OMNIOPSD: RATIONALE-PRIVILEGED ON-POLICY SELF-DISTILLATION FOR AFFECTIVE COMPUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have advanced affective computing by enabling nuanced reasoning over visual, acoustic, and linguistic evidence. A key approach to improving their generalization is outcome-based reinforcement learning, which converts existing label-level annotations into rewards for final-answer correctness. However, its sparse sequence-level supervision provides limited fine-grained guidance for interpreting multimodal affective cues and may reward correct labels even when those cues are misinterpreted. To address this challenge, we introduce **OmniOPSD**, the first on-policy self-distillation framework for omni-modal affective reasoning. At its core, OmniOPSD transforms label-level supervision into evidence-rich, teacher-only rationales by leveraging reference affective judgments as semantic priors for offline multimodal evidence elicitation from frontier MLLMs. Through rationale-privileged self-distillation, it couples on-policy student rollouts with dense, evidence-informed token-level supervision from a local self-teacher, allowing the student to learn along its own trajectories rather than imitate fixed reasoning traces. This decoupling of offline evidence acquisition from local policy learning enables training without additional human annotation, frontier-model logits, or online frontier-model calls, while preserving student-only inference from the original task input. At both 3B and 7B scales, OmniOPSD outperforms supervised fine-tuning (SFT) and group relative policy optimization (GRPO) in mean macro-F1 on both in-domain and out-of-domain evaluations. Among the compared models, it ranks first by MER-UniBench mean across all three input-modality settings, reaching 84.13 with audio–video–text input, and achieves the highest MME-Emotion recognition score (Rec-S) of 42.4.

1 INTRODUCTION

Affective computing advances human-centered artificial intelligence by modeling human emotion, sentiment, and intent, with applications spanning education, mental healthcare, and human–computer interaction (Poria et al., 2017). Inferring these states requires interpreting facial expressions, vocal prosody, and language whose affective meaning is often ambiguous and context dependent (Barrett et al., 2019). Multimodal large language models (MLLMs) extend affective computing beyond fixed-category prediction toward richer reasoning (Lian et al., 2025a;b). This shift calls for models that not only predict affective states accurately but also explain how multimodal evidence supports their judgments, promoting transparent and interpretable inference (Zhang et al., 2026a).

Yet multimodal affective data resources primarily provide labels, with limited human-authored evidence annotations (Busso et al., 2008; Zadeh et al., 2018; Poria et al., 2019; Zhang et al., 2024). Acquiring such annotations requires labor-intensive interpretation of heterogeneous, context-dependent cues, making high-quality reasoning supervision difficult to scale (Cheng et al., 2024a; Lin et al., 2024b; Lian et al., 2025a; Zhang et al., 2026b). The practical question is therefore how to turn existing label-level annotations into reasoning-level supervision for affective MLLMs.

A natural response is to use reference judgments to construct outcome rewards for reinforcement learning with verifiable rewards (RLVR), often optimized with group relative policy optimization (GRPO) (Shao et al., 2024; Lian et al., 2025d; Wu et al., 2026; WANG et al., 2026). This enables on-policy learning from student-generated responses without prescribing reference rationales. However, final-answer and format rewards provide only sequence-level feedback: responses are rewarded for

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.

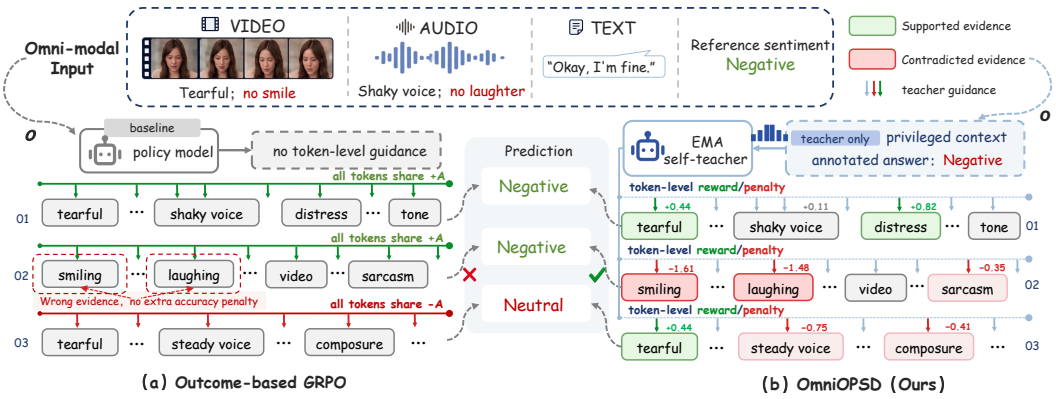


Figure 1: Comparison of (a) outcome-based GRPO and (b) OmniOPSD. GRPO assigns the same sequence-level advantage to all tokens, including those in erroneous reasoning. OmniOPSD provides dense, differentiated token-level supervision from a rationale-privileged self-teacher.

correct final answers even when their explanations misinterpret or hallucinate supporting cues (Han et al., 2026b). Explicit reasoning does not necessarily improve affective prediction (Lian et al., 2025d), highlighting the need for evidence-informed guidance along student-generated trajectories.

On-policy distillation (OPD) provides dense token-level supervision along student-generated trajectories (Agarwal et al., 2024; Lu & Lab, 2025). Distillation from frontier teachers faces limited logit access and tokenizer mismatches (Boizard et al., 2025), while logit-free alternatives require online teacher queries (Zhou et al., 2026). On-policy self-distillation (OPSD) uses a local self-teacher whose informational advantage comes from feedback or reference solutions (Hübotter et al., 2026; Zhao et al., 2026a), or asymmetric access to original and masked or degraded images (Li et al., 2026). Extending these designs to omni-modal affective reasoning, however, is nontrivial: labels specify outcomes rather than their evidential basis, while richer perceptual access alone leaves implicit how ambiguous, potentially conflicting cross-modal cues jointly support affective judgments.

To bridge this gap, we introduce **OmniOPSD**, a **Rationale-Privileged On-Policy Self-Distillation** framework for omni-modal affective reasoning (Figure 1). Our central insight is that reference affective judgments can do more than evaluate outcomes: they can anchor the acquisition of multimodal evidence that informs learning. OmniOPSD leverages this insight to construct teacher-only rationales from offline frontier-elicited evidence and reference judgments, recasting rationales as a basis for constructing supervision rather than fixed targets for student imitation. A local self-teacher translates this privileged context into dense token-level guidance tailored to the student's own evolving trajectories, turning offline evidence into trajectory-dependent supervision. By decoupling external evidence acquisition from local policy learning, OmniOPSD bridges label-level annotation and evidence-informed on-policy learning without additional human annotation, frontier-model logits, or online frontier-model calls, while retaining student-only inference from the original task input.

Our contributions are threefold:

1. We introduce OmniOPSD, which leverages reference judgments as semantic priors to elicit multimodal evidence descriptions from frontier MLLMs offline, combining both into teacher-only privileged context, not imitation targets. To our knowledge, OmniOPSD is the first on-policy self-distillation framework for omni-modal affective reasoning.

2. We formulate rationale-privileged on-policy self-distillation, coupling on-policy student rollouts with dense token-level supervision from an EMA self-teacher conditioned on privileged context. This requires neither frontier-model logits nor additional human annotation.

3. At 3B and 7B scales, OmniOPSD outperforms supervised fine-tuning (SFT) and GRPO in mean in-domain and out-of-domain macro-F1. Among the compared models, it achieves the highest MER-UniBench mean score (84.13), ranks first by mean score in all three input-modality settings, and attains the highest MME-Emotion Rec-S (42.4).

2 Related Work

2.1 Multimodal Affective Reasoning

Masked pretraining and transfer learning strengthen facial, acoustic, and textual affect representations (Sun et al., 2023; Chen et al., 2024; 2026; Ma et al., 2024; Tian et al., 2020). Building on these representations, multimodal emotion recognition fuses visual, acoustic, and linguistic features to capture complementary affective cues and contextual dependencies (Cheng et al., 2023; Wang et al., 2025; Fang et al., 2025). Large language models extend affective modeling toward cause analysis and descriptive understanding (Cheng et al., 2024b; Lian et al., 2025b), supported by emotion-oriented instruction data and supervised fine-tuning (Xie et al., 2024; Cheng et al., 2024a; Lian et al., 2025a). Recent GRPO-based post-training further explores task rewards, perceptual grounding, and tool-assisted or slow–fast reasoning (Zhao et al., 2025a; Lian et al., 2025d; Han et al., 2026b; Liu et al., 2026; Han et al., 2026a). OmniOPSD complements these advances by using offline evidence descriptions as privileged teacher context for dense supervision on student-generated responses.

2.2 On-Policy Distillation and Self-Distillation

On-policy distillation applies teacher supervision to student-generated trajectories, typically through token-level distribution matching (Agarwal et al., 2024; Lu & Lab, 2025). When teacher logits are unavailable for direct distribution matching, logit-free alternatives instead derive feedback from sampled teacher continuations (Zhou et al., 2026). Context-conditioned on-policy self-distillation uses reference solutions or textual feedback to guide a self-teacher's predictions on student-generated prefixes (Hübotter et al., 2026; Zhao et al., 2026a; Cao et al., 2026). Further work examines generalization risks from rich teacher context (Kim et al., 2026) and develops reflection-guided, localized correction (Zhao et al., 2026b). Multimodal extensions ground teacher supervision in image evidence through image crops or textual descriptions of question-relevant visual cues (Yuan et al., 2026; Tian et al., 2026). Alongside these visual grounding strategies, Wang et al. (2026) separate image-based perceptual supervision from reference-conditioned reasoning guidance to reduce answer-driven shortcuts. OmniOPSD studies rationale-privileged on-policy self-distillation for omni-modal affective reasoning, using offline multimodal evidence descriptions as teacher context.

3 Method

This section presents OmniOPSD, a rationale-privileged on-policy self-distillation framework for omni-modal affective reasoning (Figure 2). We define the task and privileged context, describe student rollouts and teacher re-scoring, and introduce the distillation objective and teacher update.

3.1 Task and Privileged Context

We consider multimodal affective tasks such as emotion, sentiment, and intent recognition, where predictions depend on interpreting the available visual, acoustic, and linguistic signals in context. Let $\mathcal{D} = \{(o_i, q_i, a_i, r_i)\}_{i=1}^{N}$ contain N training examples indexed by i, where o_i contains the available omni-modal input, q_i is the task instruction, and a_i is the annotated or consensus answer. Each reference *rationale* $r_i = \mathrm{Serialize}(e_i, a_i)$ pairs a_i with frontier-generated reasoning e_i. When available, this reasoning describes potentially relevant facial behavior, vocal delivery, linguistic content, and temporal context, including how verbal and nonverbal cues inform the affective interpretation. These descriptions give the teacher an auxiliary interpretation of the input (Appendix C).

The student and teacher receive the same omni-modal input and task instruction, with the reference rationale included only in the teacher context:

$$c_i^S = \mathrm{Prompt}_S(o_i, q_i), \tag{1}$$

$$c_i^T = \mathrm{Prompt}_T(o_i, q_i, r_i). \tag{2}$$

During these updates, r_i serves exclusively as teacher context and is not a supervised target sequence. The student constructs its response from the omni-modal input and task instruction in c_i^S.

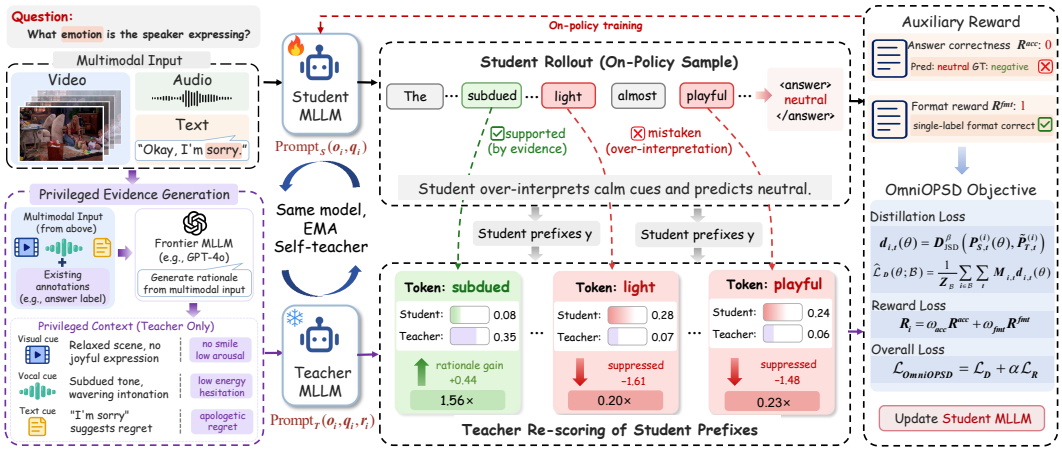


Figure 2: **Overview of OmniOPSD.** The framework supports video, audio, and text inputs and trains the student through distillation from a teacher with privileged information. The example illustrates guidance on affective reasoning despite an incorrect final prediction.

3.2 Student Rollout and Teacher Re-Scoring

The student is prompted to explain its interpretation of the available multimodal cues and then provide the task-specific answer. At optimizer step k, the student with parameters θ_k samples a completion $y_i = (y_{i,1}, \ldots, y_{i,L_i})$ from c_i^S, where L_i is the completion length, $y_{i,t}$ its t-th token, and $y_{i,<t}$ the prefix before position t. Sampling follows decoding configuration G, with settings for the main benchmark run given in Appendix E.2. The teacher obtains its target distributions by re-scoring these student-generated prefixes under c_i^T, without decoding a separate completion. Reference rationales thus inform token-level continuation targets along the student's own interpretations of affective cues, regardless of final-answer correctness.

The student and local teacher share an architecture, tokenizer, and vocabulary $\mathcal{V}$, with $u_\theta(c, y_{<t})$ denoting the next-token logits and $\tau > 0$ the distillation temperature. For a fixed rollout, the differentiable student distribution and teacher target are

$$P_{S,t}^{(i)}(\theta) = \mathrm{softmax}\big(u_\theta(c_i^S, y_{i,<t})/\tau\big)\,, \tag{3}$$

$$\widetilde{P}_{T,t}^{(i)} = \mathrm{stopgrad}\Big[\mathrm{softmax}\Big(u_{\widetilde{\theta}_k}(c_i^T, y_{i,<t})/\tau\Big)\Big]\,, \tag{4}$$

where $\bar{\theta}_k$ denotes the EMA teacher parameters, with $\widetilde{\theta}_k = \theta_k$ before their initialization and $\widetilde{\theta}_k = \bar{\theta}_k$ afterward. Student gradients are evaluated at $\theta = \theta_k$, with teacher targets fixed during backpropagation (Appendix B.3).

3.3 Token-Normalized Self-Distillation

At each valid completion position, OmniOPSD minimizes the generalized Jensen–Shannon divergence (JSD) between the student and the fixed teacher target. For mixing coefficient $\beta \in (0, 1)$,

$$d_{i,t}(\theta) = D_{\mathrm{JSD}}^{\beta}\Big(P_{S,t}^{(i)}(\theta), \widetilde{P}_{T,t}^{(i)}\Big)\,, \tag{5}$$

with the mixture and the decomposition into Kullback–Leibler (KL) divergences given in Appendix A.1. For a local batch $\mathcal{B}$, the token-normalized loss is

$$\widehat{\mathcal{L}}_D(\theta; \mathcal{B}) = \frac{1}{Z_{\mathcal{B}}} \sum_{i \in \mathcal{B}} \sum_t M_{i,t}\, d_{i,t}(\theta), \tag{6}$$

where $M_{i,t} \in \{0, 1\}$ selects valid completion tokens and $Z_{\mathcal{B}} = \sum_{i \in \mathcal{B}} \sum_t M_{i,t} > 0$ counts them.

Distillation covers all valid completion tokens, including those used for the explanation and final answer, so rationale-conditioned teacher preferences can shape the full generated response.

With sampled tokens and teacher targets fixed, differentiating Eq. (6) yields an on-policy semi-gradient over visited prefixes (see Appendix A for JSD properties and derivations).

3.4 Optimization and Teacher Update

For emotion, sentiment, and intent tasks, we optionally complement distillation with rewards R^{acc} for agreement with the annotated answer and R^{fmt} for output-format compliance. Their weighted sum, with fixed hyperparameters ω_{acc} and ω_{fmt}, defines the sequence reward (Appendix E.2):

$$R_i = \omega_{\mathrm{acc}} R^{\mathrm{acc}}(y_i, a_i) + \omega_{\mathrm{fmt}} R^{\mathrm{fmt}}(y_i). \tag{7}$$

A sequence-level advantage derived from R_i weights the per-completion, length-normalized surrogate $\widehat{\mathcal{L}}_R(\theta; \mathcal{B})$ (Appendix B.2). On the same local batch, the combined objective uses $\alpha \geq 0$ to weight this surrogate relative to distillation:

$$\widehat{\mathcal{L}}_{\mathrm{OmniOPSD}}(\theta; \mathcal{B}) = \widehat{\mathcal{L}}_D(\theta; \mathcal{B}) + \alpha\, \widehat{\mathcal{L}}_R(\theta; \mathcal{B}), \qquad \alpha \geq 0. \tag{8}$$

An exponential-moving-average (EMA) teacher smooths the rationale-conditioned scoring parameters. Scoring initially uses the current student parameters without gradients. The updated student initializes the teacher state after the first optimizer step, and subsequent steps use EMA updates with decay $\mu \in [0, 1)$ (Appendix B):

$$\bar{\theta}_{k+1} = \mu \bar{\theta}_k + (1 - \mu)\theta_{k+1}, \qquad \mu \in [0, 1). \tag{9}$$

With reference rationales cached offline, this optimization stage requires no frontier-model logits or online frontier-model calls. At inference, the student performs the affective task using only the available omni-modal input and task instruction.

