EMOGPT: TOWARDS EXPLAINABLE MULTIMODAL EMOTION RECOGNITION WITH CROSS-MODAL CHAIN-OF-THOUGHT SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding human emotions from multimodal signals requires reasoning over potentially conflicting cues across facial expressions, vocal prosody, and language. However, existing Multimodal Large Language Models (MLLMs) rely primarily on categorical supervision or unstructured descriptions, limiting their ability to perform interpretable cross-modal reasoning, especially in cases of semantic–prosodic inconsistency such as sarcasm. In this work, we introduce structured cross-modal Chain-of-Thought (CoT) supervision for multimodal emotion recognition. We present **EmoGPT**, a framework that learns to generate emotion predictions grounded in explicit reasoning over visual, acoustic, and textual evidence. To enable this, we construct **ECoT-65K**, a large-scale Emotional Chain-of-Thought dataset, designed to capture modality-specific evidence and their interactions. Beyond supervision, we propose two inductive mechanisms to improve multimodal reasoning. First, **Affective Visual Routing (AVR)** introduces a mixture-of-experts projector that adaptively routes visual tokens based on emotion-relevant semantics. Second, **Emotional Saliency Bias (ESBias)** introduces zero-initialized, layer-specific Q/K/V/O offsets that complement LoRA with emotion-sensitive affine adaptation. Extensive experiments on nine benchmarks under three modality configurations show that EmoGPT achieves the highest mean accuracy in every setting, improving the AVT mean by 4.51 points over the previous state of the art; notably, fine-tuning on ECoT-65K alone already surpasses the previous state of the art without any architectural change. A human audit further shows that 94.3% of the retained reasoning annotations are rated as grounded in multimodal evidence, advancing multimodal emotion recognition toward interpretable and robust cross-modal reasoning.

1 INTRODUCTION

Emotional information conveyed through facial expressions, vocal prosody, and verbal content plays a central role in human communication Ekman (1992); Russell (1980), making multimodal emotion recognition (MER) important for affective computing, mental health monitoring, and human-computer interaction Poria et al. (2017). Recent Multimodal Large Language Models (MLLMs) Liu et al. (2023); Xu et al. (2025b) provide a unified backbone for visual, acoustic, and textual reasoning, but current emotion-oriented MLLMs Lian et al. (2025); Cheng et al. (2024) are still trained mainly with categorical labels or modality-parallel captions. Such supervision can teach a model *what* emotion to output, but not *how* to reconcile conflicting evidence when modalities disagree, as in sarcasm or masked affect. We therefore reformulate MER as a structured cross-modal reasoning problem and aim to supervise not only the final label but also the evidence path that leads to it.

Realizing this view raises three technical challenges. First, existing emotion datasets rarely provide structured reasoning traces that jointly encode Action Units, acoustic prosody, cross-modal conflict resolution, and pragmatic intent. Second, the visual projector in standard MLLM backbones applies a single pathway to all visual tokens, even though affective cues from faces, body posture, and scene context are semantically heterogeneous. Third, standard LoRA adapts projection weights through low-rank updates but introduces no explicit affine pathway for recalibrating attention and hidden representations toward subtle affective evidence.

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.

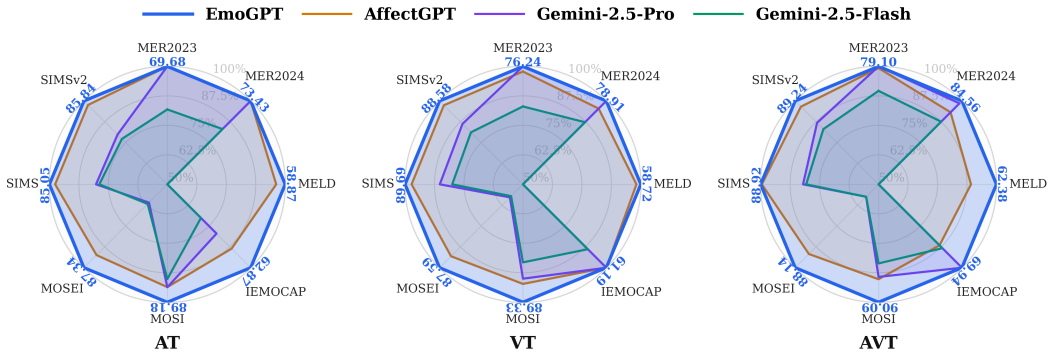


Figure 1: **Radar comparison across modality settings.** The proposed EmoGPT outperforms AffectGPT, Gemini-2.5-Flash, and Gemini-2.5-Pro across AT, VT, and AVT settings.

To address these challenges, we build EmoGPT around a structured "data–supervision–architecture" recipe that treats cross-modal reasoning as training objective. We first construct **ECoT-65K**, a 65,370-sample Emotional CoT dataset sourced from seven corpora Zadeh et al. (2016; 2018); Poria et al. (2019); Lian et al. (2024); Liu et al. (2022a; 2024); Sun et al. (2025), where each annotation separates modality extraction, inference logic, and final judgment through a five-stage generation pipeline followed by Generator–Critic auditing. We fine-tune Qwen2.5-Omni-7B Xu et al. (2025b) on this corpus with two lightweight architectural extensions: **Affective Visual Routing (AVR)**, a four-expert top-2 MoE visual projector that replaces the default patch merger to capture emotion-specific visual patterns, and **Emotional Saliency Bias (ESBias)**, a set of zero-initialized, layer-specific Q/K/V/O offsets learned jointly with LoRA to provide emotion-sensitive affine adaptation.

Together, ECoT-65K, AVR, and ESBias push the AVT mean accuracy from 67.16% to 79.28%, achieving the highest mean accuracy under all three modality configurations on MER-UniBench.

EmoGPT delivers consistent gains across three modality configurations (AVT, AT, VT) on MER-UniBench, a comprehensive benchmark comprising nine datasets spanning English and Chinese, acted and in-the-wild scenarios. Compared with the previous SOTA AffectGPT Lian et al. (2025), EmoGPT improves mean accuracy by $+4.51\%$ points under AVT, $+1.93\%$ under AT, and $+2.20\%$ under VT. Figure 1 visualizes this comparison, showing that EmoGPT, as a 7B model, consistently outperforms both open-source AffectGPT and closed-source Gemini-2.5-Flash and Gemini-2.5-Pro across all modality settings. Our contributions are summarised as follows:

- We propose the first structured cross-modal CoT framework for trimodal (audio–video–text) MER, and construct **ECoT-65K**, a large-scale Emotional CoT dataset with annotated visual AUs, acoustic prosody, pragmatic intent, and cross-modal conflict resolution.
- We propose two lightweight architectural modules, **AVR**, an emotion-specialised MoE visual projector, and **ESBias**, an explicit layer-specific Q/K/V/O adaptation module that complements LoRA with emotion-sensitive affine offsets.
- EmoGPT achieves the highest mean accuracy under all three modality configurations on MER-UniBench, outperforming the previous state of the art by up to 4.51 points on the AVT mean and advancing MER toward interpretable cross-modal reasoning.

2 Related Work

Multimodal Emotion Recognition. Multimodal emotion recognition (MER) has long been formulated as mapping visual, acoustic, and textual signals into an emotion label. Classical methods focus on feature fusion and alignment Zadeh et al. (2017); Tsai et al. (2019), while modern benchmarks still formulate MER as classification over fixed label spaces Poria et al. (2019); Busso et al. (2008); Lian et al. (2023; 2024). This label-only supervision leaves evidence paths implicit, making predictions uninterpretable and offering no principled way to resolve cross-modal conflict.

Multimodal Large Language Models for Emotion Analysis. MLLMs have recently shifted MER from classifier heads to instruction-following multimodal generation. General-purpose systems such as SALMONN and Qwen-Audio Tang et al. (2023); Chu et al. (2023) treat emotion recognition as one downstream task, while emotion-oriented models such as Emotion-LLaMA and

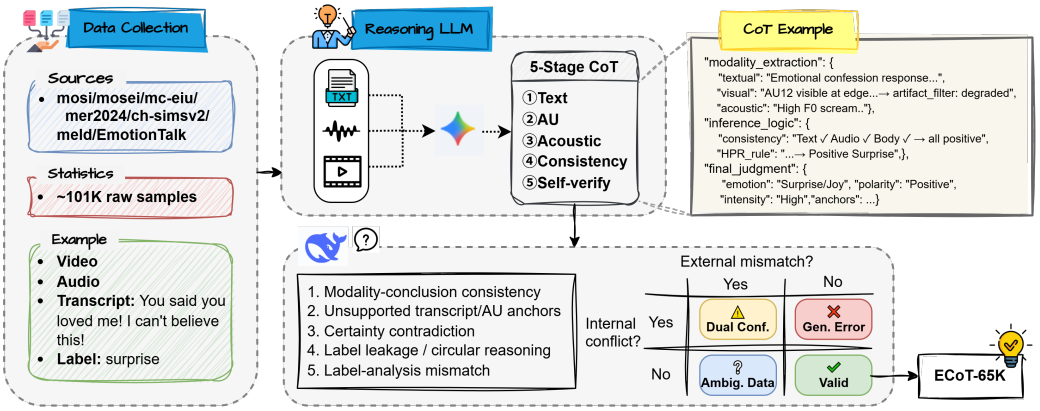


Figure 2: **ECoT-65K construction.** Gemini generates five-stage CoT annotations from several source corpora. An external DeepSeek critic then organizes annotations into four audit outcomes, *Valid*, *Gen. Error*, *Ambig. Data*, and *Dual Conf.* Only *Valid* cases are retained in ECoT-65K.

