable parameters or the subset active in text-only inference. Values are approximate component totals, not newly measured exact counts for every model.

Table 12: **Provenance-complete MER-UniBench results.** An asterisk marks results reported by the corresponding model's original paper; † and ‡ mark results reported second-hand by Emotion-LLaMAv2 and AffectGPT, respectively. "Ours", "sampling", and the EQ-MER rows are produced under our evaluation environment; sampling rows use official generation settings. Dashes denote unverified parameter counts. Literature rows may use source-specific checkpoints and protocols and are not used for the main-text ranking or bold maxima. Scores are on a 0–100 scale.

		Basic emotion (HIT)				Sentiment (weighted F1)				OV-MERD+	Overall	
Model / provenance	P_{B}	MER23	MER24	MELD	IEMO	MOSI	MOSEI	SIMS	SIMSv2	EW-F1	S	E_P
General-purpose multimodal models												
Qwen2.5-Omni-7B†	10.7	72.84	80.44	43.72	61.53	78.86	61.84	75.44	76.67	53.13	67.16	6.28
Qwen2.5-Omni-7B (ours)	10.7	55.18	44.62	37.09	35.01	63.90	37.53	79.73	79.92	47.49	53.38	4.99
HumanOmni-7B†	9.6	63.36	76.64	37.05	52.52	7.37	7.33	43.87	35.35	59.03	42.50	4.43
HumanOmni-7B (ours)	9.6	65.30	77.42	24.12	47.14	50.18	28.87	75.52	71.61	22.75	51.43	5.36
VideoLLaMA2.1-AV (ours)	7.6	43.58	49.87	29.41	25.32	54.52	50.34	78.01	75.27	11.35	46.41	6.11
VITA-1.5 (ours)	8.0	48.66	65.98	38.42	38.27	49.66	26.94	61.07	54.09	46.68	47.75	5.97
Baichuan-Omni-1.5 (ours)	11.0	50.16	51.66	29.36	44.22	73.83	70.23	82.55	83.32	43.84	58.80	5.35
MiniCPM-o-4.5 (ours)	9.3	60.67	68.18	33.96	50.18	75.45	62.41	83.62	81.76	46.51	62.53	6.72
video-SALMONN 2 (ours)	8.1	41.09	53.86	15.16	25.09	77.94	58.49	81.91	82.80	48.52	53.87	6.65
Emotion-finetuned multimodal models												
Emotion-LLaMA‡	8.2	59.38	73.62	46.76	55.47	66.13	67.66	78.32	77.23	52.97	64.17	7.83
Emotion-LLaMA (ours)	8.2	67.26	76.49	31.21	59.16	76.15	78.79	77.32	76.57	62.79	67.30	8.21
* AffectGPT	8.2	78.54	78.80	55.65	60.54	81.30	80.90	88.49	86.18	62.52	74.77	9.12
AffectGPT (ours)	8.2	69.48	65.00	54.84	49.82	83.37	76.70	80.47	79.76	55.67	68.35	8.33
* AffectGPT-R1	–	84.51	93.13	66.71	74.26	79.65	80.18	87.26	85.75	68.39	79.98	–
* Emotion-LLaMAv2 (S1)	9.3	77.28	86.90	51.32	84.05	86.28	87.69	87.72	86.02	62.90	78.91	8.48
* Emotion-LLaMAv2 (S2)	9.3	76.72	86.90	48.61	83.60	86.96	86.15	86.35	87.42	64.00	78.52	8.44
Emotion-LLaMAv2 (ours)	9.3	72.20	81.41	43.77	83.87	86.87	84.86	84.54	84.12	62.18	75.98	8.17
* Light-MER	0.9	75.66	80.04	57.02	61.46	80.04	77.23	89.05	87.74	63.28	74.61	82.90
Light-MER (ours)	0.9	76.53	85.20	61.20	67.86	75.91	76.06	86.90	84.53	61.15	75.04	83.38
Light-MER (sampling)	0.9	74.04	80.22	58.42	67.72	76.02	74.36	85.16	84.31	61.40	73.52	81.68
* Nano-EmoX	2.5	79.09	77.94	56.55	60.12	76.82	79.81	86.25	84.76	64.75	74.01	29.60
Nano-EmoX (ours)	2.5	76.21	78.13	53.98	57.22	75.06	80.33	86.41	84.54	61.02	72.55	29.02
Nano-EmoX (sampling)	2.5	70.80	73.34	49.40	57.22	74.64	78.76	84.31	83.33	60.08	70.21	28.08
Our two-stage EQ-MER												
EQ-MER: SFT	0.9	80.75	85.59	61.71	61.49	75.45	79.32	87.53	84.21	62.89	75.44	83.82
EQ-MER: + GRPO	0.9	93.25	96.02	80.61	80.70	75.84	78.95	88.18	85.39	61.08	82.22	91.36