4 Experiments

We evaluate OmniOPSD on MER-UniBench and MME-Emotion for affective recognition and reasoning, with controlled comparisons against SFT and GRPO at 3B and 7B scales. Ablations assess the contributions of on-policy sampling and rationale-privileged self-distillation. Training dynamics, token-level diagnostics, and qualitative cases further characterize learning behavior and teacher guidance along student-generated reasoning trajectories.

4.1 Experimental Setup

Datasets and evaluation metrics. MER-UniBench (Lian et al., 2025a) covers sentiment analysis, basic emotion recognition, and fine-grained emotion recognition, measured by weighted-average F1 (WAF), hit rate (HIT), and emotion-wheel F1 (EW-F1), respectively, with an unweighted mean over its nine datasets. MME-Emotion (Zhang et al., 2026a) assesses recognition and reasoning through Rec-S and Rea-S, which evaluate final predictions and intermediate reasoning steps, respectively, and are combined in CoT-S. The controlled ablations jointly train on MELD (Poria et al., 2019), MIntRec 2.0 (Zhang et al., 2024), and IEMOCAP (Busso et al., 2008) for sentiment analysis, intent recognition, and emotion recognition, respectively. Out-of-domain (OOD) evaluation further assesses generalization to Mandarin conversations through intent and emotion recognition on MC-EIU (Liu et al., 2024), and to single-label and multi-label emotion recognition on MAFW (Liu et al., 2022b). All controlled ablations use macro-F1, and the overall in-domain (ID) and OOD scores are unweighted means over three and four tasks, respectively. Appendix D provides further details on dataset characteristics, evaluation splits, task-specific label mappings, and scoring protocols.

Implementation details. Tables 1 and 2 report one Qwen2.5-Omni-7B model post-trained by an SFT cold start and one epoch of OmniOPSD updates and evaluated under all three input-modality settings and on MME-Emotion (Appendix E.2), while the ablations post-train Qwen2.5-Omni-3B and Qwen2.5-Omni-7B on the joint MELD, MIntRec 2.0, and IEMOCAP mixture. OmniOPSD uses a learning rate of 10^{-6}, EMA decay $\mu = 0.999$, and full-vocabulary JSD with $\beta = 0.5$ and temperature $\tau = 1.0$. The reward weight is $\alpha = 0.1$ for the main benchmarks and 0.2 for ablations. During OmniOPSD updates, only the teacher receives rationales combining generated reasoning and ground-truth answers. Details of rationale construction and experimental settings are provided in Appendix C and Appendix E, respectively.

Table 1: **Main results on MER-UniBench (%).** The best score in each column within each input-modality block is in bold, and OmniOPSD is post-trained from Qwen2.5-Omni-7B.

Model	Sentiment				Basic emotion				Fine-grained	Mean
	MOSI	MOSEI	SIMS	SIMS v2	MER2023	MER2024	MELD	IEMOCAP	OV-MERD+	
Input modality: Audio + Text										
OneLLM	64.01	54.09	63.39	61.98	25.52	17.21	28.32	33.44	22.25	41.14
SECap	55.76	54.18	59.51	57.41	40.95	52.46	25.56	36.92	36.97	46.64
PandaGPT	66.06	61.33	62.93	58.88	33.57	39.04	31.91	36.55	31.33	46.84
Qwen-Audio	70.09	46.90	70.73	65.26	41.85	31.61	49.09	35.47	32.36	49.26
SALMONN	81.00	67.03	68.69	65.93	55.53	45.38	45.62	46.84	45.00	57.89
AffectGPT	83.46	80.74	82.99	83.75	72.94	73.41	56.63	55.68	59.98	72.18
AffectGPT-R1	80.13	80.01	84.49	82.31	**81.69**	**93.49**	63.74	63.85	**65.49**	77.24
OmniOPSD (Ours)	**88.90**	**87.37**	**87.14**	**86.52**	77.27	86.47	**71.06**	**82.75**	56.09	**80.40**
Input modality: Video + Text										
Otter	52.89	50.44	57.56	53.12	16.41	14.65	22.57	29.08	16.63	34.82
Video-LLaVA	56.37	61.64	53.28	57.45	36.93	30.25	30.73	38.95	34.00	44.40
PandaGPT	58.50	64.25	62.07	65.25	39.13	47.16	38.33	47.21	35.07	50.77
Video-ChatGPT	54.42	63.12	64.82	65.80	44.86	46.80	37.33	56.83	39.80	52.64
VideoChat2	66.84	54.32	69.49	70.66	33.67	54.50	36.64	48.70	39.21	52.67
LLaMA-VID	61.78	63.89	69.35	67.48	50.72	57.60	42.75	46.02	45.01	56.07
VideoChat	65.13	63.61	69.52	72.14	48.73	57.30	41.11	48.38	44.52	56.71
Chat-UniVi	54.53	63.18	68.15	66.36	57.62	65.67	45.61	52.37	48.00	57.94
mPLUG-Owl	72.40	72.91	72.13	75.00	56.86	59.89	49.11	55.54	48.18	62.45
AffectGPT	82.39	81.57	87.20	86.29	74.58	75.29	57.63	62.19	61.65	74.31
AffectGPT-R1	78.78	79.07	85.91	85.85	**77.72**	**85.29**	61.09	67.42	**62.42**	75.95
OmniOPSD (Ours)	**88.29**	**88.12**	**87.53**	**87.94**	76.32	85.27	**68.16**	**79.11**	56.24	**79.66**
Input modality: Audio + Video + Text										
PandaGPT	61.92	67.61	68.38	67.23	40.21	51.89	37.88	44.04	37.12	52.92
R1-Omni	58.02	56.48	71.82	68.58	64.17	67.43	43.20	51.58	55.24	59.61
Emotion-LLaMA	66.13	67.66	78.32	77.23	59.38	73.62	46.76	55.47	52.97	64.17
Emotion-LLaMAv2	86.28	87.69	87.72	86.02	77.28	86.90	51.32	84.05	62.90	78.91
AffectGPT	81.30	80.90	88.49	86.18	78.54	78.80	55.65	60.54	62.52	74.77
AffectGPT-R1	79.39	79.24	88.25	84.97	84.32	**93.75**	63.12	74.26	**68.05**	79.48
OPPO	86.50	84.63	86.22	88.26	**87.73**	90.34	64.06	73.60	67.16	81.05
OmniOPSD (Ours)	**90.06**	**89.03**	**90.76**	**89.14**	85.79	92.43	**71.89**	**88.59**	59.52	**84.13**

4.2 Main Results on MER-UniBench

Table 1 shows that OmniOPSD achieves the highest reported mean in all three input-modality settings: 84.13 with audio–video–text, compared with OPPO's reported 81.05, and 80.40 and 79.66 with audio–text and video–text, exceeding AffectGPT-R1 by 3.16% and 3.71%, respectively. With all three modalities, OmniOPSD leads on six of nine datasets: all four sentiment datasets, including 90.06 WAF on CMU-MOSI and 90.76 on CH-SIMS, as well as MELD and IEMOCAP, with HIT scores of 71.89 and 88.59. Its 59.52 EW-F1 on OV-MERD+ still trails AffectGPT-R1's 68.05, identifying fine-grained open-vocabulary prediction as a remaining limitation of the overall training approach. These results highlight OmniOPSD's strong sentiment and basic emotion recognition across modality combinations with a single post-trained checkpoint.

4.3 Main Results on MME-Emotion

Table 2 shows that OmniOPSD achieves the highest Rec-S (42.4%) among all compared models and the highest CoT-S (51.4%) among the compared open-source audio–video–text models. Relative to Qwen2.5-Omni, its gains primarily reflect more accurate final judgments while preserving reasoning quality. Compared with R1-Omni and AffectGPT, it achieves higher Rec-S and Rea-S, with higher CoT-S on seven of eight tasks. Its overall CoT-S also approaches that of GPT-4o. Among all compared models, OmniOPSD achieves the highest CoT-S on fine-grained sentiment analysis (56.1%), noise-robust emotion recognition (75.3%), and sentiment analysis (64.9%). Together, these results highlight OmniOPSD's strengths in both fine-grained sentiment understanding and robust emotion recognition, demonstrating its effectiveness across diverse multimodal affective tasks.

Table 2: **Main results on MME-Emotion (%).** The best result among the closed-source models and among the open-source audio-video-text models in each score column is in bold, and Qwen2.5-Omni is the base model of OmniOPSD.

Model	Task-level CoT-S								Mean		
	ER-Lab	ER-Wild	FG-ER	FG-SA	ML-ER	Noise-ER	IR	SA	Rec-S	Rea-S	CoT-S
Closed-source MLLMs: Video + Text											
GPT-4o	**50.6**	50.8	**50.2**	**43.6**	56.2	53.0	**59.0**	60.2	27.8	**79.8**	53.8
GPT-4.1	33.9	42.1	47.8	39.1	55.6	47.0	52.2	54.0	28.8	65.2	47.0
Gemini-2.0-Flash	39.8	49.0	44.7	27.5	55.0	60.0	46.3	51.6	36.3	60.0	48.1
Gemini-2.5-Flash	33.9	45.3	29.2	31.3	34.4	63.8	45.1	50.1	34.7	52.7	43.7
Gemini-2.5-Pro	42.7	**52.0**	49.8	40.5	**65.5**	**69.4**	56.4	**63.4**	**39.3**	72.7	**56.0**
Open-source MLLMs: Audio or Video + Text											
Qwen2-Audio	26.3	30.3	21.8	41.1	63.2	48.2	36.6	59.5	34.1	50.4	42.3
Qwen2-VL-7B	9.1	31.9	15.4	22.6	27.3	47.8	43.6	47.3	29.2	38.1	33.7
Qwen2.5-VL-7B	42.0	42.1	40.9	34.4	49.4	50.3	55.3	53.6	28.4	64.8	46.6
Video-LLaVA	9.9	11.7	18.7	39.7	54.9	29.9	27.6	45.8	25.8	32.8	29.3
Open-source MLLMs: Audio + Video + Text											
Video-LLaMA	24.8	35.1	34.1	34.6	49.2	33.0	41.4	41.0	26.1	48.5	37.3
Video-LLaMA2	11.9	35.0	13.3	27.2	35.5	29.0	15.5	32.5	29.2	27.7	28.4
Qwen2.5-Omni	35.2	36.7	26.3	25.7	38.0	42.1	43.6	46.3	17.4	59.3	38.4
Emotion-LLaMA	10.5	16.1	7.5	2.9	3.1	29.2	12.0	13.1	25.1	0.4	12.8
HumanOmni	12.4	26.4	8.6	3.9	18.5	34.2	8.0	16.0	36.0	0.3	18.1
R1-Omni	39.3	43.6	27.3	29.8	**49.6**	59.7	37.7	44.2	26.3	58.6	42.4
AffectGPT	41.2	25.1	22.9	32.1	38.3	38.1	22.9	35.0	11.9	50.6	31.2
OPPO	**54.6**	**45.5**	**42.0**	40.7	39.6	66.2	34.8	60.6	31.0	**68.1**	49.5
OmniOPSD (Ours)	47.1	45.3	28.3	**56.1**	34.8	**75.3**	**48.2**	**64.9**	**42.4**	60.3	**51.4**

4.4 Ablation and Training Analysis

Post-training effectiveness and transfer. Table 3 shows that OmniOPSD outperforms SFT and GRPO in mean ID and OOD macro-F1 at both model scales. Its ID means reach 0.4021 (3B) and 0.4146 (7B), improving over the base models by 4.80% and 4.93%, with the best scores in five of the six ID comparisons. Among post-training methods, it leads on three of four OOD tasks per scale, exceeding SFT's mean by 1.80% and 1.50% and GRPO's by 1.30% and 0.33%, respectively. SFT remains competitive in-domain but yields the lowest OOD mean among the post-training methods. Both on-policy methods transfer better, with OmniOPSD outperforming GRPO. In this comparison, OmniOPSD offers the strongest combination of ID adaptation and OOD retention, supporting the value of dense teacher guidance on student-generated trajectories.

Inference-time reasoning and on-policy sampling. Table 4 holds model parameters fixed within each method and varies inference mode. Thinking improves OmniOPSD's ID/OOD means by $2.51\%/1.72\%$, whereas both means decrease slightly for GRPO. AffectGPT-R1's separately trained answer-only variant also outperforms its reasoning variant on OV-MERD+ (68.39% vs. 66.49%) (Lian et al., 2025d). Despite different protocols, both comparisons suggest that explicit reasoning alone need not improve prediction. GRPO's sparse outcome rewards may leave intermediate interpretations weakly constrained, allowing hallucinated affective cues to mislead final predictions. OmniOPSD addresses this supervision gap through dense token-level targets from a rationale-informed teacher on student-generated prefixes, guiding intermediate reasoning even when the final answer is wrong. Table 5 further shows that online prefix refresh exceeds both cached controls in ID and OOD means, leading on all three ID tasks and two of four OOD tasks (Appendix F, Tables 11 and 13). Together, these results support OmniOPSD's complementary strengths: on-policy sampling for stronger OOD generalization and dense teacher supervision for evidence-informed reasoning.

Rationale-privileged self-distillation and efficiency. Figure 3 compares separately trained models under matched protocols. Adding rationales to the teacher's label context improves macro-F1 by 1.38%, 0.56%, and 1.34% on MELD, MIntRec 2.0, and IEMOCAP (Table 12), indicating that ratio-

Table 3: **Ablation results across ID and OOD affective benchmarks.** The left block reports ID results after joint training on MELD, MIntRec 2.0, and IEMOCAP, while the right block reports OOD generalization to Mandarin conversations on MC-EIU and to single-label and multi-label emotion recognition on MAFW. Bold numbers indicate the best result within each base-model block, and underlining indicates the second-best result.

Base model	Method	In-domain (ID)			Out-of-domain (OOD)			
		MELD	MIntRec 2.0	IEMOCAP	MC-EIU Intent-ZH	MC-EIU Emotion-ZH	MAFW Emotion-SL	MAFW Emotion-ML
Qwen2.5-Omni-3B	Base	0.5941	0.2936	0.1747	**0.2597**	0.2747	0.2127	**0.2636**
	SFT	0.6355	0.3161	0.2275	0.2270	0.2593	0.1920	0.2383
	GRPO	0.6093	0.2921	0.2208	0.2352	0.2670	0.2048	0.2298
	OmniOPSD	**0.6427**	**0.3248**	**0.2388**	0.2551	**0.2834**	**0.2159**	0.2343
Qwen2.5-Omni-7B	Base	0.6051	0.3007	0.1900	**0.2539**	**0.2839**	**0.2453**	**0.3164**
	SFT	0.6451	**0.3581**	0.2343	0.2214	0.2450	0.1965	0.2330
	GRPO	0.6424	0.3179	0.2267	0.2196	0.2607	0.2194	0.2431
	OmniOPSD	**0.6501**	0.3510	**0.2426**	0.2248	0.2360	0.2250	0.2700

Table 4: **Effect of inference-time thinking.** We report mean ID and OOD macro-F1 for Qwen2.5-Omni-3B with and without thinking.

Method	Mode	ID Avg.	OOD Avg.
GRPO	w/o Think	0.3752	0.2347
	w/ Think	0.3741	0.2342
OmniOPSD	w/o Think	0.3770	0.2300
	w/ Think	**0.4021**	**0.2472**

Table 5: **Effect of on-policy prefix refresh.** We report mean ID and OOD macro-F1 for Qwen2.5-Omni-3B trained with cached student prefixes or online prefix refresh.

Prefix source	ID Avg.	OOD Avg.
Single cached prefix	0.3853	0.2394
Frozen prefix pool	0.3855	0.2331
Online refresh	**0.4099**	**0.2438**

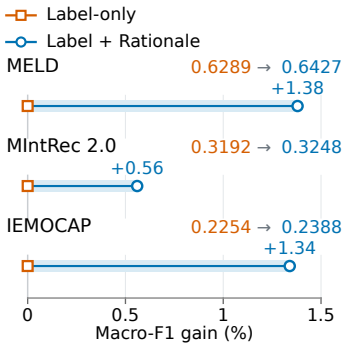


Figure 3: **Rationale supervision.** Macro-F1 gains (%) over Label-only for separately trained OmniOPSD (3B) models.