AffectGPT Cheng et al. (2024); Lian et al. (2025) fine-tune MLLMs on dedicated emotion corpora. However, most existing corpora provide labels or captions rather than a structured reasoning chain, so they do not teach the model how to reconcile conflicting modalities before prediction.

Chain-of-Thought Reasoning in Multimodal Learning. Chain-of-Thought (CoT) reasoning has become a useful way to supervise intermediate inference in large models Wei et al. (2022); Zhang et al. (2023); Hsieh et al. (2023); Xu et al. (2025a). The concurrent EmotionThinker Wang et al. (2026) is especially relevant in showing that emotion reasoning benefits from structured intermediate supervision, but it focuses on speech-only settings. Our setting is harder because trimodal MER additionally requires explicit visual AU cues and a decision rule for cross-modal conflict, which are the two components ECoT-65K is designed to supply.

3 METHOD

Given synchronized visual, acoustic, and textual inputs (V, A, T), EmoGPT predicts an emotion label together with a structured reasoning trace grounded in cross-modal evidence. Figure 4 shows how AVR and ESBias are integrated into the CoT-supervised backbone to provide routed visual projection and emotion-sensitive affine adaptation throughout the language backbone.

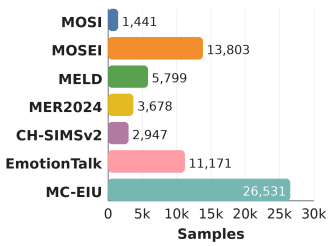
We focus the main text on the core pipeline and defer exhaustive prompts, training configuration, and extended analyses to Appendix C.

3.1 ECOT-65K: A LARGE-SCALE EMOTIONAL CHAIN-OF-THOUGHT DATASET

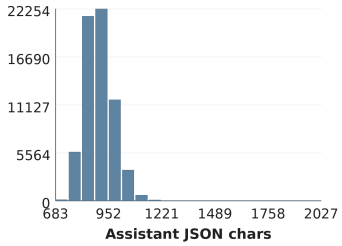
Categorical labels supervise only final predictions, whereas descriptive captions provide parallel modality-wise observations without an explicit inferential path from evidence to judgment. Neither format directly supervises distractor identification, cross-modal conflict resolution, or evidence grounding. We therefore construct ECoT-65K, whose targets are structured JSON traces linking emotion judgments to visual, acoustic, and textual evidence. ECoT-65K contains 65,370 video–reasoning pairs drawn from seven public corpora spanning English and Chinese, as well as acted and in-the-wild settings. Among the datasets compared in Table 1, it is the only one covering all six annotation dimensions. Each trace follows a three-block schema: `modality_extraction` stores observable evidence, `inference_logic` records cross-modal reconciliation and intent reasoning, and `final_judgment` contains the emotion, intent, polarity hint, intensity, certainty, and key evidence anchors. As illustrated in Figure 2, annotations are produced through five stages: textual sentiment analysis, AU-based visual analysis, acoustic prosody analysis, cross-modal consistency analysis, and final label-consistency verification. When modalities disagree, the prompt first assesses the ambiguity of each modality and then uses text if unambiguous, audio if the text is inconclusive, and visual evidence as the final fallback. This heuristic regularizes annotation generation rather than defining a universal ranking of modality reliability; non-verbal cues may dominate in cases such as sarcasm or masked affect. The generator is instructed to first extract modality-specific evidence, reconcile cross-modal cues, and derive a provisional emotion judgment; only then does it compare that judgment with the reference label in the final self-verification stage. Following

Table 1: Comparison of representative emotion datasets by annotation dimension. **Desc.**: natural-language descriptions; **CoT**: structured reasoning traces; **AU**: facial Action Units; **Pros.**: prosodic features; **Conflict**: cross-modal conflict rules; **Filter**: critic audit for logical or label-alignment errors. ECoT-65K is the only dataset covering all six dimensions.

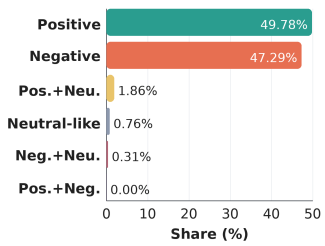
Group	Dataset	Mod.	#Samp.	Desc.	CoT	AU	Pros.	Conflict	Filter
Categorical	MELD Poria et al. (2019)	A,V,T	13.7K	×	×	×	×	×	×
	IEMOCAP Busso et al. (2008)	A,V,T	7.4K	×	×	×	×	×	×
	MER2023 Lian et al. (2023)	A,V,T	5K	×	×	×	×	×	×
Descriptive / CoT	MAFW Liu et al. (2022b)	A,V,T	10K	✓	×	×	×	×	×
	MERR Cheng et al. (2024)	A,V,T	28K	✓	×	×	×	×	×
	MER-Caption+ Lian et al. (2025)	A,V,T	31K	✓	×	×	×	×	×
	EmotionCoT-35K Wang et al. (2026)	A,T	35K	✓	✓	×	✓	×	×
	ECoT-65K (Ours)	A,V,T	65K	✓	✓	✓	✓	✓	✓



(a) Source-corpus composition across the seven contributing datasets.



(b) Distribution of assistant-side structured CoT target lengths, measured as the character count of the JSON supervision string.



(c) Coarse polarity distribution obtained by normalising the free-form `mapping_hint` field into a small set of polarity buckets.

Figure 3: Visual statistics of the ECoT-65K training set. The polarity distribution in (c) reflects our post-hoc normalisation of CoT annotations rather than the original labels of the source datasets.

self-verification, an external critic receives the transcript, reference label, and generated JSON to check textual and internal consistency, unsupported links between evidence and conclusions, certainty contradictions, circular reasoning, and label alignment. Because the critic does not inspect the raw audiovisual streams, it serves as a logical and label-alignment filter rather than a complete factual verifier; multimodal factual grounding is evaluated separately through the human audit in Section 4.1. Critic outcomes are summarized using two binary indicators—internal logical conflict and label mismatch—yielding four audit cells: *Valid* (neither), *Gen. Error* (internal conflict only), *Ambig. Data* (label mismatch only), and *Dual Conf.* (both). Only *Valid* cases are retained in ECoT-65K. Figure 3 summarizes the source-corpus distribution, CoT length distribution, and coarse polarity distribution; complete statistics, prompts, and examples are provided in Appendix B.

3.2 Affective Visual Routing

Visual signals in emotion analysis are semantically heterogeneous. A single clip typically contains three functionally distinct token categories: facial regions encoding Action Units, body regions encoding gesture and posture, and background regions encoding scene context. Each category contributes a different conditional distribution over emotion, and in some cases (e.g., costumes, occlusions) the background region should be suppressed entirely. A single-pathway projector forces all three categories to share one set of parameters (Liu et al., 2023; Ye et al., 2023), and prior work on mixture-of-experts architectures (Shazeer et al., 2017; Lin et al., 2026) suggests that routing-based designs better accommodate semantically heterogeneous inputs.

We replace the original Qwen2.5-Omni visual merger with **Affective Visual Routing** (AVR), a sparsely-gated Mixture-of-Experts projector whose experts specialise for different affective token categories. Formally, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times d_v}$ denote the vision-encoder output of a clip, and let $\{E_i : \mathbb{R}^{d_v} \to \mathbb{R}^{d_\ell}\}_{i=1}^{M}$ be $M = 4$ MLPs mapping visual features into the language-model hidden space. A lightweight router $G_\phi(\cdot) \in \mathbb{R}^M$ produces per-token routing logits, which are converted into a sparse gating distribution via top-k selection with $k = 2$:

$$\pi(\mathbf{x}_n) = \mathrm{softmax}\big(\mathrm{TopK}(G_\phi(\mathbf{x}_n),\, k)\big), \tag{1}$$

where $\mathrm{TopK}(\mathbf{z}, k)$ keeps the k largest entries of $\mathbf{z}$ and sets the rest to $-\infty$ before the softmax, producing a k-sparse probability vector over the M experts. The projected visual token is then the

