SOCIALBENCH: EVALUATING MULTIMODAL PERCEPTION, UNDERSTANDING, AND REASONING IN SOCIAL SETTINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Social understanding requires perceiving behavioral cues, interpreting them in context, and reasoning about implicit intentions. Evaluating only final answers offers an incomplete view of these complementary skills. We introduce SOCIALBENCH, a unified benchmark organizing 16 existing datasets into 30 task variants across three levels: Perception, Understanding, and Theory of Mind. It combines direct assessments of social cues with contextual and inferential tasks across image, video, audio, and text inputs. Evaluation of five multimodal large language models reveals strengths in selected semantic categorization tasks alongside difficulties in fine-grained spatial and temporal prediction and implicit social inference. On Qwen3-Omni, one epoch of task-specific supervised fine-tuning (SFT) substantially improves several initially weak tasks, while joint training brings further gains at the cost of regressions elsewhere. To examine how these improvements relate to internal representations, we train supervised linear predictors on activations extracted before and after SFT. These predictors outperform generated answers on several social tasks, revealing predictive information that the model's responses do not fully reflect. After SFT, generated answers often improve more than probe predictions, narrowing but not consistently eliminating this gap. Thus, neither zero-shot nor fine-tuned responses alone fully characterize the social competencies evaluated here. SOCIALBENCH provides a framework for studying social capabilities through generated performance, supervised adaptation, and linear decodability together.

1 INTRODUCTION

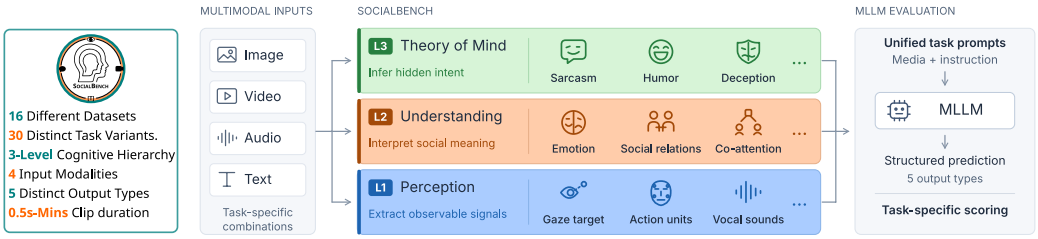


Figure 1: Overview of SOCIALBENCH.

Understanding what people *mean*, rather than just what they say or do, is central to social interaction. AI agents intended to assist, teach, or collaborate with people must therefore account for the social context in which their actions take place. Mathur et al. (2024) identify social intelligence as an important requirement for both virtual and embodied agents, encompassing the ability to perceive, interpret, and respond to human behavior and affect.

Social scene understanding encompasses multiple levels of interpretation, from perceiving observable cues such as gaze, facial movements, and vocal expressions, to interpreting affect and social relationships, and reasoning about intentions in context. Evaluating these abilities separately provides a more

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.

differentiated account of a model's social understanding, revealing where its strengths and limitations lie across these levels.

MLLMs have made substantial progress on broad multimodal benchmarks such as MMBench, MMStar, and MMMU (Liu et al., 2024; Chen et al., 2024; Yue et al., 2024), demonstrating increasingly strong visual understanding and reasoning capabilities. However, how these advances translate across different levels of social understanding is less clear.

Social understanding has often been studied through individual tasks, supported by dedicated datasets such as MuStARD for sarcasm detection, MELD for emotion recognition, and GazeFollow for gaze following, that were studied in isolation (Castro et al., 2019; Poria et al., 2019; Recasens et al., 2015). Recent survey work highlights the need to connect these research areas to support more comprehensive social understanding (Lee et al., 2024). Among recent efforts, HI-ToM (Wu et al., 2023) evaluates higher-order Theory of Mind reasoning through textual scenarios, without requiring models to interpret naturalistic behavioral signals. MoMentS (Villa-Cueva et al., 2025) and Social-IQ 2.0 (Wilf et al., 2023) incorporate video context but primarily assess high-level social inference through end-to-end question answering. This monolithic evaluation provides limited insight into how models perform across the underlying levels of social understanding: an incorrect answer alone does not reveal whether a model also struggles to perceive relevant cues, such as gaze direction or facial movements. HitEmotion (Luo et al., 2026) takes a step toward hierarchical evaluation, but focuses on emotion-related tasks rather than a broader range of social and cognitive abilities.