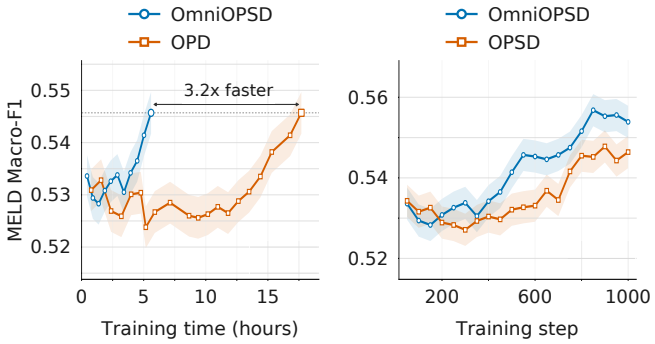


Figure 4: **Comparison with OPD and OPSD on MELD.** Left: OmniOPSD versus OPD over training time, with a dashed line at OPD's final macro-F1 of 0.546. Right: OmniOPSD versus OPSD by optimizer step.

nales supply useful affective evidence beyond the annotated answers. Figure 4 evaluates a separate configuration from Table 3. In the left panel, OmniOPSD reaches 0.546 MELD macro-F1 in 5.6 hours, versus 17.7 hours for OPD, using approximately 68% less training time in this run. With rationales cached offline, local teacher scoring avoids frontier-model logits and online frontier-model calls during optimization, reducing teacher-inference overhead. In the right panel, OmniOPSD exceeds the OPSD baseline without frontier-generated rationales at every evaluation from step 200 onward, finishing at 0.554 versus 0.546. This improvement supports the combined training design in Eq. (8): dense self-distillation targets guide intermediate reasoning, while reinforcement learning rewards final-answer accuracy and format compliance. Together, these comparisons support efficient self-distillation with complementary process and outcome supervision.

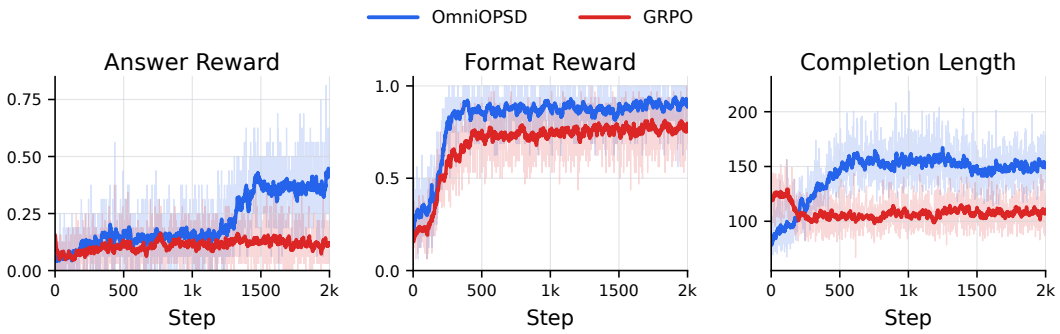


Figure 5: **Training dynamics of GRPO and OmniOPSD on Qwen2.5-Omni-3B.** Three panels track the answer-accuracy reward, the format reward, and the completion length over training.

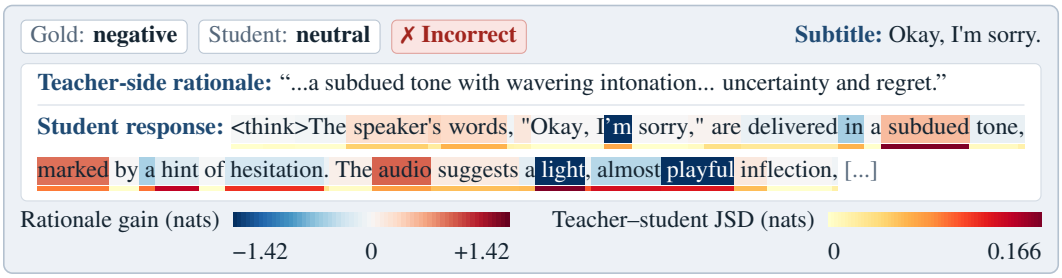


Figure 6: **Affect-related guidance on an incorrect MELD rollout.** Background colors show log-probability changes from adding the rationale to the label-only teacher context, while aligned strips show full-vocabulary teacher–student JSD on the same student prefixes (full response: Figure 8). This illustrates OmniOPSD's dense token-level supervision: the teacher favors affective cues consistent with its rationale and suppresses conflicting interpretations, even on an incorrect rollout.

Training dynamics. Figure 5 shows OmniOPSD's sharp late-stage answer-reward gains (left) and consistently higher smoothed format rewards than GRPO (middle). Completion length (right) increases and stabilizes near 150 tokens for OmniOPSD, whereas GRPO's rises slightly before declining toward 100–110. These trends suggest that OmniOPSD encourages thinking through dense teacher guidance while improving answer correctness and format adherence.

Dense affective guidance and qualitative cases. Figure 6 shows an incorrect MELD rollout rescored with checkpoint, labels, and student prefixes fixed. Rationale conditioning raises the probability of *subdued* and decreases those of *light* and *playful*, consistent with the teacher's affective account despite the wrong final prediction. Across 600 held-out responses, the dataset-balanced mean teacher–student JSD is 0.01973 for reasoning text and 0.00071 for format tags (Appendix G.1, Table 14). Together with the rationale ablation, these diagnostics support affect-related guidance on student-generated reasoning, including unsuccessful rollouts. Appendix G.2 (Tables 15–19) provides author-selected cases illustrating OmniOPSD's multimodal affective reasoning strengths.

5 Conclusion

We presented OmniOPSD, a rationale-privileged on-policy self-distillation framework for omni-modal affective reasoning. OmniOPSD uses offline frontier-generated multimodal evidence descriptions and annotated answers as teacher-only privileged context. An EMA teacher provides dense token-level supervision on student-generated prefixes, while the student uses only the original task input during rollout and inference. Results show that OmniOPSD outperforms SFT and GRPO in mean in-domain and out-of-domain macro-F1 at both 3B and 7B scales. Among the compared models, it achieves the highest MER-UniBench mean in all three modality settings and the highest MME-Emotion Rec-S. Ablations support rationale conditioning and on-policy prefix refresh. Future work will address rationale reliability and fine-grained open-vocabulary emotion recognition.

Ethics statement

All datasets and their associated labels are obtained through the required application procedures and used within the scope of the granted research permissions. Training and evaluation use the provided audio, video, and text content and their original labels, without altering the underlying data or annotations. The experiments focus on affective understanding and do not involve discriminatory profiling or decision-making. Benchmark performance reflects agreement with context-dependent annotations rather than direct measurement of individuals' internal mental states.

Reproducibility statement

Appendices A–B present the formal assumptions and training algorithm. Appendices C–E document the prompts, datasets, and detailed protocols for training and evaluation. To support fair comparison, we apply consistent testing procedures to the methods evaluated in our experiments within each benchmark and compute metrics according to the corresponding evaluation protocols. Appendix F provides supplementary experimental analyses that complement the main results.

AI use statement

In this work, we used generative AI tools for retrieval and discovery (e.g., finding related work), equation and notation checks, experimental code development, multimodal cue extraction, result organization, output normalization, benchmark evaluation, and grammar and spelling edits. The authors take full responsibility for the manuscript, code, experiments, and all reported results and scientific claims.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for LLMs. *Transactions on Machine Learning Research*, 2025.

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.

Di Cao, Dongjie Fu, Hai Yu, Siqi Zheng, Xu Tan, and Tao Jin. X-opd: Cross-modal on-policy distillation for capability alignment in speech llms. *arXiv preprint arXiv:2603.24596*, 2026.

Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*, 16(2):624–638, 2024.

Yin Chen, Jia Li, Yu Zhang, Zhenzhen Hu, Shiguang Shan, Meng Wang, and Richang Hong. Static for dynamic: Towards a deeper understanding of dynamic facial expressions using static expression data. *IEEE Transactions on Affective Computing*, 17(1):438–451, 2026. doi: 10.1109/TAFFC.2025.3623135.

Zebang Cheng, Yuxiang Lin, Zhaoru Chen, Xiang Li, Shuyi Mao, Fan Zhang, Daijun Ding, Bowen Zhang, and Xiaojiang Peng. Semi-supervised multimodal emotion recognition with expression mae. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9436–9440, 2023.

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

Zebang Cheng, Fuqiang Niu, Yuxiang Lin, Zhi-qi Cheng, Xiaojiang Peng, and Bowen Zhang. MIPS at SemEval-2024 task 3: Multimodal emotion-cause pair extraction in conversations with multimodal language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 667–674. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.semeval-1.97.

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*, 2024c.

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.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14314–14324, 2025.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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.

Zhiyuan Han, Beier Zhu, Wenwen Tong, Chengwei Qin, Xinyi Wang, Jiayu Zhang, Jiangnan Chen, Hewei Guo, Dongchuan Ran, Lewei Lu, et al. Mer-r1: Multimodal emotion reasoning via slow-fast thinking synergy. *arXiv preprint arXiv:2606.27652*, 2026a.

Zhiyuan Han, Beier Zhu, Wenwen Tong, Pengyang Shao, Peipei Song, Xinyi Wang, Jiangnan Chen, Lewei Lu, and Xun Yang. Omni-perception policy optimization for multimodal emotion reasoning. *arXiv preprint arXiv:2606.25325*, 2026b.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2881–2889, 2020.

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.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms? *arXiv preprint arXiv:2603.24472*, 2026.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10143–10152, 2019.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. IntentQA: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11963–11974, 2023b.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023c.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 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.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *Computer Vision–ECCV 2024*, 2024b.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-supervised visual on-policy distillation. *arXiv preprint arXiv:2608.14144*, 2026. doi: 10.48550/arXiv.2608.14144.

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, 2023a.

Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable multimodal emotion reasoning. *arXiv preprint arXiv:2306.15401v1*, 2023b.

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. In *Proceedings of the 42nd International Conference on Machine Learning*. ML Research Press, 2025a.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, et al. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13837–13842, 2025b.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, et al. Ov-mer: Towards open-vocabulary multimodal emotion recognition. In *International Conference on Machine Learning*, pp. 37015–37050. PMLR, 2025c.

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, Xiaobai Li, and Bin He. AffectGPT-R1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*, 2025d.

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, 2024a. doi: 10.18653/v1/2024.emnlp-main.342.

Wang Lin, Yueying Feng, WenKang Han, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, and Jingyuan Chen. E^3: Exploring embodied emotion through a large-scale egocentric video dataset. *Advances in Neural Information Processing Systems*, 37:118182–118197, 2024b.

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.

Shifeng Liu, Zhengye Zhang, Sirui Zhao, Xinglong Mao, Zhehan Kan, Zhixiang Wei, Shiwei Wu, Chaoyou Fu, Tong Xu, and Enhong Chen. Actfer: Agentic facial expression recognition via active tool-augmented visual reasoning. *arXiv preprint arXiv:2604.08990*, 2026.

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.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026.

Yu Luo, Jianbo Ye, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International journal of computer vision*, 128(1):1–25, 2020.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.931.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024. doi: 10.18653/v1/2024.acl-long.679.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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.

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.

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

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. MAE-DFER: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6110–6121, 2023. doi: 10.1145/3581783.3612365.

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. In *International Conference on Learning Representations*, 2024.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4067–4076. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.374.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. ViCuR: Visual cues as recoverable privilege for multimodal on-policy distillation. *arXiv preprint arXiv:2606.05718*, 2026.

Dingdong WANG, Shujie LIU, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotionthinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 153708–153733, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation. *arXiv preprint arXiv:2606.19120*, 2026.

Yusong Wang, Xuanye Fang, Huifeng Yin, Dongyuan Li, Guoqi Li, Qi Xu, Yi Xu, Shuai Zhong, and Mingkun Xu. Big-fusion: Brain-inspired global-local context fusion framework for multimodal emotion recognition in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1574–1582, 2025.

Jiulong Wu, Yucheng Shen, Lingyong Yan, Haixin Sun, Deguo Xia, Jizhou Huang, and Min Cao. Facial-r1: Aligning reasoning and recognition for facial emotion analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26939–26947, 2026.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26596–26605, 2024.

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.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19323–19331, 2024. doi: 10.1609/aaai.v38i17.29902.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 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.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

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.

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.

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 *The Fourteenth International Conference on Learning Representations*, 2026a.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023. doi: 10.18653/v1/2023.emnlp-demo.49.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. MIntrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*, 2024.

Hanwen Zhang, Yao Liu, Peiyuan Jiang, Lang Junjie, Xie Jun, Yihui He, Yajiao Deng, Siyu Du, and Qiao Liu. Xemogpt: An explainable multimodal emotion recognition framework with cue-level perception and reasoning. *arXiv preprint arXiv:2602.05496*, 2026b.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025a.

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*, 2025b.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026a.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. Rosd: Reflective on-policy self-distillation for language model reasoning across domains. *arXiv preprint arXiv:2605.28014*, 2026b.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Peng Bo, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. Omniopd: Logit-free on-policy distillation via speculative verification. *arXiv preprint arXiv:2606.01476*, 2026.

Appendix overview. The appendices provide the mathematical foundations, experimental protocols, and supporting analyses of OmniOPSD. Appendix A derives the properties and gradients of the generalized JSD objective and clarifies the on-policy semi-gradient and local token normalization. Appendix B presents the training algorithm, optional reward surrogate, and implementation details for completion alignment and EMA teacher updates. Appendix C describes the construction of teacher-side privileged rationales and specifies the student, teacher, and evaluation prompts and output formats. Appendix D documents data sources, evaluation subsets, and scoring protocols for the controlled ablations, MER-UniBench, and MME-Emotion. Appendix E details the training configurations, implementation, and evaluation settings for the main benchmarks and controlled ablations. Appendix F presents additional experiments, including an analysis of MME-Emotion scores and response length, as well as studies of teacher-side rationale conditioning, inference-time thinking, and on-policy prefix refresh. Appendix G uses controlled re-scoring of fixed student responses to characterize token-level teacher guidance and presents qualitative cases of multimodal affective reasoning.

A THEORETICAL SUPPORT

This appendix provides theoretical support for the token-normalized distillation objective. Section A.1 establishes the bounds, equality conditions, and gradients of the generalized JSD, while Section A.2 characterizes the on-policy semi-gradient and clarifies the consequences of local valid-token normalization.

A.1 JSD PROPERTIES AND GRADIENTS

Following the notation of Section 3, let $p = P_{S,t}^{(i)}(\theta)$ and $q = \widetilde{P}_{T,t}^{(i)}$ denote the student and stop-gradient teacher distributions over $\mathcal{V}$ at a fixed completion prefix, using natural logarithms throughout. For $\beta \in (0, 1)$, define the mixture

$$p_\beta = (1 - \beta)p + \beta q, \tag{10}$$

and the generalized JSD

$$D_{\mathrm{JSD}}^{\beta}(p, q) = (1 - \beta)\,\mathrm{KL}(p \,\|\, p_\beta) + \beta\,\mathrm{KL}(q \,\|\, p_\beta). \tag{11}$$

This divergence satisfies

$$0 \le D_{\mathrm{JSD}}^{\beta}(p, q) \le h(\beta), \qquad h(\beta) = -\beta \log \beta - (1 - \beta) \log(1 - \beta), \tag{12}$$

with equality at the lower bound if and only if $p = q$, where non-negativity and the equality condition follow from KL divergence. For the upper bound, introduce a binary variable $C \sim \mathrm{Bernoulli}(\beta)$ and sample a token V from p when $C = 0$ and from q when $C = 1$, so that $D_{\mathrm{JSD}}^{\beta}(p, q) = I(C; V) \le H(C) = h(\beta)$. The same bound applies to the valid-token mean in Eq. (6), with an upper bound of $\log 2$ at $\beta = 0.5$.

Differentiating with q fixed and including both occurrences of p_β gives

$$\frac{\partial D_{\mathrm{JSD}}^{\beta}(p, q)}{\partial p_v} = (1 - \beta) \log \frac{p_v}{p_{\beta,v}}. \tag{13}$$

The non-logarithmic terms cancel because $[(1 - \beta)p_v + \beta q_v]/p_{\beta,v} = 1$, and applying the chain rule to the student logits $u = u_\theta(c_i^S, y_{i,<t})$ with $p = \mathrm{softmax}(u/\tau)$ then yields

$$\frac{\partial D_{\mathrm{JSD}}^{\beta}}{\partial u_j} = \frac{1 - \beta}{\tau} p_j \left[\log \frac{p_j}{p_{\beta,j}} - \sum_{v \in \mathcal{V}} p_v \log \frac{p_v}{p_{\beta,v}} \right]. \tag{14}$$

The mixture remains differentiable through the student while the teacher distribution is detached, and the standard implementation applies no additional τ^2 factor. For positive softmax distributions, the logit gradient is zero if and only if $p = q$: a zero gradient requires $p_v/p_{\beta,v}$ to be constant across the vocabulary, and normalization then implies $p = p_\beta = q$. Distillation can therefore provide a token-level signal even when an optional sequence reward is zero, although this gradient property does not guarantee that the teacher's preferences improve task accuracy.