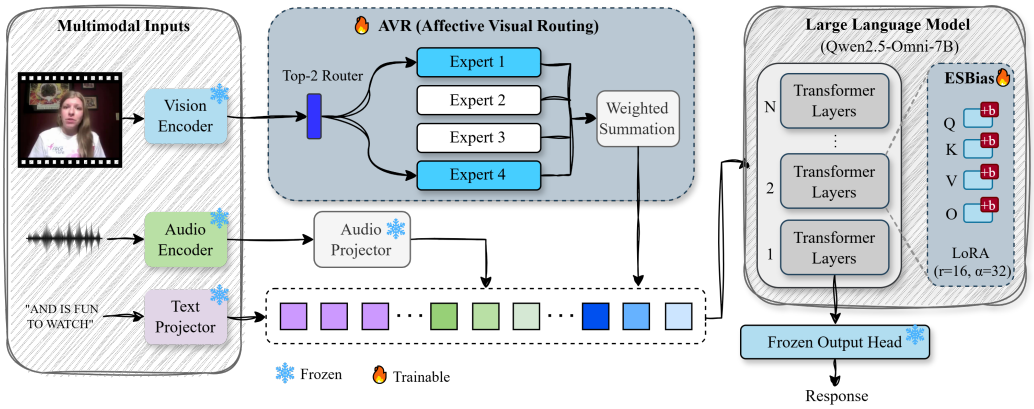


Figure 4: **Architecture of EmoGPT.** EmoGPT keeps the Qwen2.5-Omni-7B backbone and adds AVR to the visual merger and ESBias to attention projections.

expectation of the expert outputs under $\pi(\mathbf{x}_n)$,

$$\mathrm{AVR}(\mathbf{x}_n) = \sum_{i=1}^{M} \pi_i(\mathbf{x}_n)\, E_i(\mathbf{x}_n) = \mathbb{E}_{i\sim\pi(\mathbf{x}_n)}\big[E_i(\mathbf{x}_n)\big], \tag{2}$$

which can be read as a *data-conditioned affine map* whose effective weight matrix $\sum_i \pi_i(\mathbf{x})W_i$ varies with the affective category of the input token, letting different visual cues flow through specialised pathways. To prevent expert collapse, where a few experts dominate routing while the others receive near-zero gradient, we follow Shazeer et al. (2017) and add an auxiliary load-balancing loss

$$\mathcal{L}_{\mathrm{bal}} = M \cdot \sum_{i=1}^{M} \bar{f}_i\, \bar{\pi}_i, \qquad \bar{f}_i = \tfrac{1}{N}\sum_{n=1}^{N} \mathbb{1}[i \in \mathrm{Top}k(G_\phi(\mathbf{x}_n))], \quad \bar{\pi}_i = \tfrac{1}{N}\sum_{n=1}^{N} \pi_i(\mathbf{x}_n), \tag{3}$$

which penalises the correlation between token assignment frequency $\bar{f}_i$ and mean routing probability $\bar{\pi}_i$, encouraging a uniform utilisation of all experts. The final training objective is $\mathcal{L} = \mathcal{L}_{\mathrm{CoT}} + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}}$ with $\lambda_{\mathrm{bal}} = 10^{-2}$. AVR expands the merger from 44M to 178M parameters while keeping the per-token FLOPs at only $2/M = 50\%$ of a dense M-expert counterpart, since only two of four experts are active per token, preserving inference efficiency despite the larger parameter count.

AVR provides an *input-conditioned* projection whose effective weight matrix adapts to the affective semantics of each visual token, whereas the original single-MLP merger imposes a *shared* projection on all tokens. Equation (2) makes this explicit: the effective map $\sum_i \pi_i(\mathbf{x})W_i$ is a point on the convex hull of the four expert matrices, parameterised by the router, and therefore AVR strictly contains the original merger as the degenerate case $\pi = [1/M, \ldots, 1/M]$. More importantly, the AVR induces structurally mirrors the `visual` sub-chain of our CoT supervision, which itself separates AU evidence, body-posture evidence, and artefact cues; the architectural prior of AVR therefore aligns with the annotation prior of ECoT-65K, making the CoT easier for the backbone to follow.

3.3 Emotional Saliency Bias

LoRA parameterises the adaptation of an attention projection $W \in \mathbb{R}^{d_{\mathrm{in}}\times d_{\mathrm{out}}}$ as

$$W' = W + \frac{\alpha}{r}\mathbf{B}\mathbf{A}, \qquad \mathbf{B} \in \mathbb{R}^{d_{\mathrm{in}}\times r}, \quad \mathbf{A} \in \mathbb{R}^{r\times d_{\mathrm{out}}}, \quad r \ll \min(d_{\mathrm{in}}, d_{\mathrm{out}}), \tag{4}$$

where the low-rank update adapts projections while preserving pretrained weights. This makes LoRA a purely *multiplicative* perturbation: its contribution to any attention logit factorises through the content interaction of the two tokens within a low-rank subspace. LoRA can therefore rotate and rescale the query–key geometry, but no adapter of the form in Equation (4) can inject a *query-independent* preference for a particular token—precisely the capacity that affective saliency requires, since an AU4 brow-lowerer or a high-F_0 scream should attract attention *regardless* of the current query. We therefore complement LoRA with lightweight, emotion-sensitive additive adaptation.

We introduce **Emotional Saliency Bias** (ESBias), which attaches four zero-initialized, layer-specific vectors $\mathbf{b}_Q^{(\ell)}$, $\mathbf{b}_K^{(\ell)}$, $\mathbf{b}_V^{(\ell)}$, and $\mathbf{b}_O^{(\ell)}$ to the query, key, value, and output projections of each language-

model attention layer to enable lightweight, emotion-sensitive saliency recalibration:

$$\overline{\mathbf{Q}} = \mathbf{X}W'_Q + \mathbf{1}\mathbf{b}_Q, \qquad \overline{\mathbf{K}} = \mathbf{X}W'_K + \mathbf{1}\mathbf{b}_K,$$
$$\mathbf{V} = \mathbf{X}W'_V + \mathbf{1}\mathbf{b}_V, \qquad \mathbf{Z} = \mathrm{Attn}(\overline{\mathbf{Q}}, \overline{\mathbf{K}}, \mathbf{V})W'_O + \mathbf{1}\mathbf{b}_O. \tag{5}$$

Here, $\overline{\mathbf{Q}}$ and $\overline{\mathbf{K}}$ denote the query and key representations before rotary positional encoding, which is applied within $\mathrm{Attn}(\cdot)$. We omit any pretrained projection biases from Equation (5); these parameters remain frozen, and ESBias refers only to the newly introduced offsets. All ESBias vectors are zero-initialized, so the modified model exactly recovers the pretrained forward pass at initialization, and are jointly optimized with LoRA under ECoT supervision. In Qwen2.5-Omni-7B, the query and output offsets each contain 3,584 parameters per layer, whereas the key and value offsets each contain 512 under grouped-query attention. Across 28 language-model attention layers, ESBias adds 229,376 trainable parameters, less than 0.004% of the backbone.

To characterize how ESBias complements LoRA, let $\mathbf{P}_{ij} = \mathbf{R}_i^\top \mathbf{R}_j$ denote the relative rotary transform between query position i and key position j. The unscaled attention logit s_{ij} expands as

$$s_{ij} = (\mathbf{x}_i W'_Q + \mathbf{b}_Q)\mathbf{P}_{ij}(\mathbf{x}_j W'_K + \mathbf{b}_K)^\top = \underbrace{\mathbf{x}_i W'_Q \mathbf{P}_{ij} {W'_K}^\top \mathbf{x}_j^\top}_{\text{LoRA-adapted relational}} + \underbrace{\mathbf{x}_i W'_Q \mathbf{P}_{ij} \mathbf{b}_K^\top}_{\text{query-conditioned affine}}$$
$$+ \underbrace{\mathbf{b}_Q \mathbf{P}_{ij} {W'_K}^\top \mathbf{x}_j^\top}_{\text{key-content preference}} + \underbrace{\mathbf{b}_Q \mathbf{P}_{ij} \mathbf{b}_K^\top}_{\text{content-independent positional prior}}. \tag{6}$$

The four terms expose complementary components of attention adaptation. The first is the LoRA-adapted content–content interaction governing pairwise query–key similarity. The second couples query content with the learned key offset. More importantly, the third is independent of the query *content* and introduces a layer-specific preference over key representations, modulated by their relative positions through $\mathbf{P}_{ij}$. The fourth is independent of token content and contributes a learned relative-position prior. Thus, whereas LoRA modifies the bilinear content–content interaction through low-rank weight updates, ESBias introduces explicit trainable content–bias and bias–bias interactions that are not independently adapted by LoRA alone.

The value and output offsets recalibrate the attended representation and output hidden state, respectively. We therefore view LoRA and ESBias as a multiplicative–additive decomposition of attention adaptation: LoRA adapts relational geometry, while ESBias provides explicit affine pathways encoding saliency preferences. The offsets are made emotion-sensitive through ECoT supervision, which optimizes them toward affective evidence useful for cross-modal emotion reasoning.