Taken together, these efforts motivate a unified evaluation of social understanding that spans multiple levels of interpretation, social phenomena, and input modalities. Such an evaluation can reveal where models' strengths and limitations lie across a social hierarchy. To this end, we introduce SOCIALBENCH, a benchmark organized around Perception, Understanding, and Theory of Mind (ToM), as illustrated in Fig. 1.

Our analysis goes beyond zero-shot responses to examine how supervised adaptation changes performance and what task-relevant information can be decoded from model representations. These complementary evaluations distinguish what models generate without task-specific training, what adaptation can improve, and what a supervised linear probe can extract. Our contributions and main findings are threefold:

- **A structured benchmark of social understanding.** SOCIALBENCH brings together 16 datasets and 30 task variants across three levels: Perception, Understanding, and ToM. A common prompting and output protocol, paired with task-specific metrics, supports systematic evaluation across social competencies and modalities.
- **An evaluation of social capabilities and their adaptation.** Zero-shot models perform well on selected semantic categorization tasks but struggle with fine-grained spatial and temporal prediction and implicit social inference. On Qwen3-Omni, task-specific supervised fine-tuning (SFT) substantially improves several initially weak tasks. Joint training further benefits some tasks with limited single-task gains, but reduces performance on others. Output-format controls show that removing the JSON requirement does not generally improve performance on the tested classification tasks.
- **Social information beyond generated answers.** Across multiple social competencies, Qwen3-Omni's internal representations support more accurate predictions than its generated answers reveal. Supervised linear predictors trained on representations extracted before and after SFT outperform the corresponding answers on several social tasks, including physical interaction and social relationship recognition. SFT often narrows this gap without consistently eliminating it. Probing before and after SFT helps separate two effects of adaptation: making task-relevant information more linearly accessible, and improving how the model turns already-accessible information into answers.

  To support further research, we release our training and evaluation code at `https://anonymous.4open.science/r/SocialBench-B10E/README.md`.

2 RELATED WORK

Social benchmarks. Social understanding has traditionally been studied through task-specific datasets covering gaze following (Recasens et al., 2015), facial action units (Ben et al., 2022;

Kollias & Zafeiriou, 2019), speaker diarization (Chung et al., 2020), emotion recognition (Poria et al., 2019; Kosti et al., 2020), social relationships (Li et al., 2017), sarcasm (Castro et al., 2019), humor (Hasan et al., 2019), and deception (Pérez-Rosas et al., 2015). These datasets provide complementary assessments, but do not by themselves establish a common evaluation protocol across social competencies.

Broader benchmarks examine social interpretation and reasoning. HI-ToM (Wu et al., 2023) evaluates higher-order belief reasoning through textual scenarios, while Social-IQ 2.0 (Wilf et al., 2023) and MoMentS (Villa-Cueva et al., 2025) use video question answering to assess social understanding and mental-state inference. However, this monolithic evaluation does not distinguish errors in perceiving social cues from errors in interpreting them, motivating separate assessment of these component skills.

Other benchmarks explicitly structure their evaluation. MimeQA (Li et al., 2025a) examines nonverbal understanding through mime videos, with questions ranging from perceptual grounding to scene-level and global social interpretation. HitEmotion (Luo et al., 2026) organizes affective tasks into recognition, understanding, and reasoning, including humor and sarcasm. OmniBench (Li et al., 2025b) instead focuses on general understanding across visual, acoustic, and textual inputs. SOCIALBENCH brings direct assessments of behavioral cues, contextual interpretation, and implicit social inference into a common framework, spanning classification, regression, and spatial and temporal prediction. This supports comparison of individual social competencies alongside their response to supervised adaptation.

Fine-grained social perception and adaptation. BLINK (Fu et al., 2024) exposes difficulties in visual perception among the VLMs evaluated, including depth estimation, spatial reasoning, and multi-view correspondence. More directly related to social understanding, SocialFusion (Tahboub et al., 2026) reports reduced linear decodability of social features in visual encoders following vision-language pre-training, together with negative transfer in multi-task adaptation of the VLMs studied. Its proposed architecture combines a frozen visual encoder with a lightweight connector and language-model adapters, achieving positive transfer across five social perception tasks. These findings motivate examining both representations and adaptation when evaluating social capabilities. Our study considers a broader set of social prediction targets across image, video, audio, and text inputs, comparing zero-shot, task-specific, and joint adaptation performance.