EQ-MER is approximately 0.9B including Qwen3-0.6B, CLIP-ViT-Base, Chinese-HuBERT-Base, and Prefusion. Light-MER reports 854.93M total parameters (Zheng et al., 2026), shown as 0.9B here; this exact Light-MER count is not asserted for EQ-MER. The inventory estimates Emotion-LLaMAv2 at 9.3B, Emotion-LLaMA and AffectGPT at 8.2B, Nano-EmoX at 2.5B, MiniCPM-o-4.5 at 9.3B, Baichuan-Omni-1.5 at 11.0B, video-SALMONN 2 at 8.1B, Qwen2.5-Omni-7B at 10.7B, HumanOmni-7B at 9.6B, VITA-1.5 at 8.0B, and VideoLLaMA2.1-AV at 7.6B. For Nano-EmoX, the 2,452M parameters in Table 3 are rounded to 2.5B for Figure 3 and all parameter-efficiency calculations. Published and evaluated scores for a given model use the same rounded inventory denominator. Consequently, the score-per-parameter comparison is approximate and is not a substitute for deployment profiling.

B.3 MME-EMOTION

Tasks and prompts. MME-Emotion uses the official task wordings, candidate labels, and modality adaptation from the paper's Appendix D.2 and the official repository (Zhang et al., 2026); each model answers the same 6,500 real-video questions. The per-task phrasings are kept verbatim: ER-Lab, ER-Wild, and Noise-ER ask for the *most prominent emotion*; FG-ER asks for *one or several emotions*; ML-ER asks for *multiple emotions* under multi-label rules; SA, FG-SA, and IR ask for the most prominent *sentiment*, *fine-grained sentiment*, and *intent* respectively and are not rewritten

as emotion questions. Every item appends its original candidate labels and the instruction "Please think step by step. Enclose your final answer within `<answer>` tags." Models receive the real audiovisual evidence through their native input paths, and the actual modalities and frame counts are logged per item; missing modalities are never replaced by textual descriptions.

Scoring protocol. We follow the official "answer → extract reasoning steps → judge each step" pipeline: step extraction reads only the tested model's answer; step judging consumes up to 10 frames sampled at roughly one frame per second from the same video, pre-extracted Qwen2-Audio audio cues, the original ground truth, and the extracted steps; the final step is judged for class correctness while preceding steps are judged for reasoning evidence. In our runs, **Qwen3.8** replaces the paper's GPT-4.1 (step extraction) and GPT-4o (step judging); Qwen3.8 exceeds both on public capability leaderboards (Artificial Analysis, 2026c), and the full model-selection rationale is given in Table 9. The judge never generates answers on behalf of the models under test.

Metrics and failure handling. Per-item scores are $\mathrm{CoT}_i = (\mathrm{Rec}_i + \mathrm{Rea}_i)/2$, and the Overall is the sample-level micro-average over all 6,500 items rather than a mean over the eight task types. Rec_i is the correctness of the final category (0/1 for single-label items; the fraction of true labels hit for ML-ER), and Rea_i is the average correctness of the reasoning steps preceding the final answer (0 when none exist). Model-side failures—empty answers, format violations, or truncation at the generation limit—contribute zero on the fixed 6,500-item denominator; environment or judge failures never count as model zeros and leave the corresponding scores undefined until rerun. We do not adopt the official script's skip-then-average behavior for unparseable judge outputs.