3.4 Training Strategy

We train EmoGPT in a single SFT stage on ECoT-65K with next-token loss over the full structured reasoning trace. For each training example, let $\mathbf{x}_i = (V_i, A_i, T_i)$ denote the multimodal input and let $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,L_i})$ denote the target structured JSON CoT sequence from ECoT-65K. The supervision loss is the assistant-token negative log-likelihood

$$\mathcal{L}_{\mathrm{CoT}} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\frac{1}{L_i}\sum_{t=1}^{L_i}\log p_\theta(y_{i,t} \mid \mathbf{x}_i, y_{i,<t}), \tag{7}$$

where the loss is computed only on the assistant-side reasoning target rather than on the system or user prompt tokens. For the full EmoGPT model, we combine this generation loss with the AVR load-balancing term in Equation (3):

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{CoT}} + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}}, \qquad \lambda_{\mathrm{bal}} = 10^{-2}. \tag{8}$$

Only the LoRA adapters, AVR experts, and ESBias parameters are updated, while the pretrained perceptual encoders remain frozen; full hyperparameters are listed in Appendix C.1.

4 Experiments

We evaluate EmoGPT on benchmarks spanning English and Chinese, acted and in-the-wild settings: MER2023 (Lian et al., 2023), MER2024 (Lian et al., 2024), MELD (Poria et al., 2019),

Table 2: Main results on MER-UniBench (Lian et al., 2025) under three modality configurations, following the official evaluation protocol. Qwen-Omni (finetuned) denotes Qwen2.5-Omni-7B (Xu et al., 2025b) fine-tuned on ECoT-65K with LoRA only; EmoGPT further adds AVR and ESBias.

Mod.	Model	MER2023	MER2024	MELD	IEMOCAP	MOSI	MOSEI	SIMS	SIMSv2	OV-MERD+	Mean
AT	OneLLM (Han et al., 2024)	25.52	17.21	28.32	33.44	64.01	54.09	63.39	61.98	22.25	41.13
	SECap (Xu et al., 2024)	40.95	52.46	25.56	36.92	55.76	54.18	59.51	57.41	36.97	46.64
	Qwen-Audio (Chu et al., 2023)	41.85	31.61	49.09	35.47	70.09	46.90	70.73	65.26	32.36	49.26
	SALMONN (Tang et al., 2023)	55.53	45.38	45.62	46.84	81.00	67.03	68.69	65.93	45.00	57.89
	AffectGPT (Lian et al., 2025)	**72.94**	73.41	56.63	55.68	83.46	80.74	82.99	83.75	**59.98**	72.18
	Light-MER (Zheng et al., 2026)	59.33	58.30	53.10	52.18	80.48	79.95	80.74	79.80	53.75	66.40
	Qwen-Omni (finetuned)	66.32	70.15	56.04	59.73	87.05	85.48	82.31	82.94	54.36	71.60
	EmoGPT	69.68	**73.43**	**58.87**	**62.87**	**89.18**	**87.34**	**85.05**	**85.84**	54.75	**74.11**
VT	Otter (Li et al., 2025a)	16.41	14.65	22.57	29.08	52.89	50.44	57.56	53.12	16.63	34.82
	Video-LLaVA (Lin et al., 2024)	36.93	30.25	30.73	38.95	56.37	61.64	53.28	57.45	34.00	44.40
	Video-ChatGPT (Maaz et al., 2024)	44.86	46.80	37.33	56.83	54.42	63.12	64.82	65.80	39.80	52.64
	VideoChat2 (Li et al., 2024a)	33.67	54.50	36.64	48.70	66.84	54.32	69.49	70.66	39.21	52.67
	LLaMA-VID (Li et al., 2024b)	50.72	57.60	42.75	46.02	61.78	63.89	69.35	67.48	45.01	56.07
	VideoChat (Li et al., 2025b)	48.73	57.30	41.11	48.38	65.13	63.61	65.74	72.14	44.52	56.30
	Chat-UniVi (Jin et al., 2024)	57.62	65.67	45.61	52.37	54.53	63.18	68.15	66.36	48.00	57.94
	mPLUG-Owl (Ye et al., 2023)	56.86	59.89	49.11	55.54	72.40	72.91	72.13	75.00	48.18	62.45
	AffectGPT (Lian et al., 2025)	74.58	75.29	57.63	**62.19**	82.39	81.57	87.20	86.29	**61.65**	74.31
	Light-MER (Zheng et al., 2026)	58.77	56.31	51.51	47.38	82.39	79.36	82.80	81.32	51.51	65.71
	Qwen-Omni (finetuned)	73.06	75.84	56.18	58.27	86.74	85.13	86.59	85.86	56.48	73.80
	EmoGPT	**76.24**	**78.91**	**58.72**	61.19	**89.33**	**87.59**	**89.69**	**88.58**	58.37	**76.51**
AVT	Qwen2.5-Omni-7B (Xu et al., 2025b)	72.84	80.44	43.72	61.53	78.86	61.84	75.44	76.67	53.13	67.16
	HumanOmni-7B (Zhao et al., 2025)	63.36	76.64	37.05	52.52	7.37	7.33	43.87	35.35	59.03	42.50
	HumanOmniV2-9B (Yang et al., 2025)	68.59	57.41	42.74	60.95	76.59	75.44	83.58	81.65	61.62	67.62
	Emotion-LLaMA (Cheng et al., 2024)	59.38	73.62	46.76	55.47	66.13	67.66	78.32	77.23	52.97	64.17
	AffectGPT (Lian et al., 2025)	78.54	78.80	55.65	60.54	81.30	80.90	88.49	86.18	62.52	74.77
	Light-MER (Zheng et al., 2026)	75.66	80.04	57.02	61.46	80.04	77.23	**89.05**	87.74	**63.28**	74.61
	Qwen-Omni (finetuned)	74.67	80.32	58.74	65.42	86.75	84.29	84.21	84.53	59.77	75.41
	EmoGPT	**79.10**	**84.56**	**62.38**	**69.94**	**90.09**	**88.14**	88.92	**89.24**	61.13	**79.28**

IEMOCAP (Busso et al., 2008), CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), CH-SIMSv2 (Liu et al., 2022a), and OV-MERD+ (Lian et al., 2025). Following the MER-UniBench protocol of AffectGPT (Lian et al., 2025), we exclude *neutral* when computing accuracy and evaluate three input settings: audio–text (AT), video–text (VT), and audio–video–text (AVT). All baselines, spanning general-purpose multimodal LLMs and emotion-specialised MLLMs, are evaluated on the same test splits and prompt templates, and our system uses deterministic vLLM decoding at temperature zero.

4.1 Main Results

Table 2 reports per-dataset scores and their mean under AT, VT, and AVT settings. Fine-tuning Qwen2.5-Omni-7B on ECoT-65K alone raises its AVT mean from 67.16 to 75.41, exceeding the reported AffectGPT result by 0.64 points without architectural changes. Adding AVR and ESBias further increases the AVT mean to 79.28, while achieving 74.11 and 76.51 under AT and VT, respectively. Overall, EmoGPT surpasses AffectGPT by 4.51, 1.93, and 2.20 points under AVT, AT, and VT, respectively, and achieves the highest mean score among all compared systems under all three modality configurations. The gains are distributed across monologue sentiment, dialogue emotion, Chinese multimodal sentiment, and open-vocabulary evaluation rather than being driven by one benchmark. Under AVT, EmoGPT improves over AffectGPT by 6.73 points on MELD and 9.40 points on IEMOCAP, indicating particular benefits in dialogue-centric settings where pragmatic intent and conflicting multimodal cues are important. On MOSI and MOSEI, ECoT-65K fine-tuning already produces strong AVT scores of 86.75 and 84.29, which AVR and ESBias further refine to 90.09 and 88.14. This progression suggests that ECoT supervision establishes a strong reasoning-oriented foundation, while the two modules improve the processing and prioritization of affective evidence. On OV-MERD+, EmoGPT remains competitive despite not being trained specifically on open-vocabulary targets, suggesting transfer beyond closed-set labels. Its gains also persist under AT and VT, indicating that the improvements are not confined to full-modality input or driven by a single modality.

Table 3: **Human audit of ECoT-65K annotation quality.** F/P denotes Fully/Partially Supported. Broad Support is computed over assessable responses, excluding Unable responses. All values are percentages.

Dimension	F/P	Unsup.	Unable	Broad
Textual evidence	77.1 / 15.7	4.3	2.9	95.6
Visual evidence	68.6 / 25.7	5.7	0.0	94.3
Acoustic evidence	62.9 / 25.7	5.7	5.7	93.9
Cross-modal reasoning	55.7 / 34.3	4.3	5.7	95.5
Emotion judgment	60.0 / 30.0	5.7	4.3	94.0
Intent judgment	62.9 / 25.7	7.1	4.3	92.5
Evidence/reasoning avg.	64.5 / 26.2	5.5	3.8	**94.3**
Overall usability	55.7 / 30.0	8.6	5.7	90.9

Human audit of annotation quality. To complement model-based critic filtering, we randomly sample 70 retained ECoT annotations for human assessment. Each annotation is independently rated by three annotators drawn from a pool of 20, yielding 210 ratings per dimension. Annotators assess modality-specific evidence, cross-modal reasoning, emotion and intent judgments, and overall usability using four categories: Fully Supported, Partially Supported, Unsupported, and Unable to Assess. As shown in Table 3, the six evidence-and-reasoning dimensions achieve an aggregate Broad Support rate of 94.3% among assessable responses, while only 5.5% of all ratings are Unsupported. Overall annotation usability obtains 90.9% Broad Support, providing human evidence that the retained annotations are generally grounded in the corresponding multimodal inputs.