Internal representations and generated answers. Linear probes assess how well task targets can be predicted from internal representations (Alain & Bengio, 2017), although their performance also depends on the supervised readout and its capacity (Hewitt & Liang, 2019). In VLMs, Fu et al. (2025b) show that visual information can remain accessible within the model despite poor generated-answer performance, and investigate how fine-tuning different components affects this discrepancy. Theodoridis et al. (2026) compare linear probes across the vision encoder, projector, and language model with output predictions on controlled driving scenes. Our analysis examines this relationship across multimodal social capabilities, comparing task-specific linear readouts and generated answers before and after supervised adaptation of Qwen3-Omni. It characterizes which social predictions are accessible from representations and how their performance relates to improvements in the model's generated responses.

3 THE SOCIALBENCH BENCHMARK

SOCIALBENCH brings together 16 existing datasets and 30 task variants to evaluate social understanding in MLLMs across image, video, audio, and text inputs. Table 1 summarizes the tasks, input modalities, evaluation metrics, sample counts, and task-specific reference methods.

3.1 SCOPE AND SOCIAL HIERARCHY

Prior benchmarks have organized multimodal emotional understanding into levels ranging from recognition to contextual interpretation and more complex reasoning (Luo et al., 2026; Hu et al., 2026). We adopt a similar organizing principle for a broader set of social competencies, grouping tasks into three levels: Perception, Understanding, and Reasoning / Theory of Mind. Tasks are

Table 1: **Dataset Overview: A Task-Centric Cognitive Hierarchy.** Tasks are categorized by cognitive depth rather than source dataset. **L1 (Perception)**: Physical signal extraction. **L2 (Understanding)**: Contextual & relational interpretation. **L3 (ToM)**: Inference of hidden mental states. Modality codes: **I**: Image, **V**: Video, **A**: Audio, **T**: Text. †Small test set

Level	Dataset	Task	# Variants	Mod.	Metric	$\lvert\mathcal{D}\rvert$	SOTA
L1 · Perception	GazeFollow (Recasens et al., 2015)	Gaze target detection	1	I	Dist. ↓	4,782	Ryan et al. (2025)
	VAT (Chong et al., 2020)	Att. target localization	2	I / V	Dist. ↓	1,569	Ryan et al. (2025)
	VocalSound (Gong et al., 2022)	Vocal sound classification	1	A	Acc. ↑	3,591	Elizalde et al. (2023)
	VoxConverse (Chung et al., 2020)	Speaker count estimation	1	A	Exact Acc.	10,088	—
	Proxemics (Yang et al., 2012)	Physical distance/touch	2	I	mAP	118	Jiménez-Velasco et al. (2024)
	MMEW (Ben et al., 2022)	Action unit (AU) detection	1	I	macro F1	56	—
	AffWild2 (Kollias & Zafeiriou, 2019)	Action unit (AU) detection	1	V+A	macro F1	2,804	Yu et al. (2026)
		Valence-arousal estimation	1	V+A	CCC ↑	4,340	Ryumina et al. (2026)
	MSP-Podcast (Lotfian & Busso, 2019)	VAD estimation	1	A	CCC ↑	61,116	Lertpetchpun et al. (2025)
L2 · Understanding	VoxConverse (Chung et al., 2020)	Speaker diarization	1	A	DER ↓	10,088	Plaquet & Bredin (2023)
	MMEW (Ben et al., 2022)	Emotion recognition	2	I / V	UAR ↑	236	Fu et al. (2025a)
	AffWild2 (Kollias & Zafeiriou, 2019)	Expression recognition	1	V+A	macro F1	1,028	Sun & Gao (2026)
	MELD (Poria et al., 2019)	Emotion in dialogue	3	V+A+T	wtd. F1	2,602	Dutta et al. (2026)
	EMOTIC (Kosti et al., 2020)	Contextual emotion (cat/VAD)	2	I	mAP / AAE	7,279	Yang et al. (2024)
	PISC (Li et al., 2017)	Social relation classification	1	I	mAP	3,961	Huang et al. (2025)
	MSP-Podcast (Lotfian & Busso, 2019)	Emotion recognition	1	A	macro F1	61,116	Chatzichristodoulou et al. (2025)
	VideoCoAtt (Fan et al., 2018)	Co-attention localization	1	V	L^2 ↓	18,397	Sumer et al. (2020)
		Co-attention detection	1	V	Acc2	35,495	Sumer et al. (2020)
L3 · ToM	MuStARD (Castro et al., 2019)	Sarcasm detection	2	V+A+T	Acc2	138	Guo et al. (2026)
	UR-Funny (Hasan et al., 2019)	Humor detection	2	V+A+T	Acc2	994	Gao et al. (2026)
	RLDD (Pérez-Rosas et al., 2015)†	Deception detection	1	V+A	Acc. ↑	25	Gogate et al. (2017)
	MEVIEW (Husak et al., 2017)†	Concealed emotion recog.	1	V	UAR ↑	20	Liong et al. (2023)

assigned by their prediction target and context, rather than by their source dataset; a single dataset can therefore contribute tasks at multiple levels.