A.2 ON-POLICY OPTIMIZATION

At optimizer step k, completions are sampled from $\pi_{\theta_k, G}(\cdot \mid c_i^S)$ under the decoding configuration G and re-scored by both branches, with the teacher additionally conditioned on the privileged rationale. Holding the sampled tokens, completion masks, and teacher targets fixed, the computed gradient for a local forward batch $\mathcal{B}$, with token divergence $d_{i,t}(\theta)$ and $Z_{\mathcal{B}} = \sum_{i \in \mathcal{B}} \sum_t M_{i,t} > 0$, is

$$\widehat{g}_k = \frac{1}{Z_{\mathcal{B}}} \sum_{i \in \mathcal{B}} \sum_t M_{i,t} \nabla_\theta d_{i,t}(\theta) \Big|_{\theta = \theta_k}. \tag{15}$$

Because the probability of sampling each prefix is held fixed, Eq. (15) is a semi-gradient through the re-scored student distributions on states visited by the current rollout policy, rather than the full derivative of an objective whose sampling distribution varies with θ.

Within $\mathcal{B}$, each retained token receives the same coefficient $1/Z_{\mathcal{B}}$, giving a mean over valid completion tokens in which longer retained completions contribute more terms. This generally differs from an average of independently length-normalized sequence losses, although both reduce to the response's token mean when the local batch contains one completion. Writing $N_{\mathcal{B}}(\theta) = \sum_{i\in\mathcal{B}}\sum_t M_{i,t}d_{i,t}(\theta)$, the expected finite-batch loss, with the expectation over batch selection and student rollouts at fixed θ_k, is

$$\mathbb{E}_{\mathcal{B}}\left[\widehat{\mathcal{L}}_D(\theta;\mathcal{B})\right] = \mathbb{E}_{\mathcal{B}}\left[\frac{N_{\mathcal{B}}(\theta)}{Z_{\mathcal{B}}}\right] \neq \frac{\mathbb{E}_{\mathcal{B}}[N_{\mathcal{B}}(\theta)]}{\mathbb{E}_{\mathcal{B}}[Z_{\mathcal{B}}]}. \tag{16}$$

The ratio of expectations describes a population token-weighted objective and generally differs from the expected local mean implemented in Eq. (6) when batch token counts vary.

B OmniOPSD Algorithm and Implementation Details

This appendix translates the objective in Section 3 into the complete training procedure. Section B.1 presents the EMA-teacher algorithm, Section B.2 defines the optional reward surrogate, and Section B.3 documents completion alignment, loss reduction, and teacher-state handling.

Following Section 3, $\mathcal{B}$ denotes the local batch processed by one standard logits forward pass, and $M_{i,t} \in \{0,1\}$ marks valid completion positions, with prompt and padding positions assigned zero. Completion i contains $T_i = \sum_t M_{i,t}$ valid tokens, and $Z_{\mathcal{B}} = \sum_{i\in\mathcal{B}} T_i$ counts the valid tokens in $\mathcal{B}$.

B.1 Training Algorithm

Algorithm 1 uses an EMA teacher with symmetric JSD ($\beta = 0.5$) and distillation temperature $\tau = 1$, sampling student responses under decoding configuration G and re-scoring their prefixes to obtain full-vocabulary distributions for distillation.

The sampled completion is factorized autoregressively under the decoding configuration G:

$$y_i \sim \pi_{\theta_k,G}(\cdot \mid c_i^S), \qquad \pi_{\theta_k,G}(y_i \mid c_i^S) = \prod_{t=1}^{L_i} \pi_{\theta_k,G}(y_{i,t} \mid c_i^S, y_{i,<t}). \tag{17}$$

The reduction in Eq. (6) first sums the divergence over the vocabulary and then averages over the $Z_{\mathcal{B}}$ valid completion positions, reducing to a single response's token mean when the local forward pass contains one example. Gradient accumulation and distributed synchronization combine these local losses with the trainer's gradient scaling, without replacing each $Z_{\mathcal{B}}$ with a token count pooled over the optimizer step, so the result generally differs from a global token mean when local valid-token counts vary.

B.2 Optional Reward Objective

We convert the weighted sum R_i of answer-accuracy and output-format rewards in Eq. (7) into a fixed advantage for the current update:

$$A_i = \mathrm{stopgrad}\left(\frac{R_i - b}{s}\right). \tag{18}$$

The resolved experimental configuration selects the zero, batch-mean, or running-mean baseline b and the scaling rule, with statistics computed over the reward-scoring batch, which may span workers. The scale s is one when scaling is disabled or the reward variance is zero, and otherwise equals the empirical reward standard deviation plus 10^{-4}.

With $\ell_{i,t}(\theta) = \log P_{S,t}^{(i)}(\theta)[y_{i,t}]$ and $\ell_{i,t}^{\mathrm{old}} = \mathrm{stopgrad}(\ell_{i,t}(\theta))$, the implementation uses the ratio surrogate

$$\rho_{i,t}(\theta) = \exp\left(\ell_{i,t}(\theta) - \ell_{i,t}^{\mathrm{old}}\right), \tag{19}$$

$$\widehat{\mathcal{L}}_R(\theta;\mathcal{B}) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \frac{\sum_t M_{i,t}\min(\rho_{i,t}A_i,\ \mathrm{clip}(\rho_{i,t}, 1-\epsilon_-, 1+\epsilon_+)A_i)}{\max(1, T_i)}. \tag{20}$$

Algorithm 1 OmniOPSD with an EMA Teacher

Require: Dataset $\mathcal{D} = \{(o_i, q_i, a_i, r_i)\}$, student parameters θ_0, EMA state $\bar{\theta} \leftarrow \varnothing$, decay μ, decoding configuration G, and reward weight $\alpha \geq 0$.

for each optimizer step $k = 0, 1, \ldots$ **do**
  Initialize the local accumulated gradient to zero.
  for each local forward batch $\mathcal{B}$ in the accumulation window **do**
    Construct $c_i^S = \mathrm{Prompt}_S(o_i, q_i)$ and $c_i^T = \mathrm{Prompt}_T(o_i, q_i, r_i)$ for $i \in \mathcal{B}$.
    Sample $y_i \sim \pi_{\theta_k, G}(\cdot \mid c_i^S)$ and detach the sampled token IDs.
    Encode (c_i^S, y_i) and (c_i^T, y_i), then select and align their valid completion tokens using $M_{i,t}$.
    Re-score the student completion with gradients to obtain $P_{S,t}^{(i)}(\theta)$, evaluated at $\theta = \theta_k$.
    Set $\widetilde{\theta}_k \leftarrow \theta_k$ if $\bar{\theta} = \varnothing$, and $\widetilde{\theta}_k \leftarrow \bar{\theta}_k$ otherwise.
    Under NoGrad, re-score the same completion with $\widetilde{\theta}_k$ and c_i^T to obtain $\widetilde{P}_{T,t}^{(i)}$.
    Compute $\widehat{\mathcal{L}}_D(\theta; \mathcal{B})$ by the valid-token mean in Eq. (6).
    if $\alpha > 0$ **then**
      Score the completion rewards and construct detached advantages A_i using Eq. (18).
      Compute $\widehat{\mathcal{L}}_R(\theta; \mathcal{B})$ using Eq. (20).
    else
      $\widehat{\mathcal{L}}_R(\theta; \mathcal{B}) \leftarrow 0$.
    end if
    Accumulate the gradient of $\widehat{\mathcal{L}}_D(\theta; \mathcal{B}) + \alpha \widehat{\mathcal{L}}_R(\theta; \mathcal{B})$ with the trainer's accumulation scaling.
  end for
  Average synchronized gradients across workers and update the student to θ_{k+1}.
  Set $\bar{\theta}_{k+1} \leftarrow \theta_{k+1}$ if $\bar{\theta} = \varnothing$ and otherwise apply Eq. (9).
end for
return the student π_θ.

The denominator guard assigns zero contribution to empty completions, while detaching the old log probabilities from the same forward pass makes $\rho_{i,t} = 1$ numerically. For positive clipping radii ϵ_-, ϵ_+, clipping is locally inactive, so the gradient of Eq. (20) equals the gradient of

$$-\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{A_i}{\max(1, T_i)} \sum_t M_{i,t} \ell_{i,t}(\theta). \tag{21}$$

This equivalence holds locally for gradients with the sampled tokens, lengths, masks, and advantages fixed, although the two loss values need not agree. Length normalization generally changes the objective relative to maximizing the unnormalized expected sequence reward, and a score-function interpretation additionally requires the re-scoring and rollout distributions to match, which decoding transformations such as top-p filtering can violate. We therefore use this term as a reward surrogate in the combined objective of Eq. (8).

B.3 Implementation Details

Aligned completion scoring. With a shared tokenizer and vocabulary, the student and teacher receive the same generated response token IDs inserted into their separately encoded prompts, whose lengths can differ. We shift logits and labels by one position for next-token prediction and exclude prompt and padding positions with labels equal to -100, checking the remaining student and teacher token IDs for equal counts and identical ordering before comparing their distributions. These comparisons use full-vocabulary distributions, independently of any top-p or top-k filtering in the decoding configuration G. Only the student forward retains gradients: teacher scoring runs under `torch.no_grad()` with gradient checkpointing disabled for that pass.

The standard full-vocabulary logits path uses the local reduction in Eq. (6), and when the multimodal fallback splits an input batch into chunks, each chunk uses the same reduction, and chunk losses are combined in proportion to their numbers of examples.

EMA parameter state. The teacher state stores detached copies of trainable parameters and is initialized lazily: teacher scoring uses the current student parameters until the first synchronized optimizer step, after which the updated student parameters are copied to initialize $\bar{\theta}$. After gradient accumulation, subsequent updates follow Eq. (9) once per optimizer step.

During teacher scoring, the trainer temporarily swaps the EMA values into the local model and restores the current student values afterward, so a single local model instance serves as both student and teacher. In LoRA training, the EMA therefore covers the trainable adapters while the frozen backbone is shared, whereas in full-parameter training it covers all parameters marked trainable, while frozen modules such as the visual encoder and multimodal aligner remain shared. Alongside the EMA configuration in Algorithm 1, the implementation supports an online stop-gradient teacher that uses the current student parameters on every scoring pass without maintaining an EMA state, and both configurations use only the student and its original omni-modal task prompt at inference.

C Privileged Rationale Construction and Prompt Templates

This appendix specifies how privileged teacher context is constructed and how it differs from student and evaluation inputs. Section C.1 defines the rationale components and serialization, Section C.2 gives the student and teacher training prompts, and Section C.3 records the inference prompts and output formats used for evaluation.

C.1 Rationale Construction

For each MMEVerse training example i, let a_i denote its annotated or consensus answer and let e_i denote the evidence-aware reasoning text generated offline by GPT-4o. We serialize these components to form the complete privileged rationale r_i:

$$\begin{aligned} r_i &= \text{Serialize}(e_i, a_i) \\ &= \texttt{<think>}\, e_i \,\texttt{</think><answer>}\, a_i \,\texttt{</answer>}. \end{aligned} \tag{22}$$

During OmniOPSD updates, the precomputed rationale conditions only the teacher, while the student receives only the omni-modal task input. The rationale is therefore neither included in the student rollout prompt nor used as a supervised target sequence during these updates.

C.2 Student and Teacher Training Prompts

Student context. The system instruction below accompanies user messages containing the omni-modal input, the subtitle when available, and a task instruction. Preprocessing randomly selects from the first ten valid templates for the requested task and modality, or from all valid templates if fewer than ten are available, and shuffles the candidate labels.

System instruction

```
You are an Affective Computing expert specializing in emotion analysis,
sentiment recognition, and intent recognition.

Core Directives:

- Comprehensive Analysis: Synthesize semantic information, pragmatic
cues, and any implied multimodal features (e.g., tone, facial
expressions, action cues) in your evaluation.

- Modality Honesty: NEVER analyze or make assumptions about missing or
unusable modalities.

- Structured Workflow: You MUST strictly adhere to the user's requested
output format. Always output your step-by-step reasoning first, followed
by your final answer.
```

The representative templates below share a subtitle prefix when a transcript is available, with braces marking example-specific fields: `{answers}` and `{modalities}` specify the candidate labels and modality description, and `{target}` denotes `sentiment` or `speaker's intent`.

Student task templates

Shared prefix

```
The subtitle of this video is: {subtitle}
```

Emotion recognition

```
Choose from: {answers}. Based on {modalities}, reason through your interpretation in <think></think> and write your final answer in <answer></answer>.
```

Sentiment and intent recognition

```
Choose from: {answers}. Based on {modalities}, explain your reasoning in <think></think> and provide the {target} in <answer></answer>.
```

Multi-label emotion recognition

```
Choose one or more applicable emotions from: {answers}. Based on {modalities}, reason through your interpretation in <think></think> and write your final answer(s) in <answer></answer>. If multiple emotions apply, separate them with a comma.
```

IntentQA uses the question and its five options directly:

IntentQA instruction

```
Question: {question and options} Based on the video content, explain your reasoning in <think></think> and output only the letter (A/B/C/D/E) of the correct option in <answer></answer>.
```

Teacher context. The teacher retains the student's system instruction and omni-modal user message, appending the suffix below to the last user message with `{solution}` set to r_i in Eq. (22).

Privileged teacher suffix

```
Below is a reference answer for the same task.

Reference answer:

{solution}

Study the reference answer carefully. Then solve the original task again in your own words while preserving the original output format requirements.
```

The teacher uses this prompt to re-score the student-generated prefixes without generating a separate response (Appendix B).

C.3 Evaluation Prompts and Output Format

At evaluation time, the model receives the task input without the privileged rationale and can produce either a direct label or reasoning followed by a tagged answer. In the single-label templates below, `{target}` denotes `sentiment`, `emotion`, or `speaker's intent`, with dataset-specific variants supplying the corresponding label spaces and task wording.

Single-label evaluation templates

Direct prediction (only_label)

```
Given the visual, auditory, and spoken content in the video, what is the most appropriate {task} label? Choose from: {label space}. Output only the label.
```

Reasoning and answer (think)

```
Choose from: {label space}. Based on the speaker's expressions, vocal
tone, and spoken content, explain your reasoning in <think></think> and
provide the {target} in <answer></answer>.
```

Multi-label instructions request applicable emotions separated by semicolons, and open-vocabulary variants omit the candidate list. The evaluator applies task-specific answer extraction and normalization, with an optional API-assisted step for outputs that fail local parsing. Evaluation settings and metrics are described in Appendices D and E.

D Dataset and Benchmark Details

This appendix documents the data and evaluation protocols underlying the reported results. Section D.1 describes the source datasets, prepared evaluation inventories, and controlled-ablation splits, while Sections D.2 and D.3 then specify the benchmark subsets, metrics, and evaluation procedures for MER-UniBench and MME-Emotion, respectively. We first provide the provenance of the models compared in Tables 1 and 2.

Compared models. The external baselines in the MER-UniBench comparison comprise OneLLM (Han et al., 2024), SECap (Xu et al., 2024), PandaGPT (Su et al., 2023), Qwen-Audio (Chu et al., 2023), SALMONN (Tang et al., 2024), Otter (Li et al., 2023a), Video-LLaVA (Lin et al., 2024a), Video-ChatGPT (Maaz et al., 2024), VideoChat2 (Li et al., 2024a), LLaMA-VID (Li et al., 2024b), VideoChat (Li et al., 2023c), Chat-UniVi (Jin et al., 2024), mPLUG-Owl (Ye et al., 2023), Emotion-LLaMA (Cheng et al., 2024a), Emotion-LLaMAv2 (Peng et al., 2026), R1-Omni (Zhao et al., 2025a), AffectGPT (Lian et al., 2025a), AffectGPT-R1 (Lian et al., 2025d), and OPPO (Han et al., 2026b). Our model is initialized from Qwen2.5-Omni-7B (Xu et al., 2025).

The MME-Emotion comparison additionally includes GPT-4o (OpenAI, 2024), GPT-4.1[1], Gemini-2.0-Flash[2], Gemini-2.5-Flash, and Gemini-2.5-Pro (Gemini Team, 2025), Qwen2-Audio (Chu et al., 2024), Qwen2-VL-7B (Wang et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), Video-LLaMA (Zhang et al., 2023), Video-LLaMA2 (Cheng et al., 2024c), Qwen2.5-Omni (Xu et al., 2025), Emotion-LLaMA (Cheng et al., 2024a), HumanOmni (Zhao et al., 2025b), R1-Omni (Zhao et al., 2025a), AffectGPT (Lian et al., 2025a), and OPPO (Han et al., 2026b), alongside Video-LLaVA (Lin et al., 2024a).

D.1 Data Sources and Ablation Protocol

IEMOCAP. IEMOCAP (Busso et al., 2008) contains English scripted and improvised dyadic interactions performed by ten actors across five sessions. Its approximately 12 hours of recordings comprise 10,039 utterances with speech, video, motion capture, and transcripts. Annotations include nine named emotion categories and an *Other* option, together with valence, activation, and dominance ratings.

MELD. MELD (Poria et al., 2019) extends EmotionLines with audio and video from the English television series *Friends*. It contains 13,708 utterances in 1,433 multiparty dialogues, with seven emotion categories and three sentiment polarities: positive, negative, and neutral. The official training, development, and test sets contain 9,989, 1,109, and 2,610 utterances.

MIntRec 2.0. MIntRec 2.0 (Zhang et al., 2024) supports multimodal intent recognition and out-of-scope detection in English conversations from *Superstore*, *The Big Bang Theory*, and *Friends*. It provides text, audio, video, and speaker annotations for 15,040 utterances in 1,245 dialogues. The dataset includes 9,304 utterances across 30 in-scope intent classes and 5,736 out-of-scope utterances. The released in-scope partitions contain 6,165 training, 1,106 validation, and 2,033 test utterances.