B.4 EQ-AppBench

Subtests and modes. EQ-AppBench comprises four subtests grouped into two modes by the structure of their answer space (Table 13). *Judgment* subtests are scored deterministically and use no LLM judge: SocialIQA-Test is parsed from the single answer letter in the RESPONSE segment, and EmoBench uses the official EU/EA prompts, JSON outputs, and local parsing. *Response* subtests are scored by an external Qwen3.8 judge over the raw generated output: CounselBench-Eval contributes its 100 real help-seeking questions (Li et al., 2025b), and EQ-Support-Test provides 3,236 structured emotional-reasoning items drawn from the EQ-Support test split. Both response subtests share the same generation prompt and the `upr_xml_v1` output contract, and the same judge prompt; the judge receives the visible task context and the candidate's raw generated output and returns a single line, `Overall EQ Score: N` with $0 \le N \le 10$. Reference answers, expert-selected responses, future dialogue, and internal trace state never enter the model or judge inputs. Since CounselBench-Eval carries no separate emotion field, its judge payload reports `Emotion: Not separately provided` rather than fabricating one.

Table 13: EQ-AppBench subtests, modes, and scoring.

Mode	Subtest	Items	Screen-30%	Scoring
Judgment	SocialIQA-Test	2,224	667	Deterministic letter match
Judgment	EmoBench	800	240	Official deterministic score
Response	CounselBench-Eval	100	30	Unified judge, 0–10
Response	EQ-Support-Test	3,236	971	Unified judge, 0–10

Scoring and aggregation. Each subtest score is normalized to 0–100. The mode and overall scores are

$$J = \frac{S + E}{2}, \qquad R = 10 \times \frac{\bar{G}_{\mathrm{ES}} + \bar{G}_{\mathrm{CB}}}{2}, \qquad B = \frac{J + R}{2}, \tag{7}$$

where S and E are the SocialIQA-Test accuracy and the official EmoBench score, and $\bar{G}_{\mathrm{ES}}$ and $\bar{G}_{\mathrm{CB}}$ are the item-level mean judge scores on EQ-Support-Test and CounselBench-Eval. Subtests are weighted equally within each mode and the two modes are weighted equally, so dataset size does not determine task weight; in particular, the 3,236-item EQ-Support-Test does not dominate the 100-item CounselBench-Eval within the Response

axis. Reported fields are `socialiqa_score`, `emobench_score`, `judgment_score`, `eq_support_test_overall`, `counselbench_eval_overall`, `response_score`, and `eq_bench_score`.

Unified judging protocol. During protocol development we observed that strict structured extraction—requiring a closed `<response>` field or a complete UPR structure—caused widespread parse failures for several baselines (e.g., near-zero compliant outputs for AffectGPT on CounselBench-Eval). We attribute this to weaker instruction-following stability in small models rather than to response quality itself. To maximize fairness across models, we adopt a unified protocol in which the judge scores the candidate's raw generated output directly, without requiring structural field extraction; empty, failed, and over-length generations still contribute zero on the fixed item denominator with raw outputs retained. Over-length cases arise in two ways: most are model-side failures in which the candidate degenerates into hallucinated repetition and never emits EOS, while a small number—chiefly long ESConv-derived dialogues in EQ-Support-Test—exceed the judge's payload limit on the input side and are marked `SKIP_OVER_4096`; both contribute zero identically, without silent truncation.

Checkpoint screening and formal scoring. For checkpoint selection we pre-generate a frozen 30% panel for each subtest (Table 13), shared across checkpoints and baselines; screening scores are used only for selection and diagnostics, never published as formal results. A formal score requires complete generation, parsing, and judging on all four subtests. Model outputs that are empty, failed, or truncated contribute zero on the fixed item denominator with their raw outputs and failure categories retained; missing items, transfer failures, and judge skips or parse errors leave the corresponding subtest, mode, and overall scores undefined rather than imputed from a partial-coverage mean. A legacy five-dimension counsel rubric (coverage, empathy, specificity, factuality, safety, each 0–2) is retained as a diagnostic for CounselBench-Eval and is not part of the formal score.

C EQ-MER Reward and Training Configuration

Supervised fine-tuning objective. The MER-Caption+ training set $\mathcal{D}_{\mathrm{MER}}$ contains examples (x, y) pairing multimodal input and textual context x with a target open-vocabulary emotion answer y. SFT minimizes the negative log-likelihood of the target answer tokens:

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{MER}}}\left[\sum_{j=1}^{|y|}\log p_\theta(y_j \mid y_{<j}, Z(x), T_t(x))\right]. \tag{8}$$

Here, $|y|$ is the answer-token count, j indexes an answer token, $y_{<j}$ is its preceding answer prefix, and p_θ is the model distribution with trainable parameters θ. The multimodal tokens $Z(x)$ and text tokens $T_t(x)$ jointly condition generation. Across the SFT stages, the P2C Bridge and LoRA are trained while the audiovisual encoders and base language-model weights remain frozen; Table 14 records the schedule.