L1 – Perception. Tasks at this level target observable behavioral cues and local affective properties, such as gaze targets, facial action units, vocal sounds, speaker counts, and vocal valence and arousal. They assess the extraction of specific signals from the input.

L2 – Understanding. Tasks at this level target the interpretation of behavior in context or the organization of interactions. They include recognizing emotions, identifying social relationships and shared attention, and tracking who speaks when. Their prediction targets describe affective, relational, or interaction-level structure.

L3 – Reasoning / Theory of Mind. Tasks at this level concern implicit meaning, communicative intent, or concealed affect, including sarcasm, humor, deception, and concealed emotion recognition. We group these tasks as proxies for aspects of social reasoning associated with Theory of Mind; success on them alone does not establish that a model represents another person's mental state.

This hierarchy provides an organizing framework for comparing model performance across levels of social interpretation. It does not assume that models solve tasks through a fixed sequence of stages, nor that an error at one level can be causally attributed to a deficit at another.

3.2 Dataset selection.

We selected datasets according to three criteria: (i) modality coverage, prioritizing audio and video datasets alongside image-based tasks to assess both visual and vocal social cues; (ii) social relevance, focusing on human behavior, affect, relationships, and communicative intent; and (iii) complementary coverage across the three levels of our hierarchy and different prediction types, including classification, regression, and spatial or temporal localization. Where applicable, we define task variants that differ in the available modalities or contextual information, allowing us to examine how these inputs affect performance.

3.3 Unified evaluation format.

We formulate all tasks as prompted generation. Each input combines a task-specific instruction with the relevant media and any additional context, such as a target person's bounding box or preceding dialogue. Models are asked to return a JSON object following a task-specific schema. This common output convention supports deterministic parsing across tasks, while predictions are evaluated using the corresponding task-specific metrics.

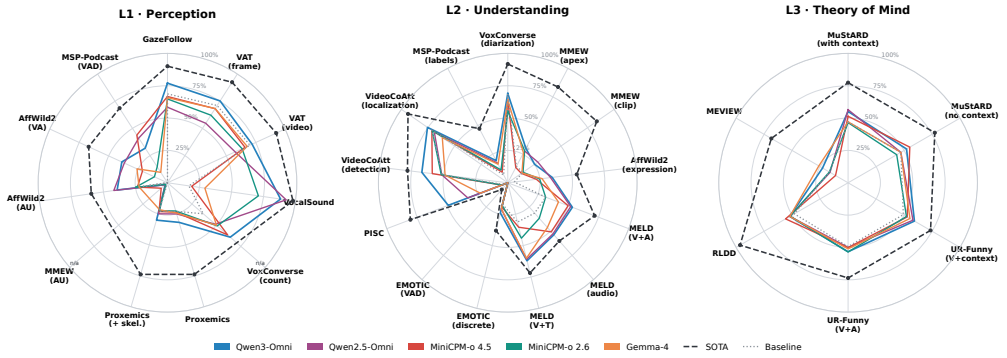


Figure 2: **Zero-shot performance across SOCIALBENCH levels.** Each axis is normalised to $[0, 1]$ (outer ring = theoretical maximum): higher-is-better metrics retain their native scale, while errors e use $b/(b+e)$, where b is the naive baseline error. The dashed line connects available published SOTA values; "n/a" marks tasks without a published SOTA. L1: Perception, L2: Understanding, L3: Theory of Mind.

3.4 OUTPUT TYPES.

The benchmark covers five types of structured predictions:

- **Single-label classification:** one category, as in emotion recognition or sarcasm detection.
- **Multi-label classification:** a set of categories, as in facial action unit detection.
- **Continuous regression:** numerical affect ratings, such as valence and arousal.
- **Spatial localization:** a point or bounding box, as in gaze-target or shared-attention localization.
- **Temporal sequence labeling:** a sequence of speaker labels, as in diarization.

Output conventions, including coordinate scales and temporal discretization, are detailed in Appendix B.2, alongside representative prompt templates and JSON responses.

3.5 MULTIMODAL AND CONTEXTUAL VARIANTS.

Several datasets contribute multiple input configurations or prediction tasks. Input variants change the evidence available to the model: VAT compares individual frames with video clips, while MMEW emotion recognition compares an apex frame with a 16-frame clip. MELD is evaluated with video and audio, audio alone, or video and a transcript. Proxemics is evaluated with and without skeleton information, and MuStARD and UR-Funny with and without preceding dialogue or scene context. Other variants change the prediction target: EMOTIC and MSP-Podcast support categorical emotion recognition and continuous affect estimation, while VoxConverse supports speaker counting and diarization. A summary is provided in Table 2 in the appendix.