4.2 Dataset Comparison

ECoT-65K not only expands data volume but changes the supervision structure seen during fine-tuning. Table 1 and the Qwen-Omni (finetuned) row in Table 2 together show this point: with no architectural modification, fine-tuning on ECoT-65K alone brings Qwen2.5-Omni-7B to 75.41% AVT mean accuracy, already surpassing AffectGPT (Lian et al., 2025), the strongest emotion-specialised baseline, under a simpler fine-tuning recipe. This result demonstrates that structured cross-modal CoT supervision alone provides a strong advantage over caption-style supervision under a simpler fine-tuning recipe. The full EmoGPT model further shows that pairing this supervision with AVR and ESBias translates the supervision-level advantage into a larger model-level improvement. This supports the paper's central positioning. Our main contribution is not a heavy new backbone, but a supervision pipeline that makes a standard multimodal backbone far more competitive, plus two lightweight modules that unlock the remaining headroom.

4.3 Ablation Studies

We focus the main paper on two ablations that directly test the design of our method: module ablation for AVR and ESBias, and parameter ablation for the AVR routing design.

Module ablation. Table 4 shows that AVR and ESBias individually improve the CoT-only baseline, while their combination performs best across all three modality settings. AVR yields larger gains when visual tokens are available, consistent with its role in the visual projection pathway, whereas ESBias improves AT, VT, and AVT, consistent with its placement across all thinker attention layers. These results support their functional complementarity: AVR adapts visual representations before language-model reasoning, while ESBias recalibrates attention interactions and output representations throughout the thinker.

Table 4: Ablation of AVR and ESBias. The first row corresponds to Qwen-Omni (finetuned).

AVR	ESBias	AT Mean	VT Mean	AVT Mean	Avg	Δ
×	×	71.60	73.80	75.41	73.60	–
✓	×	72.24	75.67	77.34	75.08	+1.48
×	✓	73.42	74.83	76.47	74.91	+1.31
✓	✓	**74.11**	**76.51**	**79.28**	**76.63**	**+3.03**

AVR parameter ablation. Table 5 explores the number of experts M and routing sparsity k. Moving from two experts to four consistently improves performance, while increasing to eight experts no longer helps, suggesting that the main affective categories are already well covered by four specialists. At matched expert counts, top-$k = 2$ is better than top-$k = 1$, indicating that a token often benefits from mixing two specialized pathways rather than committing to only one. The final choice $M = 4,\ k = 2$ therefore provides the best efficiency, accuracy trade-off. Emotion recognition requires disentangling many recurrent visual factors rather than an open-ended set of object categories, so a compact expert pool is sufficient.

Table 5: AVR parameter ablation. $M = 1$ denotes the original single-MLP merger.

M	Top-k	Params	AVT Mean
1	—	44M	75.41
2	1	89M	75.48
2	2	89M	76.42
4	1	178M	77.91
4	2	178M	79.28
8	2	356M	78.79

4.4 Fine-grained Emotion Analysis

To complement the dataset-level accuracy reported in Table 2, we conduct a fine-grained per-class recall analysis on MELD. Table 6 reports per-class recall for five models under the AVT setting. EmoGPT achieves the highest UAR and WAR overall, reflecting broad and balanced recognition capability across emotion categories. Several patterns are worth noting. First, *disgust* is a universally difficult category: most models score below 12%, which we attribute to its low sample frequency (68 samples) and visual subtlety—disgust shares facial muscle activations with anger and contempt, making it hard to distinguish without explicit cross-modal grounding. HumanOmniV2 is an exception on *disgust* (30.88%), but this comes at the cost of markedly lower recall on *anger* (30.14%) and

Table 6: Per-class recall (%) on the MELD test set under the AVT setting. EmoGPT achieves the best UAR and WAR overall, with notable gains on minority classes (*Fea*, *Sad*). UAR and WAR denote unweighted and weighted average recall, respectively.

Method	Ang	Joy	Sad	Dis	Fea	Sur	UAR	WAR
Emotion-LLaMA (Cheng et al., 2024)	**58.55**	58.41	0.00	8.82	28.00	46.26	33.34	43.34
HumanOmni-7B (Zhao et al., 2025)	44.64	39.05	9.13	11.76	8.00	48.04	26.77	35.23
HumanOmniV2-9B (Yang et al., 2025)	30.14	82.34	12.50	**30.88**	17.20	44.13	36.20	45.39
AffectGPT (Lian et al., 2025)	54.03	72.64	41.54	8.33	45.60	52.14	45.71	54.64
EmoGPT	56.52	**83.63**	**48.75**	26.47	**57.60**	**52.67**	**54.27**	**61.11**

fear (17.20%), suggesting that its predictions may be biased toward disgust-like facial patterns at the expense of other categories. In contrast, EmoGPT maintains strong recall across all three of these categories (56.52%, 26.47%, and 57.60% for *anger*, *disgust*, and *fear* respectively), suggesting that structured cross-modal CoT supervision encourages a more balanced use of visual, acoustic, and textual evidence rather than over-relying on any single modality signal.

4.5 Case Study

Figure 5 presents a conflict-heavy scene from *Friends*: superficially positive wording conflicts with high-arousal, directive prosody, while the Roman-soldier costume acts as a visual distractor. AffectGPT and Emotion-LLaMA produce plausible descriptions but do not explicitly reconcile or prioritize the conflicting cues, resulting in excitement- or anxiety-like interpretations. In contrast, EmoGPT separates textual, visual, and acoustic evidence, treats the costume as non-expressive, grounds anger in AU4/AU7 and emphatic stress, and reaches the correct judgment through the structured reasoning schema learned during training. This example is consistent with the larger gains on dialogue-rich benchmarks in Table 2, where conflicting contextual and paralinguistic cues are common. It also illustrates the contribution of ECoT-65K beyond data volume: explicit supervision for identifying distractors, reconciling conflicting evidence, and grounding final decisions. ECoT-65K targets evidence externalization and conflict resolution, while AVR and ESBias respectively refine visual projection and emotion-sensitive evidence prioritization.



Transcript: "Well, I'm totally over it Chandler. Friends forever! Don't come out here!" Ground Truth: **Anger**

AffectGPT

In the text, caption content: "Alright, I'm all done, Chandler Forever Friends, don't come out here." Based on the character's facial expressions, body language, and surroundings in the video clues, as well as the individual's high-pitched, excited tone in the audio clues, we can infer that this phrase might be an expression of their excitement and anxiety. The character's facial expressions and body language indicate that they are experiencing a strong emotional state, possibly due to something unexpected or urgent happening. The individual's high-pitched, excited tone and the mention of "Forever Friends" and "don't come out here" in the caption content suggest that this might be a scenario involving a competition or challenge. The person may be a participant or competitor in a thrilling event or challenge, feeling excited and anxious about the outcome.

Emotion-LLaMA

The person in the video is saying "I'm totally over it Chandler" which suggests that they are feeling a strong sense of frustration or annoyance towards something. The person is also saying "Don't come out here!" which could be an indication of fear or anxiety about something. The tone of the person's voice is quite high-pitched and trembling, which is consistent with feelings of anxiety or nervousness. The person's use of the phrase "Friends forever!" suggests a strong sense of loyalty and commitment to someone or something, which could be a source of anxiety or stress. Overall, based on the clues provided in the video, it seems likely that the person is feeling anxious about a particular situation or relationship

EmoGPT

Textual: The phrase "totally over it" establishes a negative polarity, while "Don't come out here!" conveys an explicit rejection command. Visual: **AU4 (Brow Lowerer) and AU7 (Lid Tightener)** indicate anger and focus; rapid head movement and pointing gesture reinforce the commanding tone. The Roman soldier costume is identified as non-expressive artifact and filtered out. Acoustic: Fast speech rate with high intensity; **emphatic stress on "totally" and "Don't" with falling-command intonation**, indicating high-arousal negative affect. Prediction: All three modalities converge on **Anger** (High intensity, High certainty). Key anchors: "totally over it", "Don't come out here!", AU4/AU7, emphatic stress on "Don't".

Figure 5: EmoGPT resolves a conflict-heavy clip by separating textual, visual, and acoustic evidence, while AffectGPT and Emotion-LLaMA are misled by surface cues.

5 Conclusion

We present EmoGPT as a structured cross-modal reasoning framework for MER, built around the ECoT-65K dataset and two lightweight architectural modules, AVR and ESBias, on top of a standard multimodal backbone. The experiments show that structured cross-modal CoT supervision is a strong signal on its own, and that AVR and ESBias further unlock the remaining performance headroom, lifting the AVT mean accuracy to 79.28%—the highest mean score under all three modality settings on MER-UniBench. We hope this work moves MER toward evidence-grounded reasoning, while future work can extend the idea to other backbones and larger datasets.