Label-count penalty and answer reward. Equation 4 defines the reward and its label-count penalty P. The counts n_{pred} and n_{ref} refer to extracted emotion labels, not answer tokens or general word counts. No label-count penalty applies when $n_{\mathrm{pred}} \le n_{\mathrm{ref}}$; missing labels are instead reflected in the semantic matching score. Excess predictions receive a count-ratio penalty when $n_{\mathrm{ref}} \le 1$ and a logarithmic-ratio penalty when $n_{\mathrm{ref}} \ge 2$. We use this term to discourage label proliferation as a way to increase hit coverage, a potential reward-hacking behavior. Hit rate is not itself a separate reward term: $\mathrm{EW}_{\mathrm{legacy}}$ averages F1 over level-one category mappings from five emotion wheels. The square root softens the penalty; for example, one reference label and four predicted labels give $P = 1/4$ and a semantic-reward multiplier of $\sqrt{P} = 1/2$. The coefficient 2 scales the semantic component, while R_{fmt} is one for compliance with the prescribed short emotion-answer format and zero otherwise.

Emotion-wheel semantic matching. The term $\mathrm{EW}_{\mathrm{legacy}}$ uses fixed lexical mappings and category-set overlap. Let $L(\hat{y})$ and $L(y)$ be the emotion-label lists extracted from the generated

and reference answers. For wheel $k \in \{1, \ldots, 5\}$, the mapping M_k applies word-form normalization, synonym lookup, and assignment to level-one categories:

$$\widehat{S}_k = \mathrm{set}(M_k(L(\hat{y}))), \qquad S_k = \mathrm{set}(M_k(L(y))). \tag{9}$$

Unmapped labels are discarded, and taking sets removes duplicate categories. When synonym lookup yields multiple candidates, the implementation selects the first wheel-compatible candidate in sorted order. For nonempty sets and positive overlap, precision, recall, and F1 are

$$p_k = \frac{|\widehat{S}_k \cap S_k|}{|\widehat{S}_k|}, \qquad r_k = \frac{|\widehat{S}_k \cap S_k|}{|S_k|}, \qquad F_{1,k} = \frac{2p_k r_k}{p_k + r_k}. \tag{10}$$

Here, $\widehat{S}_k$ and S_k are the predicted and reference category sets, p_k measures prediction precision, and r_k measures reference coverage. The score averages the five per-wheel F1 values when the result is finite. To state the recorded implementation's boundary behavior explicitly, let u be the raw five-wheel mean:

$$u = \frac{1}{5}\sum_{k=1}^{5} F_{1,k}, \qquad \mathrm{EW}_{\mathrm{legacy}}(\hat{y}, y) = \begin{cases} \min(1, \max(0, u)), & u \text{ is finite,} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Legacy boundary behavior and relation to P. The training path calls the wheel scorer on one answer pair at a time. It retains the scorer's raw boundary behavior: an empty mapped reference set is skipped, an empty mapped prediction receives zero precision and recall, and a zero precision-plus-recall denominator is not protected within the per-wheel F1 calculation. Consequently, a non-finite value from one wheel can propagate through the mean and cause the entire EW term to be set to zero by the reward wrapper. This differs from setting only the affected wheel's F1 to zero before averaging. The subscript *legacy* identifies this inherited set-based scoring path. Unlike EW, which compares mapped, deduplicated category sets, P counts extracted labels before wheel mapping and deduplication. Multiple emotion words mapping to the same category can therefore leave set F1 unchanged while increasing the label-count penalty.

Group-relative advantages and update scope. Answer-level GRPO starts from B1.5 joint epoch 5 and uses 1,000 MER2025-OV training examples. For each input, we sample $K = 4$ candidate answers and standardize their rewards to obtain relative advantages:

$$A_i = \frac{R_i - \bar{R}}{\mathrm{Std}(R_1, \ldots, R_K) + \epsilon}, \qquad \bar{R} = \frac{1}{K}\sum_{j=1}^{K} R_j. \tag{12}$$

Here, R_i is candidate i's reward, $\bar{R}$ and Std are the group mean and population standard deviation, and ϵ is a numerical stabilizer. The advantage A_i measures an answer's quality relative to other candidates for the same input. Training favors relatively higher-reward answers, with KL regularization constraining policy drift. The reference distribution is computed using the same frozen backbone with LoRA disabled. This stage updates only LoRA, freezing the audiovisual encoders and the complete P2C Bridge.