4 ZERO-SHOT PERFORMANCE OF EXISTING MLLMS

Base models. We evaluate a representative set of recent *Omni* models; architectures that natively process audio and visual streams alongside text. Large-capacity models include Qwen3-Omni (Xu et al., 2025b) (30B). More compact models include Qwen2.5-Omni (Xu et al., 2025a) (7B), MiniCPM-o 4.5 (Cui et al., 2026) (9B), and MiniCPM-o 2.6 (Yao et al., 2024) (8B). We additionally include Gemma-4 E4B (Gemma Team, 2026) (4.5B effective parameters), a recent dense model from Google that natively processes image and audio inputs, which we evaluate as a timely addition to track progress on SOCIALBENCH against state-of-the-art compact models. We use default generation temperatures and do not apply quantization. We report the zero-shot performance of these models in Figure 2, with detailed per-task results available in Tables 3, 4, and 5 (Appendix C). Below, we discuss these results.

Uneven performance across social tasks. Zero-shot results reveal contrasting performance profiles within each level of SOCIALBENCH (Tables 3–5). At L1, all models exceed the speaker-counting

baseline, while VocalSound accuracy ranges from 18.9% to 92.7% against 16.7%. Spatial and physical cues are more challenging: only Qwen3-Omni improves over baseline on GazeFollow and both VAT variants, and touch detection remains close to baseline for most models. Facial AU detection also varies across datasets: all models exceed baseline on AffWild2, whereas MMEW scores remain low in absolute terms (1.8–18.5% macro F1).

L2 shows a similarly contrasted profile. Most models exceed baseline on dialogue emotion recognition, and both Qwen models and Gemma-4 E4B achieve substantial gains on social relationship classification (22.7–49.2% mAP on PISC against 3.6%). However, MMEW clip emotion recognition remains at or below baseline for three models. The best reported diarization score barely improves on baseline (31.0 DER versus 31.4), and all models underperform the mean-prediction baseline on EMOTIC VAD.

At L3, gains are generally modest. Humor detection with context reaches 52.1–59.2% accuracy against a 49.3% baseline, while deception and concealed-emotion recognition remain limited; their small test sets also warrant caution. Overall, selected vocal, emotional, and relational classifications show clear gains, but these do not extend consistently to spatial localization, temporal speaker assignment, or implicit social inference.

Overall, the results suggest a relative strength in semantic categorization, such as recognizing vocal events, expressed emotions, and social relationships. On the other side, Fine-grained spatial and temporal tasks, such as gaze localization and micro-expression recognition, remain challenging. Despite their semantic nature, ToM-related tasks such as sarcasm, humor, and deception detection generally yield only modest improvements over the naive baselines, with binary classification accuracies ranging from 43.4% to 59.2% across all evaluated models and task variants. This suggests that interpreting implicit social meaning and communicative intent poses additional challenges beyond recognizing explicit emotional expressions.

5 Learning Social Tasks

Having characterized the zero-shot performance of pretrained models, we now investigate how supervised adaptation changes their ability to perform social tasks. We first apply task-specific fine-tuning to assess which zero-shot limitations can be reduced under a limited training budget and which difficulties persist. We then jointly train a single model across tasks to examine whether the gains obtained through individual adaptation can coexist, and where joint training brings additional benefits or compromises. These experiments complement the zero-shot evaluation by characterizing how social-task performance responds to training with supervision.

5.1 Training Protocol

We investigate adaptation in two settings: task-specific fine-tuning, which measures improvements from supervision on individual tasks, and joint fine-tuning, which evaluates performance across tasks within a single adapted model. We use Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025b), which performs strongly across a broad range of tasks in our zero-shot evaluation. We apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to the self-attention projections and cross-modal projectors, keeping the remaining parameters frozen.

For task-specific adaptation, we train for one epoch on each task's full training set, wherever available. MEVIEW has no training examples: its task-specific reference instead uses the RLDD-trained adapter, following a transfer protocol fixed before inspecting MEVIEW results.

This one-epoch setting provides a practical reference for task-specific adaptation under a limited training budget, rather than an estimate of the best achievable performance. We use it to assess how much zero-shot performance can improve with a simple adaptation protocol, without extensive task-specific optimization.

For joint adaptation, we train a single adapter using square-root sampling, with large datasets capped at 50,000 training examples. We selected this sampling strategy by comparing four alternatives at 1,000 gradient steps. Full training settings and sampling ablations are provided in Appendix D.2.