Limitations. ECoT-65K annotations are MLLM-generated and critic-audited; the human audit in Section 4.1 supports the overall quality of the retained traces, while residual noise may remain for ambiguous, low-intensity, or culturally dependent samples.

AI USE STATEMENT

Generative AI tools were used in several disclosed and scoped components of this work. During ECoT-65K construction, `gemini-3-flash-preview` served as the Generator to produce candidate structured annotations from video, audio, and transcript inputs under author-designed prompts, while `DeepSeek-V3` served as the Critic to identify inconsistent or label-misaligned annotations. Following the MER-UniBench protocol, Qwen2.5 was used as a post-hoc label extractor when free-form model outputs needed to be mapped to discrete benchmark labels. General-purpose AI assistants were also used for limited language polishing, LaTeX editing, and code review. All prompts, filtering criteria, model designs, experiments, statistical calculations, and scientific claims were designed or independently verified by the authors. AI-generated code was inspected and tested before use, and no generative AI tool was used to fabricate experimental results, citations, or conclusions. The authors take full responsibility for all content in this submission.

REPRODUCIBILITY STATEMENT

We provide the ECoT-65K construction procedure, source-corpus statistics, annotation prompts, filtering criteria, and annotation examples in Appendix B. The implementation details of AVR and ESBias are specified in Section 3, while the training configuration, optimization hyperparameters, hardware environment, and evaluation prompts are documented in Appendices C and D. All reported evaluations follow fixed benchmark splits and deterministic decoding. We will publicly release the code and ECoT-65K dataset upon publication, subject to the licenses and usage terms of the original source corpora.

REFERENCES

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

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

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

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–13710, 2024.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024b.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9610–9614, 2023.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–48, 2024.

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

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*, 2024.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 247–258, 2022a.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 24–32, 2022b.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

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, 2019.

James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161, 1980.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Haoqin Sun, Xuechen Wang, Jinghua Zhao, Shiwan Zhao, Jiaming Zhou, Hui Wang, Jiabei He, Aobo Kong, Xi Yang, Yequan Wang, et al. Emotiontalk: An interactive chinese multimodal emotion dataset with rich annotations. *arXiv preprint arXiv:2505.23018*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2019.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotion-thinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098, 2025a.

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*, 2025b.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. SECap: Speech emotion captioning with large language model. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 19323–19331. AAAI Press, 2024. doi: 10.1609/aaai.v38i17.29902.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1103–1114, 2017.

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, 2018.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

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? In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026. doi: 10.1145/3767308.3835072. URL https://doi.org/10.1145/3767308.3835072.

A THE USE OF LARGE LANGUAGE MODELS

We use large language models in three scoped parts of the project. First, during dataset construction, `gemini-3-flash-preview` is used as the *Generator* to produce candidate structured cross-modal CoT annotations from video, audio, and transcript inputs under author-designed prompts. Second, during quality control, `DeepSeek-V3` is used as the *Critic* in a Generator–Critic audit pipeline to examine Gemini-generated annotations for internal logical conflicts and label-alignment mismatches. We summarize the outcomes as four categories—*Valid*, *Gen. Error*, *Ambig. Data*, and *Dual Conf.*—and retain only *Valid* cases in ECoT-65K. Third, for evaluating generative model outputs on benchmarks that require discrete labels, we follow AffectGPT and use Qwen2.5 as a post-hoc label extractor to map free-form emotion descriptions to fixed benchmark labels when needed.

In addition, general-purpose LLM assistants were used for limited language polishing and LaTeX editing of the manuscript under author supervision. All research decisions—including problem formulation, model design, prompt design, dataset selection, filtering criteria, implementation, experiment execution, metric computation, result verification, and scientific claims—were made and checked by the authors. No LLM was used to fabricate experimental results, citations, or conclusions.

B ECOT-65K: ADDITIONAL DETAILS

B.1 SOURCE CORPORA

ECoT-65K is constructed from seven public corpora spanning English and Mandarin, monologic opinion videos and multi-party conversations, as well as both in-the-wild and actor-performed settings. Below we briefly summarise the role of each source corpus in the released 6w5 training split. Table 7 distinguishes *original corpus statistics* reported in the source papers from the number of samples that are actually retained in ECoT-65K after our preprocessing and filtering.

Table 7: Summary of source corpora used to construct ECoT-65K. **Original corpus statistics** are taken from the source papers. **Used in ECoT-65K** reports the number of samples retained in our released 6w5 training split after our preprocessing and filtering.

Source Corpus	Original Corpus Statistics	Lang.	Annotation Schema	Canonical Label Space	Collection Domain	Used in ECoT-65K
CMU-MOSI	2,199 opinion segments	EN	sentiment intensity + subjectivity	sentiment in [-3, 3]	YouTube opinion vlogs	1,441
CMU-MOSEI	23,453 utterance-level segments	EN	sentiment intensity + emotion tags	sentiment in [-3, 3] + 9 emotions	YouTube opinion videos	13,803
MELD	∼13K utterances / 1,433 dialogues	EN	utterance-level emotion + sentiment	7 emotions / 3 sentiments	TV-sitcom dialogues	5,799
MER2024	5,030 labeled Train&Val clips	ZH	categorical emotion recognition	6 emotions	movie and TV clips	3,678
CH-SIMSv2	4,402 supervised clips (+10,161 unlabeled)	ZH	multimodal + unimodal sentiment intensity	7-point sentiment scale	11 real-world scenarios	2,947
MC-EIU	56,012 utterances / 4,970 conversational clips	EN+ZH	joint emotion-intent annotation	7 emotions / 9 intents	bilingual TV-series dialogues	26,531
EmotionTalk	19,250 utterances / 744 dialogues	ZH	emotion, sentiment, and style captioning	7 emotions / 5 sentiments / 4 caption axes	acted dyadic dialogues	11,171

CMU-MOSI. CMU-MOSI Zadeh et al. (2016) is an English opinion-video benchmark collected from online sharing platforms such as YouTube. Its original annotations focus on sentiment intensity over a continuous range, making it a representative source of short, speaker-centric sentiment clips where lexical content and facial or prosodic nuance jointly determine polarity. In ECoT-65K, MOSI contributes 1,441 training-side non-neutral clips, each converted into one structured cross-modal CoT sample.

CMU-MOSEI. CMU-MOSEI Zadeh et al. (2018) extends the MOSI setting to a much larger and more diverse English corpus of opinion videos from online media, with sentiment and emotion annotations over a wide range of topics and speakers. We use it as the main English in-the-wild monologue source in our training mixture. After preprocessing and selection, MOSEI contributes 13,803 samples to ECoT-65K.

Table 8: Source-corpus composition of the proposed ECoT-65K. **Share** denotes the percentage of samples contributed by each source corpus out of the full 65,370-sample training set.

Source Corpus	# Samples	Share (%)
MC-EIU	26,531	40.59
MOSEI	13,803	21.12
EmotionTalk	11,171	17.09
MELD	5,799	8.87
MER2024	3,678	5.63
CH-SIMSv2	2,947	4.51
MOSI	1,441	2.20
Total	**65,370**	**100.00**

MELD. MELD Poria et al. (2019) is a multimodal multi-party dialogue dataset derived from the TV series *Friends*. Each utterance is aligned with text, audio, and visual streams and annotated with discrete emotion and sentiment labels, making it especially valuable for context-rich conversational affect recognition. We use MELD to inject conversational, speaker-interactive emotion cases into our CoT data, yielding 5,799 samples in the released training split.

MER2024. MER2024 Lian et al. (2024) is a Chinese multimodal emotion benchmark built from movie and TV clips and organised around semi-supervised learning, noise robustness, and open-vocabulary emotion recognition. Relative to internet opinion-video datasets, it supplies more difficult in-the-wild Chinese affective clips with diverse recording conditions. We use its labeled emotion clips as part of the Chinese supervision pool, contributing 3,678 samples to ECoT-65K.

CH-SIMSv2. CH-SIMSv2 Liu et al. (2022a) is a Chinese multimodal sentiment dataset collected from eleven everyday media scenarios and annotated with both multimodal and unimodal sentiment labels. It is particularly useful because it emphasises the contribution of non-verbal cues rather than treating audio and visual signals as minor auxiliaries to text. In our corpus, CH-SIMSv2 contributes 2,947 sentiment-bearing clips that are rewritten into the unified CoT format.

MC-EIU. MC-EIU Liu et al. (2024) is a bilingual multimodal conversational benchmark built from English and Mandarin TV-series clips, with utterance-level annotations for both emotion and intent. This source is especially important for EmoGPT because it already couples affective state and communicative purpose inside multi-turn dialogue, which aligns naturally with our reasoning targets. MC-EIU is the largest single contributor in ECoT-65K, providing 26,531 training samples.