Recorded optimization settings. Table 14 summarizes the configurations that produced the SFT checkpoint and the final EQ-MER checkpoint. It is an implementation record rather than an additional hyperparameter study.

D EQ-AGENTIC TRAINING OBJECTIVES

Agentic-SFT objective. For each pool $m \in \{\mathrm{SIQA}, \mathrm{CC}, \mathrm{ESDCoT}\}$, the loss is the token-mean negative log-likelihood over the pool's valid assistant target tokens,

$$\mathcal{L}_m = \frac{\sum_{i \in \mathcal{B}_m} \sum_{t \in T_i} -\log p_\theta(y_{i,t} \mid x_i, y_{i,<t})}{\sum_{i \in \mathcal{B}_m} |T_i|}, \tag{13}$$

where $\mathcal{B}_m$ collects pool-m samples in an update and T_i the valid target tokens of sample i (UPR tags, segment bodies, terminator, and trailing newline); all other positions are masked. The SFT objective then averages the three pool losses equally (Equation 5).

Table 14: Compact EQ-MER training record. The selected checkpoints are B1.5 joint epoch 5 for SFT and D0.9 epoch 1 for answer-level GRPO.

Setting	SFT	Answer-level GRPO
Initialization	Foundation models; staged grounding, integration, and joint training	B1.5 joint epoch 5
Training budget	1+1+6 stage epochs; 62 steps per epoch	3 epochs; 250 steps per epoch
Batching	4 workers; 128 samples per worker; global batch 512	$K = 4$ candidates per input
Optimization	Joint peak LR 1.55×10^{-5}; LoRA rank 16, $\alpha = 72$, dropout 0.025	LoRA LR 2×10^{-6}; clip 0.1; KL coefficient 0.15
Generation	—	Temperature 1.2; top-p 0.9
Updated modules	Prefusion and Qwen LoRA; modality encoders and base LLM frozen	Qwen LoRA only; encoders and complete Prefusion frozen
Selection	Highest complete nine-dataset checkpoint: joint epoch 5	Highest complete nine-dataset checkpoint: epoch 1

Agentic-RL objective. Agentic-RL applies the standard GRPO clipped surrogate (Shao et al., 2024) with token-level aggregation:

$$\mathcal{J}_{\mathrm{RL}} = \mathbb{E}_{x,\{o_k\}} \left[\frac{1}{K} \sum_{k=1}^{K} \frac{1}{|o_k|} \sum_{t=1}^{|o_k|} \min(\rho_{k,t} A_k,\ \mathrm{clip}(\rho_{k,t}, 1-\epsilon, 1+\epsilon)\, A_k) \right] - \beta\, \mathbb{D}_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{SFT}}) \,, \tag{14}$$

with token-level likelihood ratio $\rho_{k,t} = \pi_\theta(o_{k,t} \mid x, o_{k,<t}) / \pi_{\mathrm{old}}(o_{k,t} \mid x, o_{k,<t})$, clip ϵ=0.2, and KL coefficient β=0.02 toward the frozen SFT reference π_{SFT}; π_{old} denotes the behavior policy that sampled the group.

E EQ-HARNESS: TASK PROTOCOLS, SAFETY ALIGNMENT, AND DEPLOYMENT

E.1 TASK PROTOCOLS AND EQ TRACE

Request and result contract. Each EQ-Harness call is specified by a caller-provided stable sample ID, explicit task mode, task name, original language, original task or environment record, protocol revision, prompt revision, and generation configuration. The harness uses these fields to construct the task input, select the corresponding renderer and parser, and return the raw model output, parsed result, status, parser reason, protocol revisions, token counts, EOS and truncation flags, and evaluation payload. The two EQ task modes, *judgment* and *response*, share the `upr_xml_v1` output contract, while source-specific adapters interpret the fields; the model never infers the task mode.