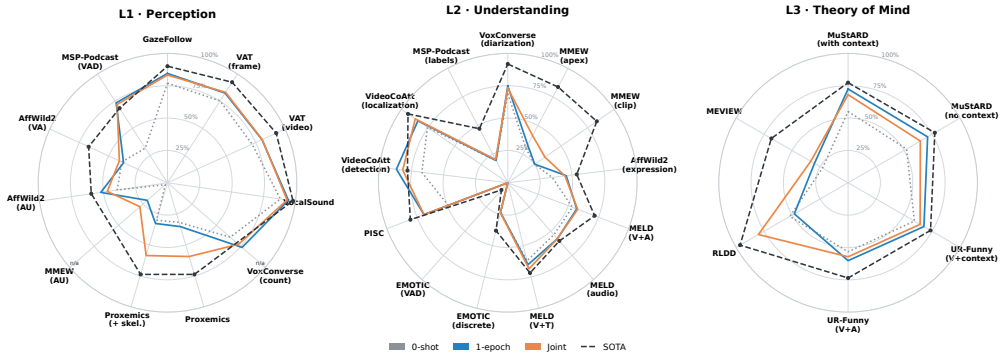


Figure 3: **Single-task vs. joint fine-tuning across SOCIALBENCH levels.** Axes use the normalisation of Figure 2 (outer ring = theoretical maximum). The dashed line connects available published SOTA values; "n/a" marks tasks without a published SOTA. The dotted gray line shows the zero-shot Qwen3-Omni baseline. Per-task numerical results are reported in Table 10 (Appendix D).

5.2 RESULTS AND DISCUSSION

Task-specific adaptation improves diverse social capabilities. Task-specific fine-tuning yields substantial improvements on most of the tasks across all three levels, although the magnitude of these gains varies considerably (Figure 3; Table 10, Appendix D). These improvements span all three levels and extend beyond the semantic categorization tasks that performed well zero-shot. Spatial prediction improves on GazeFollow and VAT, with distance errors decreasing by 0.072–0.084, while MMEW action-unit detection rises from 0.022 to 0.207 macro-F1. Substantial gains also occur in vocal affect estimation (MSP-Podcast VAD: 0.318 to 0.734 CCC), social relationship classification (PISC: 0.492 to 0.693 mAP), and contextual sarcasm detection (MuStARD: 0.551 to 0.725 accuracy). On selected tasks, adapted scores approach or exceed the published reference values: VocalSound reaches 0.944 accuracy against 0.980, while MSP-Podcast VAD reaches 0.734 CCC against 0.685. These references provide context rather than controlled head-to-head comparisons. Overall, low zero-shot scores do not preclude substantial improvements through supervised adaptation.

Limited gains and remaining performance gaps. The benefits of task-specific adaptation vary even among related tasks. While AffWild2 expression recognition improves from 0.346 to 0.455 macro-F1, MMEW clip recognition remains nearly unchanged (0.252 to 0.253 UAR). Similarly, the substantial improvement in MSP-Podcast VAD contrasts with its categorical emotion task (0.193 to 0.195 macro-F1) and with the decrease on AffWild2 valence-arousal estimation (0.388 to 0.373 CCC). These differences show that neither emotion recognition nor continuous prediction benefits uniformly from adaptation. EMOTIC categorical prediction also decreases slightly (0.240 to 0.235 mAP), as does RLDD accuracy (0.520 to 0.480), although the latter has a small test set.

Moreover, improvements do not necessarily close the gap to specialized methods. After single-task adaptation, diarization reaches 0.252 DER against a published reference of 0.082 (lower is better), Proxemics without skeleton information reaches 0.353 versus 0.738 mAP, and MMEW apex recognition reaches 0.331 versus 0.836 UAR. These remaining gaps concern temporal speaker assignment, physical interaction, and subtle facial expressions. They identify tasks that remain challenging under our one-epoch adaptation protocol, without establishing whether the limitations arise from the training budget, input representation, or model architecture.

Joint training benefits tasks with poor single-task gains. Several tasks that benefit little from task-specific adaptation improve substantially under joint training. Proxemics without skeleton information rises from 0.319 mAP zero-shot to only 0.353 after single-task adaptation, but reaches 0.594 with joint training; the skeleton variant follows the same pattern (0.301, 0.328, and 0.586). Likewise, MMEW clip recognition remains nearly unchanged after single-task adaptation (0.252 to 0.253 UAR), yet reaches 0.349 with joint training. Gains also extend to MMEW action-unit detection and apex recognition, suggesting that the benefits span multiple facial prediction targets. RLDD

exhibits an even larger reversal, from 0.480 accuracy after single-task adaptation to 0.800 under joint training, although its small test set limits interpretation.