EmotionTalk. EmotionTalk Sun et al. (2025) is an interactive Chinese multimodal emotion dataset recorded by professional actors in dyadic dialogue settings. Besides seven discrete emotion categories, it also provides five-level sentiment labels and richer descriptive annotations such as speaking-style captions. We use EmotionTalk as a high-quality acted Chinese source with clean audio-visual signals and natural dialogue structure, from which 11,171 samples enter ECoT-65K.

B.2 Sample Statistics

We report here the statistics of the released 6w5 training set used to train EmoGPT, corresponding to `train_data_v2_all8_audio_sharegpt.json`. The corpus contains 65,370 samples in total. Each sample is stored in ShareGPT format with exactly three messages (`system`, `user`, `assistant`), one paired video path, and one paired audio path. The assistant target is a structured JSON reasoning trace containing `modality_extraction`, `inference_logic`, and `final_judgment`. Here, *coarse polarity* denotes a post-hoc normalisation of the free-form `mapping_hint` field into a small set of polarity buckets for visualisation only; it is not the original official label space of the source corpora.

Table 9: Format-level statistics of the 6w5 training set. Every sample includes synchronised video and audio inputs and one structured CoT supervision target.

Field	Value
Total samples	65,370
Messages per sample	3
Video paths per sample	1
Audio paths per sample	1
Target format	Structured JSON CoT
Target blocks	`modality_extraction`, `inference_logic`, `final_judgment`

B.3 Full Annotation Prompt

Following AffectGPT, we place the annotation prompt in Appendix for reproducibility. In the same spirit as AffectGPT's prompt appendix, we report the exact prompting template used in our pipeline; our deployed version extends this style into direct structured cross-modal CoT generation. Below we provide an English translation of the instruction template used in our local data-construction pipeline; the original deployed prompt asks the generator to answer in Chinese.

System prompt (English translation).

Role and objective. You are an expert in multimodal behaviour and communicative-intent analysis. Your task is to infer high-fidelity, logically self-consistent emotion and intent labels from the video's language, facial Action Units (AUs), and acoustic cues, and return the result as a JSON object.

Core strategies and decision priority. (1) *Semantic precedence*: if the transcript contains explicit evaluative words (e.g., *goofy*, *weird*, *chaotic*, *amazing*, *smart*, *terrible*, *low quality*), infer the intent from semantics even if visual or acoustic evidence is neutral; do not collapse such cases into *Technical Fact*. (2) *Conflict resolution order*: when semantic polarity conflicts with visual or acoustic signals, resolve the final judgement in the order *Textual Sentiment* $>$ *Acoustic Prosody* $>$ *Facial AU*. Explicit verbal evaluation is treated as the most reliable indicator of user intent when physiological cues are weak or noisy. (3) *Laughter and mockery*: laughter or AU12 co-occurring with a negative target evaluation indicates *Mockery/Ridicule*; laughter with a positive evaluation indicates *Amusement/Enjoyment*, unless stronger cues such as eye-roll behaviour or AU15 suggest irony. (4) *Neutrality*: only purely factual statements with no subjective modifier may be labelled neutral; any subjective modifier disqualifies neutrality.

Five-stage procedure. Stage 1: extract the evaluation target, sentiment-bearing words, and overall textual polarity. Stage 2: identify salient facial AUs together with rough intensity and temporal cues. Stage 3: extract acoustic facts, including F_0 variation, stress position and strength, speech rate, and abrupt intonation changes. Stage 4: fuse the three modalities according to the priority rules and derive the final judgement. Stage 5: inspect whether the reasoning chain depends on the reference label; if so, remove such wording and rewrite the final JSON using only the video, audio, and transcript.

Grounding and safety rules. (1) *Mockery alert*: if laughter or smiling co-occurs with negative lexical content or explicit criticism, consider *Mockery/Ridicule*. (2) *High-arousal negative*: fast speech and high pitch do not imply excitement when the content is complaint, chaos, profanity, or frustration. (3) *Comparative trap*: comparisons with canonical works (e.g., *Inception*, *Avatar*) without explicit praise are usually implicit criticism for lack of originality. (4) Remain faithful to directly observed evidence; uncertain evidence must lower `evidence_certainty`. (5) The provided reference label may be used only for final consistency checking and must never appear in the reasoning chain. (6) Separate observation from inference: `visual_evidence.confirmed_aus` records only directly observed AUs or actions; `acoustic_evidence` records only audible facts; `final_judgment.key_anchors` may cite only transcript spans and evidence already present in the observation fields. (7) Do not hallucinate missing context or repair corrupted transcripts. If the transcript is truncated or noisy, explicitly state that it is incomplete and reason conservatively. (8) Surface emotion and final polarity may differ; for example, concern may still support positive appraisal, while smiling may still support mockery. (9) Assign *Mockery/Ridicule* or *Humorous Irony* only when at least two modalities independently support the sarcastic reading. (10) When evidence is weak or conflict-

ing, prefer conservative explanations and avoid high-confidence psychological speculation. (11) Before output, ensure `primary_emotion`, `intent_label`, `mapping_hint`, `intent_derivation`, and `rule_application` are mutually consistent, and make `mapping_hint` explicitly state the final binary polarity (e.g., "overall polarity = Positive").

Required output format. Return a JSON object with three top-level fields: `modality_extraction`, `inference_logic`, and `final_judgment`. The required subfields are `textual_features`, `visual_evidence.confirmed_aus`, `visual_evidence.artifact_filter`, `acoustic_evidence`, `consistency_audit`, `intent_derivation`, `rule_application`, `primary_emotion`, `intent_label`, `mapping_hint`, `intensity`, `evidence_certainty`, and `key_anchors`.

Sample-specific user prompt template (English translation).

Video transcript: {`transcript`}
Reference label (for final consistency checking only; do not use it as reasoning evidence and do not mention it in the output): {`ground_truth`}
First complete an independent analysis using only the video, audio, and transcript, and then perform a consistency self-check. If your independent judgement disagrees with the reference label, remain faithful to the evidence: do not rewrite evidence, repair truncated semantics, hallucinate missing context, or force a sarcasm/mockery explanation merely to match the reference label.

B.4 Annotation Examples

We provide three representative examples to illustrate how the released annotations cover aligned evidence, sarcasm-style cross-modal conflict, and incomplete-transcript cases. All examples are taken from the released 6w5 training set and keep the original structured JSON format used for supervision.

Example 1: Aligned positive evidence (MOSI). Transcript: `ANYHOW IT WAS REALLY GOOD`


```
{
  "modality_extraction": {
    "textual_features": "The transcript contains the explicit positive evaluation phrase
    'really good', with the degree adverb 'really' adding emphasis; 'Anyhow' signals a
    summarizing or wrap-up remark.",
    "visual_evidence": {
      "confirmed_aus": "A mild AU12 (lip-corner puller) and AU6 (cheek raiser) are observed,
      forming a natural smile; the speaker looks toward the camera and appears relaxed.",
      "artifact_filter": "Lighting is slightly dim, but the face remains clear; no occlusion
      or artifact is present that would interfere with expression judgment."
    },
    "acoustic_evidence": "The prosody is steady and mildly resonant, with clear emphasis on
    'really' and 'good'; speaking rate is moderate and the delivery sounds sincere."
  },
  "inference_logic": {
    "consistency_audit": "Textual semantics (positive evaluation), visual evidence (natural
    smile), and acoustic evidence (sincere emphasis) are highly consistent, with no
    cross-modal conflict.",
    "intent_derivation": "The speaker gives an affirmative evaluation of an object or
    experience in a summary-like tone; together with the facial micro-expression, this
    supports an intent of sincere praise or satisfied feedback.",
    "rule_application": "Following semantic-precedence, 'really good' establishes a positive
    base judgment; the visual and acoustic evidence further reinforces this reading and rules
    out sarcasm."
  },
  "final_judgment": {
    "primary_emotion": "Satisfaction",
    "intent_label": "Sincere Praise",
    "mapping_hint": "overall polarity = Positive",
    "intensity": "Moderate",
    "evidence_certainty": "High",
    "key_anchors": [
      "really good",
      "AU12 (subtle smile)",
      "Sincere prosody"
    ]
  }
}
```