[1]Official model page: https://openai.com/index/gpt-4-1/

[2]Official model announcement: https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

MC-EIU. MC-EIU (Liu et al., 2024) addresses joint emotion and intent understanding in English and Mandarin television conversations. Its 56,012 utterances from 4,970 conversations have aligned text, audio, and video, with seven emotion and nine intent categories. The Mandarin portion contains 11,003 utterances, divided into 7,643 training, 1,148 validation, and 2,212 test examples while keeping conversations intact.

IntentQA. IntentQA (Li et al., 2023b) studies the intentions behind human actions through video question answering. It selects causal and temporal questions from NExT-QA and adds action annotations to support reasoning about why actions occur and how their intentions connect events. The dataset contains 16,297 question–answer pairs over 4,303 videos, with 12,119 training, 2,044 validation, and 2,134 test questions. We use its training split as an additional source of five-way video question–answer examples during the training process.

MAFW. MAFW (Liu et al., 2022b) contains 10,045 audiovisual clips from films, television programmes, and short videos, with annotations for eleven underlying emotions, compound expressions, and emotion-related descriptions in English and Chinese. The original single-expression benchmark retains 9,172 clips after annotation validation, while its compound benchmark distinguishes 43 classes, comprising eleven single emotions and 32 combinations. The original evaluation uses five-fold cross-validation, while our multi-label setting represents emotions as an eleven-dimensional multi-hot vector. The evaluation inventory contains 1,839 single-label examples and 580 examples with multi-label annotations.

DFEW. DFEW (Dynamic Facial Expression in the Wild) (Jiang et al., 2020) contains 16,372 movie clips with challenging facial expressions, including occlusions, pose changes, and variations in illumination. Its seven emotion categories are happiness, sadness, neutrality, anger, surprise, disgust, and fear. The processed single-label benchmark retains 11,697 clips and uses five-fold cross-validation. Our evaluation inventory contains the 2,341 clips in its first test fold.

$\mathbf{E^3}$. E^3 (Lin et al., 2024b) studies embodied emotion in multilingual egocentric videos, with over 50 hours of visual, acoustic, and textual content in the published collection. It provides annotations for recognition, temporal localization, and reasoning about emotion. The eight categories are angry, disgusted, happy, sad, sarcastic, scared, shy, and surprised. Our recognition inventory comprises 3,288 examples: 1,662 from validation and 1,626 from test.

CAER. CAER (Context-Aware Emotion Recognition) (Lee et al., 2019) contains 13,201 audiovisual clips from 79 television shows, supporting emotion recognition from both facial expressions and surrounding context. Its seven labels are anger, disgust, fear, happiness, sadness, surprise, and neutrality. The original split allocates 70%, 10%, and 20% to training, validation, and test, respectively. Our prepared evaluation inventory contains 3,953 clips: 1,316 from validation and 2,637 from test.

BoLD. BoLD (Body Language Dataset) (Luo et al., 2020) contains 9,876 movie clips with 13,239 annotated character instances. It provides video, audio, and estimated body poses for studying bodily expressions of emotion. Each instance is annotated with 26 non-exclusive emotion categories and continuous valence, arousal, and dominance scores.

CMU-MOSI. CMU-MOSI (Zadeh et al., 2016) contains 2,199 English opinion segments from 93 online videos by 89 speakers. Each segment includes speech, visual behavior, and a transcript. Sentiment intensity is obtained by averaging five annotators' ratings on a scale from -3 to $+3$, combining assessments of linguistic and nonverbal sentiment cues. The evaluation inventory contains 686 test segments.

CMU-MOSEI. CMU-MOSEI (Zadeh et al., 2018) expands this setting to 23,453 annotated segments from 3,228 English online videos, covering 1,000 speakers and 250 topics. It provides aligned transcripts and speech, sentiment ratings in $[-3, 3]$, and intensity annotations for happiness, sadness, anger, fear, disgust, and surprise. MER-UniBench uses its sentiment annotations, with 4,659 test segments in the evaluation inventory.

CH-SIMS. CH-SIMS (Yu et al., 2020) comprises 2,281 Chinese clips from films, television series, and variety shows. Each clip has a joint multimodal sentiment annotation and separate textual, acoustic, and visual annotations, with scores in $[-1, 1]$. These modality-specific labels capture agreement and conflict between verbal and nonverbal cues. The official split contains 1,368 training, 456 validation, and 457 test clips.

CH-SIMS v2.0. CH-SIMS v2.0 (Liu et al., 2022a) expands the collection to 4,402 labeled clips and 10,161 unlabeled clips. It retains unimodal and multimodal sentiment annotations while increasing the diversity of acoustic and visual expressions. The release distinguishes supervised v2(s) and unsupervised v2(u) subsets, with the supervised subset denoted CH-SIMS_v2_s in our result tables. The evaluation inventory contains 1,034 test clips, matching the reference MER-UniBench subset.

MER2023. MER2023 (Lian et al., 2023a) is a Chinese audiovisual emotion collection based on CHEAVD and clips from films and television series. It provides valence annotations and labels for six emotion categories: neutral, anger, happiness, sadness, worry, and surprise. The challenge includes 3,373 training/validation examples and separate evaluation subsets for joint emotion–valence prediction, noise robustness, and semi-supervised learning. MER-UniBench selects the MER-MULTI subset.

MER2024. MER2024 (Lian et al., 2024) extends the collection to 5,030 labeled training/validation clips and a pool of 115,595 additional clips from Chinese films and television series. It retains six basic emotion categories and supports semi-supervised, noise-robust, and open-vocabulary evaluation. MER-UniBench uses the labeled MER-SEMI evaluation subset. The broader evaluation inventory contains 2,339 clips, comprising 1,169 MER-SEMI and 1,170 MER-NOISE examples. MER-NOISE applies additive speech noise to audio and blur to video to assess robustness to degraded inputs.

EMER. EMER (Explainable Multimodal Emotion Reasoning) introduces natural-language explanations of emotional states using audio, video, and subtitles (Lian et al., 2023b). Its initial release contains 100 non-neutral clips selected from MER2023. Human annotators describe facial expressions, body movements, vocal delivery, speech content, and contextual cues, and these descriptions are consolidated with ChatGPT and manually checked to form explanatory references. This resource supports evaluation of both emotion predictions and their supporting evidence.

MER-OV. MER-OV is the open-vocabulary track of MER2024 (Lian et al., 2024). Its annotated subset comprises 332 clips selected from the challenge's training/validation collection of Chinese film and television material. The annotation procedure combines acoustic and visual clues with subtitles to produce emotion-related descriptions, from which an LLM extracts multiple emotion labels without a predefined category set. The task assesses nuanced and coexisting emotions.

OV-MERD+. OV-MERD+ (Lian et al., 2025a) extends the original 332-example OV-MERD collection (Lian et al., 2025c) to 532 Chinese film and television clips. Each clip is associated with an emotion description and a variable-size set of emotion labels, without a fixed category inventory. The annotation framework combines model-generated acoustic and visual clues with human verification, supporting evaluation of subtle and co-occurring emotions.

Controlled ablations. The in-domain experiments jointly train on MELD, MIntRec 2.0, and IEMOCAP for sentiment analysis, intent recognition, and emotion recognition, respectively. Out-of-domain evaluation assesses generalization to Mandarin conversations on MC-EIU through intent and emotion recognition (*Intent-ZH* and *Emotion-ZH*), and to single-label and multi-label emotion recognition on MAFW (*Emotion-SL* and *Emotion-ML*). The latter uses the multi-hot representation described above. Throughout the controlled-ablation results and associated diagnostics, MELD, MIntRec 2.0, and IEMOCAP refer to the tasks defined here. MER-UniBench follows its own benchmark protocol and evaluates emotion annotations on MELD. MELD and MIntRec 2.0 use their released training partitions, while IEMOCAP uses Sessions 1–4 for training and Session 5 for evaluation. For the prefix-refresh comparison in Table 13, the primary evaluation uses the MELD test split (2,610 examples), the MIntRec 2.0 in-scope test split (2,033 examples), and 2,170 prepared

IEMOCAP Session 5 examples. The prepared transfer inventories contain 2,166 MC-EIU examples for each of Intent-ZH and Emotion-ZH, 1,839 MAFW Emotion-SL examples, and 580 MAFW Emotion-ML examples. The additional 1,109 MELD development and 1,106 MIntRec validation examples in the merged prediction files are used only for split-level diagnostics and are excluded from the reported test-only scores. All ablation scores are macro-F1:

$$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{23}$$

where C is the evaluated label inventory and zero-denominator terms are assigned zero. For multi-label evaluation, each emotion is treated as a separate binary decision. This metric differs from MER-UniBench's HIT, even when both evaluations use the same source dataset.

D.2 MER-UniBench

MER-UniBench (Lian et al., 2025a) evaluates free-form affect predictions across nine datasets spanning sentiment, basic emotion, and fine-grained emotion recognition, using the reference subsets listed in Table 6. We report audio–text, video–text, and audio–video–text results in the modality groups of Table 1, with external baseline scores taken from the corresponding publications.

Table 6: **MER-UniBench evaluation subsets.** Counts are benchmark sizes reported by Lian et al. (2025a), before metric-specific filtering.

Dataset	Task	Metric	Subset	Count
CMU-MOSI	Sentiment	WAF	Test	686
CMU-MOSEI	Sentiment	WAF	Test	4,659
CH-SIMS	Sentiment	WAF	Test	457
CH-SIMS_v2_s	Sentiment	WAF	Test	1,034
MER2023	Basic emotion	HIT	MER-MULTI	411
MER2024	Basic emotion	HIT	MER-SEMI	1,169
MELD	Basic emotion	HIT	Test	2,610
IEMOCAP	Basic emotion	HIT	Session 5	1,241
OV-MERD+	Fine-grained emotion	EW-F1	All	532

Sentiment evaluation with WAF. Reference scores below and above zero are mapped to negative and positive sentiment, respectively, while neutral ground-truth examples are excluded. After extracting sentiment from the generated response, weighted-average F1 is computed as

$$\mathrm{WAF} = \sum_{c\in\{\mathrm{negative},\mathrm{positive}\}}\frac{n_c}{N}\,\mathrm{F1}_c, \tag{24}$$

where n_c is class support among the N retained examples. The released evaluator maps an extracted neutral prediction to the non-positive class when binarizing predictions.

Emotion normalization and HIT. The evaluator extracts emotion sets from generated responses, normalizes word forms and synonyms, and maps related labels to core categories using five emotion wheels. Let G_k denote normalization under wheel k, with $A_{ik} = G_k(Y_i)$ and $B_{ik} = G_k(\widehat{Y}_i)$ for the reference and predicted sets. The released evaluator excludes examples with empty mapped references, leaving N_k examples indexed by $\mathcal{I}_k$. For basic emotion recognition, each reference contains one label, and

$$\mathrm{HIT} = \frac{1}{5}\sum_{k=1}^{5}\frac{1}{N_k}\sum_{i\in\mathcal{I}_k}\mathbb{I}[A_{ik}\cap B_{ik}\neq\varnothing]\,. \tag{25}$$

HIT checks whether the prediction includes the reference emotion without penalizing additional predicted labels.

Fine-grained emotion evaluation with EW-F1. For OV-MERD+, precision and recall are first averaged over examples within each wheel:

$$P_k = \frac{1}{N_k} \sum_{i \in \mathcal{I}_k} \frac{|A_{ik} \cap B_{ik}|}{\max(1, |B_{ik}|)}, \qquad R_k = \frac{1}{N_k} \sum_{i \in \mathcal{I}_k} \frac{|A_{ik} \cap B_{ik}|}{|A_{ik}|}. \tag{26}$$

An empty prediction receives zero precision and recall. The reported emotion-wheel F1 averages the resulting wheel-level harmonic means:

$$\text{EW-F1} = \frac{1}{5} \sum_{k=1}^{5} \frac{2P_k R_k}{P_k + R_k}, \tag{27}$$

with zero assigned when $P_k + R_k = 0$. All MER-UniBench scores are expressed as percentages, and the overall score is the unweighted mean of the nine dataset-level primary metrics.

D.3 MME-EMOTION

MME-Emotion (Zhang et al., 2026a) contains 6,500 video–question pairs covering eight tasks and 27 scenarios, with tasks, source datasets, and sample counts listed in Table 7. It resamples test examples from public affective datasets and supplies candidate labels in each question.

Table 7: **Task composition of MME-Emotion.** The listed datasets are the source corpora, and counts denote video–question pairs in the released benchmark.

Task	Datasets	Count
Emotion recognition in the lab (ER-Lab)	IEMOCAP	500
Emotion recognition in the wild (ER-Wild)	CAER, MC-EIU, E^3, MELD, MAFW, DFEW	1,800
Fine-grained emotion recognition (FG-ER)	BoLD	500
Fine-grained sentiment analysis (FG-SA)	CMU-MOSI	500
Intent recognition (IR)	MC-EIU	500
Multi-label emotion recognition (ML-ER)	MAFW, CMU-MOSEI	600
Emotion recognition under noise (Noise-ER)	MER2024	500
Sentiment analysis (SA)	CMU-MOSI, CMU-MOSEI, MELD, CH-SIMS v2.0	1,600
Total (8 tasks)	–	**6,500**

Recognition and reasoning metrics. Rec-S averages the judge's binary rating of the final prediction step over scored examples. The MME-Emotion paper describes multi-label Rec-S as the fraction of ground-truth labels recovered, but its released judge prompt rates the final step as correct only when the prediction matches the ground-truth label, and OmniOPSD's scores follow this released prompt and metric script. Rea-S first averages binary correctness over the reasoning steps of each response, excluding the final prediction, and then averages over scored examples. A response without reasoning steps receives zero. The benchmark combines recognition and reasoning with equal weight:

$$\text{CoT-S} = \tfrac{1}{2}\,\text{Rec-S} + \tfrac{1}{2}\,\text{Rea-S}\,. \tag{28}$$

The reference overall evaluation aggregates scores over examples. When every listed example receives a valid score, a metric S is therefore aggregated as $S_{\text{overall}} = \sum_{t=1}^{8} n_t S_t / \sum_{t=1}^{8} n_t$, where S_t and n_t are the score and sample count of task t. This aggregation is inferred from the published task scores: it reproduces the overall table, unlike an unweighted task mean, although the MME-Emotion paper does not explicitly give the equation. The mean Rec-S, Rea-S, and CoT-S scores in Tables 2 and 9 are example-weighted over all 6,500 examples. Two response-length statistics accompany the scores: Avg Step, the mean number of extracted steps per response including the final prediction step, and Avg Token, the mean number of response tokens.

E EXPERIMENTAL SETTINGS AND REPRODUCIBILITY

This appendix records the configurations needed to interpret and reproduce the experiments. Section E.1 specifies the controlled ablations and their privileged-information, thinking-mode, and

prefix-source controls, Section E.2 describes the two-stage main-benchmark training protocol, and Section E.3 documents the implementation, multimodal preprocessing, and evaluation settings.

E.1 Controlled Ablation Settings

The controlled experiments jointly train on MELD, MIntRec 2.0, and IEMOCAP, comparing SFT, GRPO, and OmniOPSD with the corresponding Qwen2.5-Omni-3B and Qwen2.5-Omni-7B base models (Table 3). All post-training methods use a learning rate of 10^{-6} and two epochs, with out-of-domain evaluation assessing generalization to Mandarin conversations on MC-EIU and to single-label and multi-label emotion recognition on MAFW. All scores are reported as macro-F1, with task definitions and splits described in Appendix D.1.

SFT optimizes the serialized reference responses, GRPO uses answer-accuracy and output-format rewards, and OmniOPSD uses the reference rationale to condition the teacher scoring the student's generated response, with prompts given in Appendix C. For OmniOPSD ablation runs, the reward-surrogate weight in Eq. (8) is $\alpha = 0.2$, set by the implementation parameter `reward_training_weight`.

Teacher-conditioning ablation. Table 12 evaluates two separately optimized Qwen2.5-Omni-3B models under matched training and evaluation protocols. Both use OmniOPSD and differ only in teacher-side privileged information: *Label-only* supplies the ground-truth label, whereas *Label + Rationale* also supplies the evidence-aware rationale described in Appendix C.1.

Thinking-mode comparison. Tables 4 and 11 compare GRPO and OmniOPSD with the Qwen2.5-Omni-3B backbone under the reported *w/ Think* and *w/o Think* inference modes, which enable or disable an explicit reasoning trace before the final answer. Within each method, model parameters are held fixed across the two modes. Thus, this comparison varies inference-time output behavior without retraining, whereas privileged rationales condition the teacher during training.

Prefix-refresh control. The separate comparison summarized in Table 5 and detailed in Table 13 uses three Qwen2.5-Omni-3B runs initialized from the same cold-start weights and trained with seed 42. Each run completes two epochs and 752 optimizer steps. The online condition regenerates prefixes from the current student, the frozen-pool condition is recorded as using a two-prefix initial-policy pool, and the single-cached condition reuses one initial-policy prefix.