However, joint training also reduces performance on tasks such as action-unit detection, co-attention detection, and sarcasm recognition. It therefore offers a trade-off rather than a general remedy for the remaining performance gaps.

Cross-task synergies at inference time. We additionally evaluate whether predictions from one task help another when supplied as textual hints at inference time. The protocol and results are reported in Appendix E.

6 Probing for Latent Social Signals

The previous section identifies which social tasks benefit from SFT and which remain challenging, but output scores alone provide limited insight into the changes underlying these effects. We therefore turn to the model's internal representations, using linear probing to assess how well they support task-relevant predictions (Alain & Bengio, 2017; Hewitt & Liang, 2019). By comparing probe and generated-answer performance before and after SFT, we examine how improvements in the model's responses relate to changes in linear decodability.

Probing protocol. We use linear probes to assess how accurately task targets can be predicted from the model's internal representations. A probe is a supervised linear predictor trained on frozen activations, without updating the underlying model. We fit task-specific regularized probes on the training split, separately for the base model and its single-task SFT checkpoint.

We consider representations from multiple layers and groups of input tokens. For each available group (audio tokens, visual tokens (image or video), question tokens, and prompt template tokens) we average the activations across tokens to obtain a single feature vector per example. We also evaluate the last input token's activation alone. The layer and token group are selected by cross-validation within the training split, and the selected configuration is evaluated on held-out test data. We compare probe predictions with generated answers using the same task metric.

This analysis covers 28 task variants. MEVIEW is excluded because it has no training split; speaker diarization would require a separate protocol for temporally structured predictions. Figure 4 shows the 26 variants with a published reference score available for normalization; speaker counting and MMEW action-unit detection are reported separately in the appendix. For higher-is-better metrics, scores are divided by the published reference; for error metrics, the reference error is divided by the measured error. A value of one therefore denotes reference-level performance, and higher values indicate better performance on either axis.

Linear probes reveal a gap with generated answers. Before adaptation, linear probes outperform generated answers on a broad range of tasks across all three levels, including physical interaction, affective prediction, social relationships, and sarcasm recognition. These tasks lie above the diagonal in Figure 4. For example, probe versus output performance reaches 0.575 versus 0.319 mAP on Proxemics without skeleton information and 0.746 versus 0.551 accuracy on contextual sarcasm. Weak zero-shot responses therefore do not necessarily imply that task-relevant information is unavailable in the model's representations: a supervised linear readout can often extract more accurate predictions.

SFT improves generated answers and linear decodability. The trajectories distinguish two patterns behind the adaptation gains reported in Section 5. Several tasks with strong initial probe performance move predominantly horizontally: generated answers improve with little change in linear decodability. For example, VocalSound output accuracy increases from 0.882 to 0.944 while probe accuracy changes from 0.945 to 0.954. This pattern is consistent with better use of information already accessible to a supervised linear readout. Other tasks, notably gaze localization, start with weaker performance on both axes and improve in both probe and output scores. For instance, on GazeFollow, probe distance error decreases from 0.239 to 0.172 and output error from 0.230 to 0.155. Adaptation thus combines improvements in linear decodability with output gains that can substantially exceed changes in probe performance.

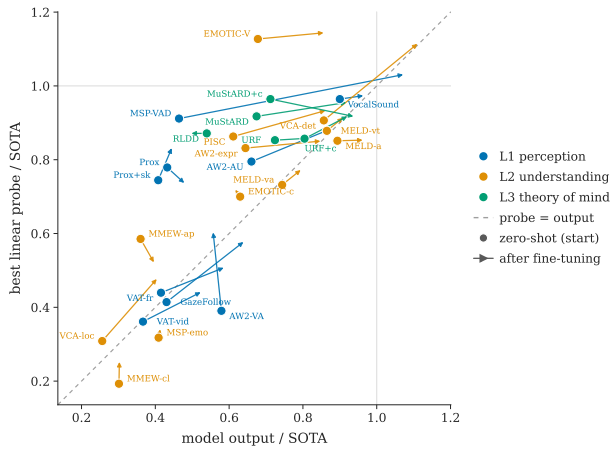


Figure 4: **Linear decodability versus generated-answer performance.** Each arrow is a task, moving from zero-shot to single-task SFT: the horizontal axis is the model's output score, and the vertical axis is the best linear-probe score, both expressed relative to the published SOTA (with lower-is-better metrics inverted). Colour identifies the cognitive level, and the diagonal marks equal probe and output scores. Only the 26 tasks with a published SOTA are shown; separate zero-shot and SFT panels are given in Figure 14 (Appendix F).