Example 2: Sarcasm-style cross-modal conflict (MOSI). Transcript: `THANKS A LOT COMMERCIAL`

```
{
  "modality_extraction": {
    "textual_features": "The utterance 'THANKS A LOT COMMERCIAL' is literally a form of
    thanks, but in context it is commonly used to express annoyance at commercial
    interruption.",
    "visual_evidence": {
      "confirmed_aus": "Clear AU4 (brow lowerer) and AU7 (lid tightener) are observed,
      indicating displeasure and tension; the mouth also shows mild AU10 (upper-lip raiser),
      suggesting disgust or contempt.",
      "artifact_filter": "Lighting is even, facial details are clear, and there is no
      occlusion; the expression signal is highly reliable."
    },
    "acoustic_evidence": "The delivery has clear sarcastic prosody, with strong emphasis on
    'THANKS' and 'LOT'; the tone is stiff and aggressive rather than genuinely appreciative."
  },
  "inference_logic": {
    "consistency_audit": "Literal semantics (thanks) conflict in polarity with the acoustic
    evidence (sarcastic tone) and the visual evidence (anger/disgust expression).",
    "intent_derivation": "Although the surface wording is gratitude, the strong sarcastic
    delivery and the frowning, disgust-like facial cues indicate a sarcastic complaint about
    the commercial interruption.",
    "rule_application": "Applying the sarcasm/mockery criterion, both the acoustic and visual
    modalities independently support a sarcastic reading, so the final judgment is
    Mockery/Ridicule."
  },
  "final_judgment": {
    "primary_emotion": "Frustration/Annoyance",
    "intent_label": "Mockery/Ridicule",
    "mapping_hint": "overall polarity = Negative",
    "intensity": "Moderate",
    "evidence_certainty": "High",
    "key_anchors": [
      "THANKS A LOT",
      "AU4 (Brow Lowerer)",
      "Sarcastic tone on 'THANKS'"
    ]
  }
}
```

Example 3: Incomplete transcript resolved by audio-visual evidence (EmotionTalk). Transcript: `Well, I ho[over/]pe[interrupted][/over]...`

```
{
  "modality_extraction": {
    "textual_features": "The transcript is a fragmentary utterance, roughly 'Well, I hope...',
    with incomplete semantics; it shows hesitation, pausing, and interrupted expression,
    suggesting that the speaker is emotionally overwhelmed or psychologically blocked.",
    "visual_evidence": {
      "confirmed_aus": "Clear AU1+4 (inner brow raiser plus brow lowerer, forming a
      worried-brow pattern), AU15 (lip-corner depressor), and AU17 (chin raiser) are observed;
      the eyes appear moist, gaze is lowered, and the face shows a typical expression of
      sadness and hurt.",
      "artifact_filter": "Lighting is soft, facial details are clear, and there is no
      occlusion; the AU evidence is both strong and internally consistent."
    },
    "acoustic_evidence": "Speech is extremely slow and marked by vocal tremor; volume is weak,
    breathing sounds heavy and intermittent, and the overall delivery conveys clear
    choking-back-tears behavior."
  },
  "inference_logic": {
    "consistency_audit": "Although the transcript is semantically incomplete and close to
    neutral on its own, the visual evidence (sad facial expression) and the acoustic evidence
    (choked, trembling delivery) are highly consistent and both indicate high-arousal negative
    sadness.",
    "intent_derivation": "The speaker appears to be trying to express a wish or request, but
    the utterance breaks off because of strong emotional fluctuation (sadness and hurt); the
    core intent is a vulnerable expression of distress or an appeal for sympathy.",
    "rule_application": "Under the modality-priority rule, when semantics are missing or
    neutral, strong visual AU evidence and acoustic evidence dominate the judgment. The
    AU1+4+15 combination is a canonical physiological pattern for sadness."
  },
  "final_judgment": {
    "primary_emotion": "Sadness/Distress",
    "intent_label": "Expressing Vulnerability",
    "mapping_hint": "overall polarity = Negative (Sadness)",
    "intensity": "High",
    "evidence_certainty": "High",
```

```
    "key_anchors": [
      "AU1+4 worried brow",
      "choked and trembling voice",
      "fragmented 'I hope'"
    ]
  }
}
```

B.5 Human Audit Protocol

We randomly sample 70 retained annotations from ECoT-65K. Each annotation is independently evaluated by three annotators drawn from a pool of 20, resulting in 210 ratings per dimension. Annotators inspect the corresponding multimodal sample together with its structured ECoT annotation and evaluate seven dimensions: textual evidence, visual evidence, acoustic evidence, cross-modal reasoning, emotion judgment, intent judgment, and overall annotation usability.

For each dimension, annotators select one of four categories:

- **Fully Supported**: the central claims are clearly supported by the available multimodal evidence;
- **Partially Supported**: the main interpretation is plausible, but some evidence is incomplete, ambiguous, or overly specific;
- **Unsupported**: one or more central claims conflict with, or cannot be grounded in, the available input;
- **Unable to Assess**: the available input quality or information is insufficient for a reliable judgment.

Broad Support measures the proportion of Fully or Partially Supported responses among assessable ratings:

$$\text{BroadSupport} = \frac{N_{\text{Fully}} + N_{\text{Partially}}}{N_{\text{Fully}} + N_{\text{Partially}} + N_{\text{Unsupported}}} \times 100\%. \tag{9}$$

Unable-to-assess responses are reported separately and excluded from the Broad Support denominator.

C Training Details

C.1 Full Training Configuration

We train EmoGPT with LlamaFactory in a supervised fine-tuning (SFT) setting with LoRA. Table 10 summarises the configuration used in `train_omni_6w5_v2.yaml`.

Two code-level modifications are applied before training. **ESBias** inserts a dedicated zero-initialized Q/K/V/O bias submodule into each of the 28 thinker attention layers and registers these modules through `additional_target: esbias`. **AVR** replaces the default visual patch merger with a four-expert top-2-routed MoE projector. The standard PEFT bias option remains disabled so that no unrelated pre-existing bias parameters are updated.

C.2 Hardware and Wall-clock Time

Training is conducted with NVIDIA RTX 4090 GPUs with 48 GB memory. The main fine-tuning runs use four GPUs, while some evaluation jobs are carried out on a single RTX 4090. The software stack consists of Python 3.11, Transformers 5.2.0, and LlamaFactory, with bf16 mixed precision, FlashAttention-2, gradient checkpointing, and DeepSpeed ZeRO-2 enabled.

D Evaluation Details

D.1 Evaluation Prompts

For closed-set benchmark evaluation, we use a fixed prompt family aligned with the training template. The same `system` prompt is shared across AVT, AT, and VT settings, while the `user` prompt

Table 10: Training configuration of EmoGPT. Paths in the original YAML are omitted; only paper-relevant settings are reported.

Category	Hyperparameter	Value
Backbone	Base model	Qwen2.5-Omni-7B
Backbone	Fine-tuning stage	SFT
Backbone	Fine-tuning type	LoRA
LoRA	Rank	16
LoRA	Alpha	32
LoRA	Dropout	0.05
LoRA	Target modules	all
LoRA	Bias	none
ESBias	Placement	Thinker attention layers
ESBias	Trainable vectors	Q/K/V/O
ESBias	Initialization	Zero
ESBias	Additional target	`esbias`
ESBias	Trainable parameters	229,376
Input processing	Template	`qwen2_omni`
Input processing	Cutoff length	8192
Input processing	Image max pixels	262,144
Input processing	Video max pixels	65,536
Input processing	Video max length	128 frames
Input processing	Audio-in-video	enabled
Input processing	Audio sampling rate	16 kHz
Dataset	Training set	`emotion_omni_v2_all8_audio`
Dataset	Preprocessing workers	8
Dataset	Dataloader workers	4
Dataset	Preprocessing batch size	10
Optimisation	Per-device batch size	1
Optimisation	Gradient accumulation steps	8
Optimisation	Learning rate	3.0×10^{-5}
Optimisation	Training epochs	3.0
Optimisation	LR scheduler	cosine
Optimisation	Warmup ratio	0.05
Optimisation	Precision	bf16
Optimisation	Vision tower	frozen
Memory/runtime	Gradient checkpointing	enabled
Memory/runtime	Attention kernel	FlashAttention-2
Memory/runtime	Distributed training	DeepSpeed ZeRO-2

is adapted to the available modalities. In all cases, decoding is performed with `temperature`=0 and `max_tokens`=1024.

Shared system prompt (English translation).

```
You are an expert in multimodal emotion and intent analysis.
Analyze by jointly considering the language in the video,
visual AUs, and acoustic features.
Decision priority:  semantic tendency > acoustic features >
visual AUs; laughter + negative semantics should be judged
as Mockery.
```

AVT prompt. For the standard trimodal setting, the `user` message contains a paired video input, a paired audio input, and the transcript text:

```
<video>
<audio>
```

```
Transcript: {transcript}
Please analyze the speaker's emotion and intent based on the
linguistic, visual, and acoustic cues in the video.
```

AT prompt. For the missing-visual setting, the video input is removed and the `user` instruction is changed accordingly:

```
<audio>
Transcript: {transcript}
Only the audio and transcript are currently available; no
video frames are provided.
Please analyze the speaker's emotion and intent based only
on vocal prosody and textual semantics.
```

VT prompt. For the missing-audio setting, the audio input is removed and the `user` instruction is changed accordingly:

```
<video>
Transcript: {transcript}
Only the video frames and transcript are currently
available; no audio is provided.
Please analyze the speaker's emotion and intent based only
on visual AU features and textual semantics.
```

The above prompts are English translations of the exact task instructions used in our vLLM-based evaluation scripts. Operationally, AVT evaluation sends a multimodal message with `video_url`, `audio_url`, and transcript text; AT removes the visual channel; VT removes the acoustic channel.