E.2 Main Benchmark Training

MER-UniBench and MME-Emotion use the same Qwen2.5-Omni-7B checkpoint, obtained through SFT cold-start training at 10^{-5} followed by OmniOPSD at 10^{-6} (Tables 1 and 2). The second stage starts from SFT step 1,000, before the first SFT epoch is complete, and runs for one epoch comprising 2,181 optimizer steps. Table 8 gives this recipe, which differs from the two-epoch ablation protocol. MER-UniBench evaluates audio–text, video–text, and audio–video–text inputs, while MME-Emotion uses audio–video–text inputs, with tasks, metrics, and aggregation rules detailed in Appendix D.

Table 8: **Training settings for the main benchmark checkpoint.** The settings correspond to Tables 1 and 2.

Setting	SFT cold start	OmniOPSD
Learning rate	10^{-5}	10^{-6}
Selected optimizer step	1,000	2,181
Tuning	Full	Full
Per-device batch size	1	1
Gradient accumulation	2	2
CLI `max_length`	8,192	24,000

Both stages use fused AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, gradient clipping at 1.0, and a cosine learning-rate schedule with a 5% warmup ratio. Training uses

DeepSpeed ZeRO-2 with the visual encoder and multimodal aligner frozen. The training and data seeds are both 42. For the main benchmark run, OmniOPSD rollouts use decoding configuration G with temperature 1.0, top-$p = 0.9$, top-$k = 50$, and at most 512 new tokens. For this main-benchmark run, `reward_training_weight` is set to $\alpha = 0.1$. The reward weights in Eq. (7) are $(\omega_{\text{acc}}, \omega_{\text{fmt}}) = (1.0, 0.5)$. These coefficients remain fixed during training. The implementation sets `reward_funcs` to `think_acc think_format` and `reward_weights` to `1 0.5`, in the same order.

E.3 Implementation and Evaluation

The ModelScope SWIFT implementation uses an EMA teacher with decay $\mu = 0.999$, full-vocabulary JSD with $\beta = 0.5$, and distillation temperature $\tau = 1.0$. Appendix B specifies loss reduction, the reward surrogate, and EMA updates.

Benchmark inference uses a 512-token response cap and batch sizes of 4 for MER-UniBench and 2 for MME-Emotion. Classification ablations use task-specific answer parsing and macro-F1, whereas MER-UniBench and MME-Emotion follow their benchmark scoring protocols. Any model-assisted answer extraction occurs during evaluation post-processing, separately from student generation (prompt formats: Appendix C.3).

The prefix-refresh evaluation uses `think` decoding at temperature zero, at most four video frames, and a usual inference batch size of 8. Predictions are parsed locally without external-model repair, and unresolved outputs are counted as errors. On its 13,564-example test-only inventory, the online, frozen-pool, and single-cached runs contain 262 (1.93%), 248 (1.83%), and 267 (1.97%) unresolved outputs, respectively.

F Additional Experiments

This appendix first relates MME-Emotion scores to response length and then reports task-level evidence for teacher-side privileged rationales, inference-time thinking, and the single-seed prefix-refresh comparison. The inference-time and prefix-refresh results are summarized in Tables 4 and 5. The classification ablations report macro-F1 on a 0–1 scale, with ID and OOD averages computed as unweighted means of task-level scores across three and four tasks, respectively, without pooling predictions or averaging across training seeds. Score differences that we report with % denote absolute differences on the percentage scale. The tasks and evaluation splits are described in Appendix D.1.

F.1 MME-Emotion Scores and Response Length

Table 9 complements the task-level CoT-S in Table 2 with model size, input modalities, and the response-length statistics defined in Appendix D.3. All 6,500 OmniOPSD responses receive valid judge scores. For 58 of them (0.9%), the multimodal judge request fails, and GPT-4o scores the steps without video frames, using only the acoustic clues and the ground-truth label, and excluding these responses changes each mean by at most 0.03%. The unrounded means are 42.4462 Rec-S, 60.2753 Rea-S, and 51.3607 CoT-S, with 6.6589 steps and 155.0285 tokens per response.

Response length across models. Across the 17 GPT-4o-judged models with reported length statistics, Avg. steps is strongly rank-correlated with Rea-S (Spearman $\rho = 0.79$) and CoT-S ($\rho = 0.75$) but not with Rec-S ($\rho = 0.17$), and Avg. tokens shows the same pattern ($\rho = 0.81$, 0.79, and 0.17). This agrees with the positive association between step count and performance reported by Zhang et al. (2026a). The association is partly structural, because a response without reasoning steps receives zero Rea-S: Emotion-LLaMA and HumanOmni answer directly, with 1.0 step and at most 0.4 Rea-S. Without these two models, the correlation weakens but remains positive ($\rho = 0.69$ for Rea-S and 0.64 for CoT-S). Length does not determine Rea-S, however. OmniOPSD produces the most steps in the table (6.7), yet its Rea-S of 60.3 is below that of GPT-4o (79.8 with 4.4 steps), Gemini-2.5-Pro, GPT-4.1, and Qwen2.5-VL-7B. Relative to Qwen2.5-Omni, OmniOPSD raises the step count from 3.7 to 6.7 and nearly doubles the token count, from 78.6 to 155.0, while Rec-S rises by 25.0% and Rea-S by 1.0%. Its steps are also short, averaging about 23 tokens, compared with about 39 for GPT-4o and 106 for Gemini-2.5-Pro. These cross-model length comparisons are approximate: OmniOPSD answers in the think-and-answer format of Appendix C.3, whereas the

Table 9: **Overall MME-Emotion results with response length.** A, V, and T denote audio, video, and text inputs, and scores are percentages. Avg. steps counts the extracted steps per response, including the final prediction step, and Avg. tokens is the mean response length. Baseline values are taken from Zhang et al. (2026a). †OPPO's scores are copied from its paper, whose authors re-judge all methods with gemini-3.1-flash-lite-preview, but its model size and response lengths are not reported. Bold marks the best score among the closed-source models and among the open-source audio–video–text models.

Model	Size	Input			Avg. steps	Avg. tokens	Rec-S	Rea-S	CoT-S
		A	V	T					
Closed-source MLLMs									
GPT-4o	–		✓	✓	4.4	169.4	27.8	**79.8**	53.8
GPT-4.1	–		✓	✓	5.2	141.2	28.8	65.2	47.0
Gemini-2.0-Flash	–		✓	✓	4.1	64.7	36.3	60.0	48.1
Gemini-2.5-Flash	–		✓	✓	4.3	261.8	34.7	52.7	43.7
Gemini-2.5-Pro	–		✓	✓	5.1	538.6	**39.3**	72.7	**56.0**
Open-source MLLMs									
Qwen2-Audio	7B	✓		✓	3.0	40.3	34.1	50.4	42.3
Qwen2-VL-7B	7B		✓	✓	2.9	68.0	29.2	38.1	33.7
Qwen2.5-VL-7B	7B		✓	✓	4.8	169.7	28.4	64.8	46.6
Video-LLaVA	7B		✓	✓	2.3	19.1	25.8	32.8	29.3
Video-LLaMA	7B	✓	✓	✓	4.5	122.5	26.1	48.5	37.3
Video-LLaMA2	7B	✓	✓	✓	2.6	37.7	29.2	27.7	28.4
Qwen2.5-Omni	7B	✓	✓	✓	3.7	78.6	17.4	59.3	38.4
Emotion-LLaMA	7B	✓	✓	✓	1.0	2.3	25.1	0.4	12.8
HumanOmni	7B	✓	✓	✓	1.0	1.3	36.0	0.3	18.1
R1-Omni	0.5B	✓	✓	✓	5.0	156.2	26.3	58.6	42.4
AffectGPT	7B	✓	✓	✓	4.9	122.8	11.9	50.6	31.2
OPPO†	–	✓	✓	✓	–	–	31.0	**68.1**	49.5
OmniOPSD (Ours)	7B	✓	✓	✓	6.7	155.0	**42.4**	60.3	**51.4**

published baselines use the benchmark's step-by-step instruction, and token counts depend on the tokenizer.

Response structure of OmniOPSD. OmniOPSD produces a nearly constant response structure. Of its $6{,}500$ responses, 97.4% contain six or seven extracted steps, including the final prediction step, and all contain between five and nine. Across tasks, the average ranges only from 6.50 steps on IR to 6.84 on FG-ER and from 149.0 to 166.8 tokens (Table 10), whereas CoT-S ranges from 28.3 to 75.3. Within the model, step count is essentially uncorrelated with both the final-step rating and the per-response reasoning score (Pearson $|r| < 0.02$), and responses with correct and incorrect final predictions contain 6.65 and 6.67 steps on average.

Step-level judgments and final predictions. The judge's step ratings closely follow the final prediction (last two columns of Table 10). When the final prediction step is judged correct, 94.1% of a response's reasoning steps are judged correct on average, but when it is judged incorrect, this share falls to 35.3%, and every reasoning step is judged incorrect in $1{,}460$ of the $3{,}741$ incorrect responses (39.0%). In incorrect responses, the last reasoning step, which names the predicted label in 90.0% of all responses, is judged correct least often (22.0%, compared with 38.4% for the first reasoning step). The gap is smaller in tasks where incorrect predictions are often partially correct: in ML-ER, 331 of 600 predicted label sets are strict subsets of the reference set, and incorrect responses retain 62.1 Rea-S, compared with 16.3 in SA. Under this protocol, Rea-S therefore reflects agreement between the reasoning and the reference label as well as the consistency of the reasoning with the multimodal cues.

Table 10: **Per-task response statistics of OmniOPSD on MME-Emotion.** Avg. steps and Avg. tokens are means per response, and scores are percentages. The last two columns average Rea-S over responses whose final prediction step is judged correct or incorrect, with the correct-response means for FG-ER and ML-ER based on 36 and 34 responses, respectively.

Task	Avg. steps	Avg. tokens	Rec-S	Rea-S	CoT-S	Rea-S by final step	
						Correct	Incorrect
ER-Lab	6.65	149.0	37.2	56.9	47.1	97.5	32.9
ER-Wild	6.70	152.7	41.4	49.3	45.3	95.1	16.9
FG-ER	6.84	166.8	7.2	49.4	28.3	84.2	46.7
FG-SA	6.62	156.9	37.8	74.5	56.1	92.0	63.8
ML-ER	6.64	157.1	5.7	63.9	34.8	93.1	62.1
Noise-ER	6.64	151.2	74.4	76.3	75.3	96.5	17.7
IR	6.50	162.0	35.4	61.1	48.2	84.6	48.2
SA	6.64	153.4	63.8	66.0	64.9	94.3	16.3
Overall	6.66	155.0	42.4	60.3	51.4	94.1	35.3

F.2 Rationale Supervision and Inference-Time Thinking

We separate two questions: whether evidence-aware rationales improve teacher-side supervision during training, and whether an explicit reasoning trace improves predictions at inference. The training ablation optimizes a separate model for each teacher-conditioning regime, whereas the output-mode study holds model parameters fixed within each method.

Inference-time output-mode comparison. Table 11 compares GRPO and OmniOPSD with Qwen2.5-Omni-3B under two inference modes: *w/o Think* omits an explicit thinking phase, whereas *w/ Think* generates thinking before the answer. For each method, both modes are decoded from the same trained parameters. Accordingly, this comparison isolates inference-time generation behavior from the contribution of privileged rationales during training. The training recipe is given in Appendix E.1.

Table 11: **Task-level results for inference-time thinking.** Macro-F1 with Qwen2.5-Omni-3B. IEMOCAP uses Session 5. Within each method, model parameters are held fixed across output modes. Bold indicates the best result in each column.

In-domain (ID)

Method	Mode	MELD	MIntRec 2.0	IEMOCAP	ID Avg.
GRPO	w/o Think	0.6854	0.3163	0.1239	0.3752
GRPO	w/ Think	0.6093	0.2921	0.2208	0.3741
OmniOPSD	w/o Think	**0.6869**	0.3157	0.1285	0.3770
OmniOPSD	w/ Think	0.6427	**0.3248**	**0.2388**	**0.4021**

Out-of-domain (OOD)

Method	Mode	MC-EIU		MAFW		OOD Avg.
		Intent-ZH	Emotion-ZH	Emotion-SL	Emotion-ML	
GRPO	w/o Think	0.2503	0.2823	0.1602	0.2460	0.2347
GRPO	w/ Think	0.2352	0.2670	0.2048	0.2298	0.2342
OmniOPSD	w/o Think	0.2515	0.2621	0.1597	**0.2466**	0.2300
OmniOPSD	w/ Think	**0.2551**	**0.2834**	**0.2159**	0.2343	**0.2472**

For OmniOPSD, the thinking variant raises ID and OOD mean macro-F1 from 0.3770 to 0.4021 and from 0.2300 to 0.2472, respectively, or by approximately 2.51% and 1.72%. The differences

are task-dependent: macro-F1 on MELD decreases from 0.6869 to 0.6427, and MAFW Emotion-ML decreases from 0.2466 to 0.2343, while the other five tasks improve. GRPO's averages change only slightly, from 0.3752 to 0.3741 on ID tasks and from 0.2347 to 0.2342 on OOD tasks.

Teacher-conditioning ablation. Table 12 provides the exact scores plotted in Figure 3, comparing separately optimized Qwen2.5-Omni-3B models under matched training and evaluation protocols. Both models use OmniOPSD. The *Label-only* variant conditions the self-teacher on the ground-truth label alone, while *Label + Rationale* additionally supplies the evidence-aware rationale defined in Appendix C.1.

Table 12: **Effect of privileged rationale supervision.** Macro-F1 for two Qwen2.5-Omni-3B models separately trained with OmniOPSD under matched training and evaluation protocols. The conditions differ only in the information available to the self-teacher. IEMOCAP uses Session 5. Bold indicates the better result in each column.

Teacher privileged information	MELD	MIntRec 2.0	IEMOCAP
Label-only	0.6289	0.3192	0.2254
Label + Rationale	**0.6427**	**0.3248**	**0.2388**

Using evidence-aware rationales to condition the self-teacher yields consistent gains over ground-truth labels alone: macro-F1 increases by 1.38%, 0.56%, and 1.34% on MELD, MIntRec 2.0, and IEMOCAP, respectively. Because each condition is optimized separately under matched protocols, this ablation measures the effect of training-time teacher conditioning rather than an inference-time prompting effect.

AffectGPT-R1 offers a complementary observation. Its separately optimized GRPO variants attain 66.49 on OV-MERD+ when generating a reasoning trace before the answer and 68.39 when predicting the answer directly (Lian et al., 2025d). In OmniOPSD, the rationale serves a different role: it is available only to the self-teacher, which supplies dense token-level targets on student-generated trajectories, and is never an imitation target for the student. Considered alongside the fixed-parameter inference comparison in Table 4, these results support rationale-informed process supervision rather than the mere inclusion of a generated reasoning trace.

F.3 On-Policy Prefix Refresh

We compare three sources of completion prefixes used for distillation. The online variant samples a fresh completion from the current student at each visit. The recorded frozen-pool configuration uses a two-prefix initial-student pool intended for fixed rotation, whereas the single-cached control repeatedly reuses one initial-student completion. The three Qwen2.5-Omni-3B runs start from the same cold-start weights, use training seed 42, and complete two epochs, corresponding to 752 optimizer steps.

Table 13 reports test-only macro-F1. MELD uses the 2,610-example test split, MIntRec 2.0 uses its 2,033-example in-scope test split, and IEMOCAP uses the 2,170 prepared Session 5 examples. The two MC-EIU tasks each contain 2,166 prepared examples, while MAFW Emotion-SL and Emotion-ML contain 1,839 and 580 examples, respectively. Evaluation uses deterministic local answer extraction, counts unresolved outputs as errors, and invokes no external model for output repair.

Online refresh obtains the highest macro-F1 on all three in-domain tests. Its ID mean is 0.4099, compared with 0.3855 for the frozen pool and 0.3853 for a single cached prefix. The corresponding OOD means are 0.2438, 0.2331, and 0.2394. Across all seven tasks, their unweighted overall means are 0.3150, 0.2984, and 0.3019, respectively. Online refresh achieves the highest macro-F1 among the three variants on five of the seven tasks. It also yields higher macro-F1 than the frozen-pool control on six of the seven tasks, with MC-EIU Intent-ZH as the sole exception. The improvement is not uniform: the single-cached control is higher on MC-EIU Intent-ZH and slightly higher on MAFW Emotion-ML (0.2308 versus 0.2305). Thus, the single-seed comparison provides descriptive evidence for an aggregate benefit from current-policy prefixes, with larger gains in-domain than out-of-domain. These results should be interpreted within the three prefix-source variants under their recorded initialization and evaluation protocol.

Table 13: **Task-level results for on-policy prefix refresh.** Test-only macro-F1 with Qwen2.5-Omni-3B under training seed 42. We compare reuse of a single cached prefix, a frozen pool of prefixes generated by the initial student, and online refresh from the current student. IEMOCAP uses Session 5. Bold indicates the best result in each column.