The protocol assembly is

$$\begin{aligned} \text{system} &= \text{COMMON_SYSTEM} \\ &\quad \oplus \text{TASK_PROMPT}_t \\ &\quad \oplus \text{OUTPUT_CONTRACT}_{m,t}, \\ \text{user} &= \text{RENDERER}_t(\text{record}), \end{aligned} \tag{15}$$

where m is the explicit task mode, taking values *judgment* or *response*, and t is the task name. The operator $\oplus$ denotes fixed-order concatenation; record is the original task or environment record. The renderer organizes the task facts without sending reference answers, Judge rubrics, or future dialogue turns to the model.

Trace, caching, and failure handling. Each trace records the sample ID, checkpoint, task mode and name, protocol, prompt, renderer, and parser revisions, a digest of the model input, raw and parsed outputs, status, token counts, EOS and truncation flags, worker or rank, and the evaluation payload. Cache identity includes the task mode, prompt, renderer, parser, checkpoint, sample ID, and decoding configuration, so different modes cannot share a cached result. Raw output is always retained. A parser failure is recorded as its original status—for example, `empty`, `malformed`, `truncated`, or `invalid_contract`—and cannot be repaired by cross-mode fallback, silent XML editing, tag insertion, or answer guessing. Judge payloads are constructed from the mode-specific parsed result and exclude reference answers, trace state, and internal diagnostics; for response tasks the judge receives the task context and the candidate's raw generated output and returns a single overall score.

E.2 TRAINING AND INFERENCE PROMPTS FOR SAFETY ALIGNMENT

EQ-Agent's safety alignment operates at two layers. At the base-model layer, the Qwen3 language backbone already undergoes dedicated safety alignment in its official post-training pipeline (Qwen Team, 2025). On top of this, EQ-Harness applies a prompt-level behavioral layer: every task shares the same `COMMON_SYSTEM` block (`common_system_compact_v2`), whose final sentence is an explicit safety clause:

> You are a highly emotionally intelligent conversational assistant.
> Recognize the emotions and needs in the provided context, consider different perspectives, and respond to the task with empathy and sound judgment.
> **Do not invent facts, ignore risks, or provide harmful guidance.**

For safety-sensitive support scenarios, the inference-time output contract (`counselchat_upr_v1`) adds further behavioral constraints: the plan must be supportive and practical, and the response must answer directly without leaking analysis or strategy labels:

> Output exactly these three XML elements in this order and nothing else: `<understand>`, `<plan>`, `<response>`.
> In `<understand>`, briefly identify the user's main concern, emotional state, and relevant context.
> In `<plan>`, briefly plan a **supportive, specific, and practical** response.
> In `<response>`, write a direct, empathetic, and context-specific answer to the user's question. **Do not put analysis, strategy labels, or meta-commentary inside `<response>`.**

Because SFT targets and inference share the same assembled prompt (Eq. 15), the safety clause is supervised into model behavior during training rather than applied only as an inference-time patch.

E.3 Deployment, Inference Acceleration, and Interactive Interface

We provide a multimodal deployment and inference-acceleration framework for personal devices, unifying audiovisual input processing, model execution, and interactive interfaces. Model residency and pre-deployment LoRA merging reduce repeated loading and adapter execution overhead. Fixed-shape TensorRT FP16 engines and device-resident tensor bindings accelerate visual and acoustic encoding, while generation-time key–value caching, streaming speech recognition, and clip-specific feature caching reduce processing waits. HTTP and Server-Sent Events interfaces deliver execution status and generated content to the client, supporting incremental text rendering and polling as a fallback.

The framework has been deployed on a Jetson Orin Nano Super with 8 GB of memory. In archived small-sample, task-specific short-answer tests, latency to the first streamed text chunk was approximately **0.45–0.49 s**, and single-turn completion took approximately **1.74–1.78 s** after model warm-up and feature preparation. Compared with the earlier approximately **14.8 s** wait for a complete answer, the best first-chunk latency represents an approximately $\mathbf{33}\times$ reduction in the wait for initial streamed output; this comparison concerns chunk arrival, not total execution-time acceleration or latency to the first visible character. In a separate server-side loading measurement, EQ-MER occupied **2.79 GiB** of GPU memory, including **2.20 GiB** allocated by PyTorch, excluding input activations and generation-time key–value caches. **Upon acceptance, we will release the inference framework, encoder-export tools, deployment scripts, and interactive-interface implementation as open source.** More broadly, upon acceptance we will open-source the complete project—models, code, data pipelines, prompts, evaluation harnesses, and documentation—to facilitate community development and reproducible research on emotionally intelligent agents.