In-domain (ID)

Prefix source	**MELD**	**MIntRec 2.0**	**IEMOCAP**	**ID Avg.**
Single cached prefix	0.6062	0.2978	0.2520	0.3853
Frozen prefix pool	0.6021	0.3058	0.2485	0.3855
Online refresh	**0.6279**	**0.3383**	**0.2636**	**0.4099**

Out-of-domain (OOD)

Prefix source	**MC-EIU**		**MAFW**		**OOD Avg.**
	Intent-ZH	Emotion-ZH	Emotion-SL	Emotion-ML	
Single cached prefix	**0.2468**	0.2768	0.2031	**0.2308**	0.2394
Frozen prefix pool	0.2378	0.2704	0.2093	0.2148	0.2331
Online refresh	0.2375	**0.2961**	**0.2110**	0.2305	**0.2438**

G Qualitative Visualizations

This appendix complements the aggregate results with two views of model behavior. Section G.1 quantifies and visualizes how rationale conditioning changes token-level teacher guidance on fixed student-generated trajectories, while Section G.2 examines representative multimodal reasoning cases across affective tasks.

G.1 Token-Level Teacher–Student Guidance

We examine rationale-conditioned supervision using response-level statistics over 600 held-out examples and six cases: one correct and one incorrect prediction each for MELD, MIntRec 2.0, and IEMOCAP, following the task definitions in Appendix D.1.

Controlled re-scoring of student trajectories. Using the Qwen2.5-Omni-3B cold-start model with the OmniOPSD LoRA adapter at checkpoint 2000, the student generates one response per video from the original omni-modal task prompt, without the ground-truth label or reference rationale. The decoding configuration G uses temperature 1, top-$p = 0.9$, top-$k = 50$, and at most 512 new tokens. We re-score these responses with the teacher conditioned on either the answer label alone or the complete privileged rationale, while holding the frozen checkpoint θ_c, media inputs, response prefixes, and correct labels fixed. This post-hoc diagnostic isolates conditioning effects without reconstructing historical EMA targets or performing optimizer updates.

For a response $y \sim \pi_{\theta_c,G}(\cdot \mid c^S)$, omitting the example index i, define the next-token distributions at its sampled prefixes as

$$\begin{aligned} p_t &= \mathrm{softmax}\big(u_{\theta_c}(c^S, y_{<t})\big)\,, \\ q_t^L &= \mathrm{softmax}\big(u_{\theta_c}(c_L^T, y_{<t})\big)\,, \qquad q_t^R = \mathrm{softmax}\big(u_{\theta_c}(c_R^T, y_{<t})\big)\,, \end{aligned} \tag{29}$$

where c^S and $c_R^T = c^T$ are the student and teacher contexts in Eqs. (1)–(2), and c_L^T replaces the rationale r in c^T with the answer-only reference <answer>a</answer>. Thus p_t and q_t^R instantiate Eqs. (3)–(4) with $\tau = 1$ and $\theta = \widetilde{\theta}_k = \theta_c$. Scoring evaluates the student prefixes using full-vocabulary logits with FP32 log-softmax at temperature 1, without the top-p/top-k truncation used for generation.

Metrics and visual encoding. Token backgrounds show the change in each sampled token's log-probability when the rationale is added to the label-only teacher context,

$$g_t = \log q_t^R(y_t) - \log q_t^L(y_t). \tag{30}$$

Red indicates that adding the rationale increases the probability of the sampled token, blue indicates a decrease, and near-white indicates little change. The ratio $\exp(g_t) = q_t^R(y_t)/q_t^L(y_t)$ gives the multiplicative change in probability. Because teacher weights, labels, and student prefixes are held fixed, g_t measures the incremental effect of rationale conditioning. It is distinct from the teacher–student sampled-token log-probability difference,

$$\Delta_t = \log q_t^R(y_t) - \log p_t(y_t). \tag{31}$$

The thin strip aligned below each token shows the symmetric, full-vocabulary JSD of Eq. (11) at $\beta = 0.5$,

$$d_t = D_{\mathrm{JSD}}^{0.5}\left(p_t, q_t^R\right). \tag{32}$$

The strip measures full-vocabulary disagreement, whereas the background measures the displayed token's probability change, and neither encodes token correctness or an optimizer gradient. All logarithms are natural, so the metrics are in nats and $0 \le d_t \le \log 2$ by Eq. (12). Color limits are fixed across the six figures from the 600-response cohort: approximately ± 1.421 for g_t and 0–0.166 for d_t, with saturation applied only to displayed colors and numerical values left unclipped.

Guidance beyond formatting. We distinguish reasoning text, answer text, and structural tags using the original token spans. For each response i and region κ with token index set $\mathcal{T}_{i,\kappa}$, we first compute the mean of a token metric $\phi_{i,t}$ within that region,

$$\overline{\phi}_{i,\kappa} = \frac{1}{|\mathcal{T}_{i,\kappa}|} \sum_{t \in \mathcal{T}_{i,\kappa}} \phi_{i,t}, \qquad \phi \in \{d, |\Delta|, |g|\}. \tag{33}$$

For the 200 held-out videos per dataset, Table 14 averages the response-level means within each dataset and weights the three datasets equally, preventing longer reasoning traces from dominating the comparison.

Table 14: **Token supervision by response region.** Dataset-balanced means are computed over 600 held-out responses, with log-probability measures in nats. Token share is the mean within-response fraction.

Region	Mean JSD	Mean $\lvert\Delta_t\rvert$	Mean $\lvert g_t\rvert$	Token share
Reasoning text	0.01973	0.2282	0.2021	91.17%
Answer text	0.00479	0.0357	0.0275	0.97%
Format tags	0.00071	0.0065	0.0045	5.78%

Uncertainty is estimated by resampling whole source clusters within each dataset, with 2000 bootstrap replicates. Reasoning text has a higher mean JSD than format tags, with a paired difference of 0.01901 and a 95% percentile interval of $[0.01841, 0.01964]$. The IEMOCAP interval contribution concerns the sampled Session 5 recordings, not variation across sessions. Tokens spanning region boundaries and termination tokens are retained in the share denominator but omitted from the three reported regions, so shares do not sum to 100%.

Context-dependent guidance on affective cues. Figures 7–12 show complete student responses with chronological video frames, transcript, teacher-conditioning excerpts, and numbered token annotations. We select the six illustrative cases after inspecting frozen outputs and source videos for acoustic, visual, and affective reasoning cues; these cases are not a random sample, whereas the statistics above use the full cohort. Teacher references are model-generated conditioning text rather than independent perceptual annotations, and terminal end-of-message tokens are omitted from the figures but retained in the underlying records.

In Figure 7, rationale conditioning favors *playful* in a positive interpretation ($g_t = 0.897$, a probability ratio of 2.45 relative to label-only conditioning). In Figure 8, the same word is suppressed ($g_t = -1.478$, ratio 0.23), while *subdued* is favored within the same incorrect response (excerpted in Figure 6). In the IEMOCAP pair, *lively* and *upbeat* receive higher probabilities in the correct Excited response, whereas the Anger example predicted as Frustration favors *strained* and suppresses *slow* and *deliberate*. The MIntRec 2.0 cases connect posture-related wording to a Complain interpretation and suppress *cheerful* and *smile* in a Greet interpretation of a Criticize example. Across the 14 numbered anchors, g_t and Δ_t have matching signs, although these quantities need not agree in general.

Interpretation. Rationale conditioning can increase the probabilities of reference-consistent cues and decrease those of incompatible wording within an incorrect student response. These changes help interpret the full-vocabulary teacher targets used by the token-normalized JSD in Eq. (6), with a detached teacher target, while g_t is not an additional reward. The diagnostic does not measure training-induced correction or the causal effects of on-policy sampling and EMA, which require the corresponding training comparisons.

G.2 Reasoning Case Studies

We present five illustrative case studies covering intent recognition, emotion recognition, multi-label emotion recognition, fine-grained sentiment, and explainable emotion reasoning. Tables 15–19 pair frames and speech with model outputs and recorded assessments, with OmniOPSD recovering the reference affect in all five cases.

Supportive intent beyond literal wording. In Table 15, the speaker explains that the issue concerns his sister's future rather than money. OmniOPSD connects this statement to the serious interpersonal context and supportive body language, correctly identifying *consoling*. Qwen3.8-Flash also recognizes the reassurance, while Gemini-3.8-Flash treats the utterance as a neutral explanation and GPT-5.6-Luna predicts *suggesting*. One intermediate step in the OmniOPSD response overemphasizes financial matters, although its final prediction matches the reference intent.

Sadness grounded in speech and context. Table 16 concerns a conversation about someone's cancer-related illness. OmniOPSD combines subdued speech, pauses, a lowered gaze, and slumped posture into a consistent interpretation of *sadness*. Gemini-3.8-Flash also identifies sadness from the quiet delivery and illness context, while MiMo-v2.5 and GPT-5.6-Luna predict *happiness*. AffectGPT returns the correct emotion but introduces unsupported visual details in a rejected reasoning step.

Recovering co-occurring emotions. The question "Supposing I don't want flowers?" in Table 17 expresses more than a single negative emotion. OmniOPSD associates forceful delivery and rejection with *anger*, while its facial interpretation supports the accompanying *disgust*. Emotion-LLaMA returns anger alone, while GPT-5.6-Luna and Gemini-3.8-Flash include anger but substitute other categories for disgust, and MiMo-v2.5 shifts toward sadness and anxiety.

Negative sentiment behind a conversational expression. In Table 18, the remark "It was really kind of dumbed down for us" carries criticism despite a relatively conversational facial expression. OmniOPSD links the evaluative wording and subdued delivery to the reference *negative* sentiment. GPT-5.6-Luna instead emphasizes the smile and predicts positive sentiment, while Qwen3.8-Flash predicts weak positive. VideoLLaMA and Gemini-3.8-Flash recover negative polarity but assign strong negative and weak negative, respectively.

Explaining sadness despite a smile. The final case, from the EMER Dataset, evaluates an explanation against annotated emotional clues, following the clue-based perspective described by Cheng et al. (2024a). In Table 19, smiling coexists with longing for an absent person and subdued speech. PandaGPT and Video-ChatGPT misread the smile as positive affect. Both Emotion-LLaMA and OmniOPSD identify sadness, with OmniOPSD linking more vocal and facial evidence and receiving a clue score of 9/10, compared with 8/10 for Emotion-LLaMA. The assessment also notes that OmniOPSD omits tearful eyes and an explicit reference to the absent person.

Gold: **positive** Student: **positive** ✓ **Correct**



Subtitle: Ooh! I'm sorry! I think, I THINK, that may have missed the table!

Teacher-side rationale (excerpt): They open with an upbeat, enthusiastic "Ooh! I'm sorry!" which carries a light, playful tone, particularly highlighted by the upward inflection.

Student response

<think>The speaker's exclamation "Ooh! I'm sorry!" accompanied by a quick, light ❶ tone suggests an energetic and playful ❷ mood, particularly emphasized by the upward ❸ inflection at the end. This upbeat expression indicates a lighthearted and joyful atmosphere. The context of playing ping pong enhances the sense of enjoyment, as the activity often brings a sense of fun and camaraderie. Visually, the person's enthusiastic expression—smiling and engaging in a playful interaction—reflects positive emotions, while the setting of a lively room contributes to a relaxed and joyful ambiance. Taken together, these cues from the speech, vocal delivery, and body language collectively support a **positive** emotional state.</think>

<answer>positive</answer>

Rationale cues

❶ **light** ▲ "light, playful tone"
❷ **playful** ▲ "spontaneous and joyful uncertainty rather than distress"
❸ **upward** ▲ "upward inflection"

Background: rationale gain g_t (nats)

−1.42 (lower) 0 (higher) +1.42

Strip: teacher–student JSD d_t (nats)

0 0.166

Figure 7: **MELD: correct prediction** (`test_dia162_utt0`). During a playful table-tennis apology, rationale conditioning favors *light*, *playful*, and *upward* relative to label-only conditioning.

Gold: **negative** Student: **neutral** ✗ **Incorrect**



Subtitle: Okay, I'm sorry.

Teacher-side rationale (excerpt): The audio suggests a subdued tone with wavering intonation, indicating vulnerability. This emotional intensity is reinforced by the hesitations and pauses in their speech, reflecting uncertainty and regret.

Student response

<think>The speaker's words, "Okay, I'm sorry," are delivered in a subdued ❶ tone, marked by a hint of hesitation. The audio suggests a light, ❷ almost playful ❸ inflection, which contrasts with the tense visual scene of children playing around the woman. The relaxed domestic setting, filled with toys and casual interactions, creates an environment that feels calm and familiar. Despite the initial implication of urgency, the overall atmosphere indicates a moment of relief and understanding. The absence of strong emotional expressions or overt conflict highlights a focus on acceptance, reinforcing the interpretation of a neutral emotional state. The combination of a light and positive delivery, coupled with the relaxed background, supports the conclusion of a neutral emotional experience.</think>

<answer>neutral</answer>

Rationale cues

❶ **subdued** ▲ "subdued tone with wavering intonation"
❷ **light** ▼ "sadness and emotional distress"
❸ **playful** ▼ "uncertainty and regret"

Background: rationale gain g_t (nats)

−1.42 (lower) 0 (higher) +1.42

Strip: teacher–student JSD d_t (nats)

0 0.166

Figure 8: **MELD: incorrect prediction** (`test_dia100_utt5`). For a negative example predicted as neutral, rationale conditioning favors *subdued* while suppressing the reference-inconsistent *light* and *playful* interpretation.

Gold: **Excited** Student: **Excited** ✓ **Correct**



Subtitle: And then the manager came in and found us rolling around on the floor biting and scratching like panthers.

Teacher-side rationale (excerpt): The speaker's vocal tone is vibrant and playful, effectively conveying excitement through a lively intonation and increased pace when describing dynamic actions like "rolling" and "biting."

Student response

<think>The speaker's lively ❶ vocal tone, featuring upbeat ❷ intonation and rhythmic speech, suggests an emotional outburst, particularly focusing on the word "Panthers." This emphasis, combined with the laughter in the audio, hints at an amusing or surprising reaction. The individual's wide smile and animated expressions reinforce a strong emotional engagement, suggesting they are not merely observing but actively participating in the scene. The context of the room, with individuals likely interacting or discussing, further enhances the sense of excitement. The combination of enthusiastic speech, laughter, and expressive facial and body language collectively depicts an individual in a state of delight and excitement.</think>

<answer>Excited</answer>

Rationale cues

❶ **lively** ▲ "lively intonation"

❷ **upbeat** ▲ "upbeat vocal delivery"

Background: rationale gain g_t (nats)

−1.42 (lower) 0 (higher) +1.42

Strip: teacher–student JSD d_t (nats)

0 0.166

Figure 9: **IEMOCAP: correct prediction** (`Ses05F_script03_1_F030`). Rationale conditioning increases the probabilities of *lively* and *upbeat* in the student's explanation of an Excited prediction.

Gold: **Anger** Student: **Frustration** ✗ **Incorrect**



Subtitle: You could go into the computer and- you know- print-

Teacher-side rationale (excerpt): The speaker's voice is strained and fragmented, reflecting a strong emotional distress, particularly anger. The emphasis on "you know" and the pause between "computer" and "print" suggests frustration with the situation being discussed.

Student response

<think>The spoken subtitle, "You could go into the computer and- you know- print," paired with a slow ❶ and deliberate ❷ delivery, particularly highlighting the phrase "you know," indicates a deliberate pause, often associated with frustration. The strained ❸ vocal tone and notable pauses suggest a struggle to articulate thoughts, indicating internal conflict. Visually, the individual's tense posture and clenched fists further reinforce feelings of frustration, while their body language of fidgeting or rubbing hands may signal nervousness or agitation. The formal setting of the interview may heighten emotional intensity, contributing to the perception of frustration rather than a more generalized state. Collectively, these cues—subdued vocal tone, expressive body language, and tense facial expression—point strongly toward a strong feeling of frustration as the person grapples with the task at hand.</think>
<answer>Frustration</answer>

Rationale cues

❶ **slow** ▼ "strained and fragmented"
❷ **deliberate** ▼ "the pause between "computer" and "print""
❸ **strained** ▲ "voice is strained and fragmented"

Background: rationale gain g_t (nats) −1.42 (lower) 0 (higher) +1.42

Strip: teacher–student JSD d_t (nats) 0 0.166

Figure 10: **IEMOCAP: incorrect prediction** (`Ses05F_impro05_F035`). In a response predicting Frustration instead of Anger, rationale conditioning suppresses *slow* and *deliberate* while favoring *strained*.

Gold: **Complain** Student: **Complain** ✓ **Correct**



Subtitle: i worked there over 20 years, and then they just get rid of me like i'm trash?

Teacher-side rationale (excerpt): His body language—slightly hunched and gesturing subtly—indicates a defensive posture, reinforcing the notion of dissatisfaction.

Student response

<think>The speaker's phrase, "I worked there over 20 years, and then they just get rid of me like I'm trash?" is delivered with a sense of disbelief and frustration. The tone conveys a clear sense of anger, emphasizing the negative treatment experienced, as if the situation is unacceptable. In a retail setting, the individual's body language—likely a defensive ❶ stance—suggests they are struggling with this realization. The expression of concern in their eyes and slightly agitated facial features further indicate their distress. The juxtaposition of their formal attire and the environment of a retail space adds context, implying this struggle could relate to their employment. Taken together, these cues—the intense speech, the negative tone, and the visible signs of emotional struggle—formally suggest the intent is to express dissatisfaction and dissatisfaction.</think>

<answer>Complain</answer>

Rationale cues

❶ **defensive** ▲ "defensive posture"

Background: rationale gain g_t (nats)

−1.42 (lower) 0 (higher) +1.42

Strip: teacher–student JSD d_t (nats)

0 0.166

Figure 11: **MIntRec 2.0: correct prediction** (`dia505_utt7`). In a complaint about dismissal, rationale conditioning favors *defensive*, linking posture-related wording to the student's Complain interpretation.

Gold: **Criticize** Student: **Greet** ✗ **Incorrect**



Subtitle: yes! you be selfish.

Teacher-side rationale (excerpt): His vocal delivery suggests a level of sarcasm; the emphasis on "Yes!" could underline an ironic agreement, further compounded by an accusatory tone regarding selfishness.

Student response

<think>The character in the blue vest, identified by their name tag as Jonah, engages with another individual, displaying a cheerful ❶ and animated demeanor. The audio reveals an enthusiastic vocal delivery, particularly the exclamation "yes!" which indicates a positive sentiment. Jonah's smile, ❷ the vibrant and engaging body language, and the friendly interaction within a retail environment all point toward an enthusiastic exchange, suggesting a friendly greeting. The emphasis on certain words aligns with a playful acknowledgment, as if responding to the other individual's statement with a lively affirmation. Together, these cues—happy expression, positive engagement, and expressive delivery—converge to illustrate Jonah's intent to greet in a warm and enthusiastic manner.</think>
<answer>Greet</answer>

Rationale cues

❶ **cheerful** ▼ "a level of sarcasm"
❷ **smile** ▼ "surprise or concern"

Background: rationale gain g_t (nats)

−1.42 (lower) 0 (higher) +1.42

Strip: teacher–student JSD d_t (nats)

0 0.166

Figure 12: **MIntRec 2.0: incorrect prediction** (`dia673_utt3`). For a selfishness remark labeled Criticize but predicted as Greet, rationale conditioning suppresses *cheerful* and *smile* relative to label-only conditioning.

Table 15: **IR: consolation in a serious conversation.**
MME-Emotion example `IR__dia_560_utt_11_ch` has the reference label *consoling*.

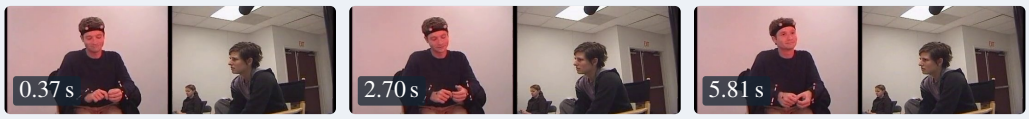


Speech (English translation): "Oh, this isn't about money. It's about how my sister is going to live the rest of her life."

Reference: **consoling**

Model	Reasoning and final prediction	Rating
VideoLLaVA ✓ Correct	**Step 1:** Observe the man eating cake and the woman watching him.	1/1
	Step 2: Interpret her attention as an attempt to provide comfort or support.	1/1
	Step 3 (prediction): The predicted intent is consoling.	1/1
Qwen3.8-Flash ✓ Correct	**Step 1:** Observe the man and woman dining together, with the woman appearing downcast and distressed.	1/1
	Step 2: Note the man's sigh, serious expression, concerned gaze, and earnest tone.	1/1
	Step 3: Interpret his statement about prioritizing his sister's lifelong well-being over money as reassurance and emotional support.	1/1
	Step 4 (prediction): The predicted intent is consoling.	1/1
Gemini-3.8-Flash ✗ Incorrect	**Step 1:** Interpret the speaker's statement about the issue concerning his sister's future.	1/1
	Step 2: Identify the utterance as a plain explanation of his perspective.	0/1
	Step 3: Exclude questioning, consoling, suggesting, encouraging, agreeing, wishing, and acknowledging intents.	0/1
	Step 4 (prediction): The predicted intent is neutral.	0/1
GPT-5.6-Luna ✗ Incorrect	**Step 1:** Interpret the dialogue as concern for the younger sister's future.	1/1
	Step 2: Note the man's earnest delivery and emphatic gestures.	1/1
	Step 3: Identify his attempt to persuade or advise the woman.	1/1
	Step 4 (prediction): The predicted intent is suggesting.	0/1
OmniOPSD (ours) ✓ Correct	**Step 1:** The speaker's words, "This is not about money," convey a sense of urgency and concern, particularly regarding the future of their sister.	1/1
	Step 2: Their vocal tone is described as neutral, with a slight emphasis on "money", indicating a serious discussion about financial matters.	0/1
	Step 3: This aligns with the visual context of a domestic setting, where the speaker's body language - such as leaning forward and maintaining eye contact - suggests a desire to console and support the other person.	1/1
	Step 4: The overall atmosphere, characterized by a serious conversation, further reinforces the intent to console, as the speaker appears to be addressing a significant emotional issue.	1/1
	Step 5: The combination of verbal, vocal, and visual cues leads to a clear interpretation of the speaker's intent as one of consolation.	1/1
	Step 6 (prediction): The predicted intent is consoling.	1/1

Table 16: **ER-Lab: sadness when discussing an illness.**
MME-Emotion example `ER-Lab__Ses05M_impro06_M002` has the reference label *sadness*.



Speech: "Yeah, he's been sick for a while. He had cancer."

Reference: **sadness**

Model	Reasoning and final prediction	Rating
AffectGPT ✓ Correct	**Step 1:** Observe crying, wet makeup, and face-clasping, indicating strong distress.	0/1
	Step 2: Note the shaking voice and monotonous delivery, suggesting fear, anxiety, or emotional detachment.	1/1
	Step 3: Combine the visual and auditory cues to identify the predominant negative emotion.	1/1
	Step 4 (prediction): The predicted emotion is Sadness.	1/1
MiMo-v2.5 ✗ Incorrect	**Step 1:** Analyze the visual cues in the video, focusing on facial expressions and body language of the participants.	1/1
	Step 2: Examine the audio transcript and tone for emotional content.	1/1
	Step 3: Identify any conflict between the semantic content of the speech and the visual expressions.	1/1
	Step 4: Prioritize the displayed facial expression as the primary indicator of emotion.	0/1
	Step 5 (prediction): Based on the sustained smile and relaxed demeanor, the predicted emotion is Happiness.	0/1
Gemini-3.8-Flash ✓ Correct	**Step 1:** Observe visual cues such as downward gaze, nervous fidgeting, and subdued facial expressions indicating grief.	1/1
	Step 2: Analyze auditory cues, noting the quiet, low, and resigned tone of voice when discussing illness.	1/1
	Step 3: Synthesize visual and auditory information with the context of discussing severe illness and death.	1/1
	Step 4 (prediction): The predicted emotion is Sadness.	1/1
GPT-5.6-Luna ✗ Incorrect	**Step 1:** Observe the initial expressions of both subjects, noting that the left subject appears downcast and the right subject looks serious.	1/1
	Step 2: Identify the change in the left subject's expression, specifically the smile and obvious grin at the five-second mark.	0/1
	Step 3: Compare the emotional cues from both subjects, with the left subject showing happiness and the right subject remaining neutral or concerned.	0/1
	Step 4: Synthesize the observations to determine the dominant emotion displayed in the video.	0/1
	Step 5 (prediction): The predicted emotion is happiness.	0/1
OmniOPSD (ours) ✓ Correct	**Step 1:** The speaker's subdued tone and slow speech rate, particularly the emphasis on "sick" and "cancer," indicate a deep sense of sadness and concern.	1/1
	Step 2: The slight pause before mentioning "cancer" suggests a moment of reflection, reinforcing the emotional weight of the statement.	1/1
	Step 3: The individual is seated with a downward gaze and a slumped posture, which are strong indicators of sadness and emotional distress.	1/1
	Step 4: Their body language, characterized by a lack of engagement and a downward gaze, further emphasizes feelings of melancholy.	1/1
	Step 5: The overall scene, set in a simple environment, contributes to the somber atmosphere.	1/1
	Step 6: Collectively, these cues from speech, audio characteristics, and visual expressions converge to suggest a profound sense of sadness.	1/1
	Step 7 (prediction): The predicted emotion is Sadness.	1/1

Table 17: **ML-ER: anger and disgust in a rejecting question.**
MME-Emotion example `ML-ER__00025` has the reference labels *anger, disgust*.

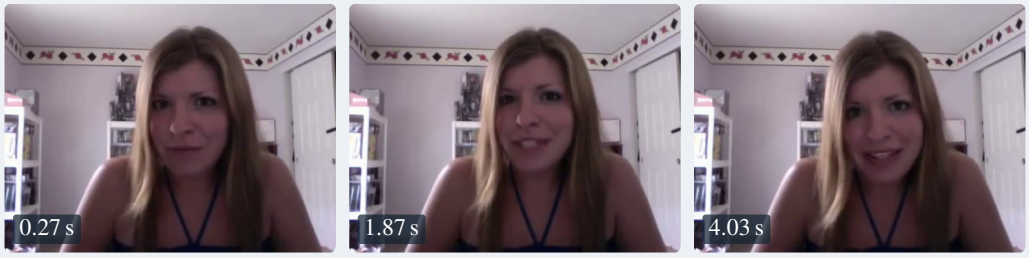


Speech: "Supposing I don't want flowers?"

Reference: **anger, disgust**

Model	Reasoning and final prediction	Rating
Emotion-LLaMA ◐ Partial	*No reasoning returned.* **Step 1 (prediction):** The predicted emotion is anger.	1/2
GPT-5.6-Luna ◐ Partial	**Step 1:** Observe the man's facial expressions, such as furrowed brow, narrowed eyes, and tense mouth. **Step 2:** Note the stern questioning and slightly dismissive look. **Step 3:** Interpret these cues as indicators of specific emotions. **Step 4 (prediction):** The predicted emotions are anger, contempt, and disappointment.	1/1 1/1 1/1 1/2
MiMo-v2.5 ✗ Incorrect	**Step 1:** Observe facial features such as furrowed brows (AU1+AU4), eye focus, and mouth shape. **Step 2:** Relate these features to emotions, mapping them to distress/sadness and worry/anxiety. **Step 3:** Integrate context from the spoken line "Supposing I don't want flowers?" to reinforce disappointment or dissatisfaction. **Step 4:** Match observed cues to candidate emotion labels. **Step 5 (prediction):** The predicted emotions are sadness, anxiety, and disappointment.	1/1 0/1 0/1 0/1 0/2
Gemini-3.8-Flash ◐ Partial	**Step 1:** Identify candidate emotion labels relevant to the video. **Step 2:** Evaluate visual cues such as facial tension, tone, and demeanor. **Step 3:** Infer the emotions conveyed based on observed cues. **Step 4 (prediction):** The predicted emotions are anger and contempt.	1/1 1/1 1/1 1/2
OmniOPSD (ours) ✓ Correct	**Step 1:** The speaker's vocal tone is characterized by a raised pitch and a sharp, forceful delivery, indicating a heightened emotional state. **Step 2:** The emphasis on "supposing" suggests a strong disagreement or frustration, reinforcing the sense of anger. **Step 3:** Visually, the person's furrowed brow and tight-lipped expression, combined with a slight upward curve of the lips, suggest a complex emotional state that could be interpreted as a mix of anger and disgust. **Step 4:** The context of the question "supposing I don't want flowers?" implies a rejection of a suggestion, further supporting the emotional intensity. **Step 5:** The overall body language, including a slight forward tilt of the head, aligns with the emotional cues of anger and disgust. **Step 6:** These consistent signals across speech, facial expressions, and body language converge to indicate a strong emotional response of anger and disgust. **Step 7 (prediction):** The predicted emotions are anger and disgust.	1/1 1/1 1/1 1/1 1/1 1/1 1/1

Table 18: **FG-SA: criticism behind a conversational expression.**
MME-Emotion example `FG-SA__nzpVDcQ0ywM_6` has the reference label *negative*.



Speech: "It was really kind of dumbed down for us"

Reference: **negative**

Model	Reasoning and final prediction	Rating
VideoLLaMA ✗ Incorrect	**Step 1:** Observe the young woman looking at the camera with a scowl.	0/1
	Step 2: Note her serious expression while speaking, reinforcing her unhappiness.	1/1
	Step 3 (prediction): The predicted sentiment is strong negative.	0/1
GPT-5.6-Luna ✗ Incorrect	**Step 1:** Observe the woman's facial expression, noting her slight smile and friendly, engaged demeanor.	0/1
	Step 2: Assess the absence of negative cues and the presence of moderate positive intensity in her expression.	0/1
	Step 3: Synthesize these observations to determine the overall sentiment.	0/1
	Step 4 (prediction): The predicted sentiment is positive.	0/1
Gemini-3.8-Flash ✗ Incorrect	**Step 1:** Analyze the speaker's verbal content for sentiment cues, such as criticism or disappointment.	1/1
	Step 2: Observe the speaker's facial expressions and body language for emotional indicators.	1/1
	Step 3: Integrate verbal and non-verbal cues to assess the overall sentiment intensity.	1/1
	Step 4 (prediction): The predicted sentiment is weak negative.	0/1
Qwen3.8-Flash ✗ Incorrect	**Step 1:** Examine each video frame for facial expressions, focusing on mouth shape, teeth visibility, and cheek movement to distinguish between smiling and talking.	1/1
	Step 2: Assess the presence and intensity of positive cues (e.g., mild smiles, raised cheeks) versus neutral or negative cues (e.g., pressed lips, furrowed brows).	0/1
	Step 3: Evaluate the overall demeanor across frames, considering whether the affect is consistently positive, neutral, or mixed.	0/1
	Step 4: Synthesize the observations to determine the most prominent fine-grained sentiment.	0/1
	Step 5 (prediction): The predicted sentiment is weak positive.	0/1
OmniOPSD (ours) ✓ Correct	**Step 1:** The speaker's subdued tone and slow speech rate, particularly the emphasis on "dumb down," suggest a sense of frustration or disappointment.	1/1
	Step 2: This is further supported by the visual cues: a slight frown and downturned mouth indicate a negative emotional state, while the overall relaxed setting contrasts with the intensity of the feelings expressed.	1/1
	Step 3: The facial expressions, including a frown and lip corner depressor, reinforce the negative sentiment, suggesting a struggle to maintain composure.	1/1
	Step 4: The combination of these vocal, facial, and contextual elements creates a coherent narrative of a person grappling with negative emotions, leading to the conclusion that the character's emotional state is indeed negative.	1/1
	Step 5 (prediction): The predicted sentiment intensity is negative.	1/1

Table 19: **EMER: sadness despite a smile.**
`sample_00002518` (EMER Dataset, 9.80 s) has the reference emotion *sadness*.



Reference: **sadness**

Subtitle (dataset translation): "If he is here, he will definitely support me."

Reference (summary): The speaker misses someone who is absent. Despite smiling, the speaker has tearful, red eyes and speaks slowly, with long pauses and a soft voice, indicating sadness.

Model / score	Response excerpt and clue-level assessment
PandaGPT **2.0/10**	Based on the image, this person's emotional state appears to be positive and confident. The woman is smiling and looking up, suggesting that she feels happy and optimistic. [. . .] *Criteria:* Positivity and confidence contradict the reference sadness. Only the smile overlaps, while tears, red eyes, and long pauses are missed.
Valley **3.0/10**	[. . .] She seems sad and disappointed in some parts of the video, especially when holding her phone and looking at it. [. . .] Overall, the woman's emotional state throughout the video seems to be a mixture of sadness and happiness. *Criteria:* Sadness provides partial overlap. Added disappointment, happiness, and contentment weaken the match to the clearly sad reference state.
Video-ChatGPT **2.0/10**	We can infer that this person feels happy and excited about the support given by the man in the video. He is smiling and looking at the camera, indicating a positive emotional state. *Criteria:* Happiness and excitement oppose sadness from missing someone. The smile is shared, but its emotional meaning is misinterpreted.
Video-LLaMA **1.0/10**	[. . .] Throughout the video, she keeps smiling and giving a speech. [. . .] Overall, the video portrays a confident and elegant woman. *Criteria:* Confidence, elegance, and smiling fail to capture sadness, tears, longing, or subdued affect. Emotional overlap with the reference is minimal.
Emotion-LLaMA **8.0/10**	[. . .] her tone and intonation convey a sense of longing, as she says, if he were here, he would definitely support me [. . .] therefore, it is inferred that his emotional state is sad *Criteria:* Sadness and longing are captured. Tearful or red eyes, slow speech with long pauses, and a soft voice are not described.
OmniOPSD (ours) **9.0/10**	The speaker's tone is notably subdued, with a slow speech rate and a downward inflection on "support," indicating a sense of sadness. [. . .] The person's expression is marked by a downward gaze and a slight frown, which are common indicators of sadness. [. . .] *Criteria:* Sadness, subdued tone, slow speech, and facial cues yield high overlap. Tears, red eyes, and the absent person are not explicitly mentioned.

Scoring (1–10): Summarize the reference and predicted emotional states, compare their overlap, and assign a higher score for greater overlap. Responses are excerpted, with [. . .] marking omissions. Chinese responses and saved assessment comments are translated, and criteria are abridged. Scores apply to complete responses.