SFT narrows, but does not eliminate, probe–output gaps. For several tasks, output performance improves more than probe performance, bringing their trajectories closer to the equal-performance diagonal. PISC provides one example: its probe–output gap shrinks from 0.203 to 0.058 mAP. Across the plotted tasks, probe and output scores also become more strongly correlated after SFT (Pearson's r: 0.73 to 0.84; Spearman's ρ: 0.72 to 0.84), indicating a stronger association between the two measures across evaluated tasks. Nevertheless, substantial gaps persist on some tasks. For example, on Proxemics without skeleton information, the adapted probe reaches 0.544 mAP against 0.353 for generated answers. Under our training protocol, SFT therefore reduces several initial discrepancies, while leaving some generated-answer scores below those achieved by supervised linear readouts. Per-task probe and generated-answer scores, their differences, and the selected probe configurations are reported in Appendix F.

7 Conclusions

We introduced SocialBench, a unified benchmark spanning social perception, understanding, and reasoning, to examine MLLMs through their generated answers, their response to supervised adaptation, and the predictions accessible from their internal representations. Our evaluation reveals strengths in selected semantic categorization tasks alongside persistent difficulties in fine-grained spatial and temporal prediction and implicit social inference. Task-specific SFT substantially improves several initially weak tasks, while joint training brings additional gains at the cost of regressions elsewhere.

Crucially, neither zero-shot nor fine-tuned answers alone fully characterize the social competencies evaluated here. Our probing results suggest that fine-tuning can improve social-task performance through two distinct effects: making task-relevant information more linearly accessible, and improving the model's ability to transform already-accessible information into task outputs. Probing only the pretrained model cannot distinguish these effects; probing before and after adaptation tracks how decodability and generated performance co-evolve. On Qwen3-Omni, SFT often narrows the probe–output gap without consistently eliminating it. These findings motivate evaluating generated performance, adaptability, and linear decodability together. By bringing these complementary perspectives into a shared framework, SocialBench provides a foundation for studying how social information is represented and expressed by MLLMs, and for measuring progress across distinct social capabilities.

REPRODUCIBILITY STATEMENT

All 16 datasets aggregated in SOCIALBENCH are publicly available; their statistics and task variants are described in Section 3 and Appendix B, and the output conventions and prompt templates, including the JSON output schemas, are given in Appendix B.2. Zero-shot evaluation uses publicly released model checkpoints with default generation temperatures and no quantization (Section 4); the effects of the output-format instruction, the response parser, and the number of sampled frames are analysed in Appendices C.1 and C.2. Fine-tuning hyperparameters, hardware, random seed, compute budget, and the joint sampling strategy are reported in Appendix D.2 (Table 11), with a LoRA sensitivity analysis in Appendix D.3. The probing protocol is described in Section 6, and 95% bootstrap confidence intervals (1,000 resamples, seed 42) for the zero-shot and fine-tuning results are provided in the supplementary material. Our evaluation and data preprocessing code is available at `https://anonymous.4open.science/r/SocialBench-B10E/README.md`.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics.

Human subjects and data. This work does not involve new data collection or human annotation. SOCIALBENCH aggregates 16 previously published datasets, used under the terms and licenses set by their original authors. Our released code provides the means to download each dataset from its original source and to preprocess it into the SOCIALBENCH format; we do not redistribute datasets with restricted access, which remain obtainable only through their original access procedures. These datasets contain faces, voices, and interactions of real people, and we rely on the consent and ethical procedures documented by their creators.

Potential misuse. SOCIALBENCH measures capabilities such as gaze following, social relation recognition, emotion recognition, and deception detection (RLDD, MEVIEW). Such capabilities could be misused to monitor attention or interpersonal dynamics without consent, or to judge truthfulness in legal, forensic, or hiring contexts. Our results show that even fine-tuned generalist models remain well below specialized systems on several of these tasks, and we caution against any such high-stakes deployment. We intend SOCIALBENCH as a diagnostic tool to expose model limitations, not as a validation of social-inference systems for deployment.

Bias and fairness. Affective and social datasets may encode demographic and cultural biases in how emotions and social signals are expressed and labelled. SOCIALBENCH inherits these biases, and aggregate scores may hide disparities across groups; high benchmark performance should therefore not be read as evidence of fair behaviour across populations.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and to provide feedback on some implementation choices of our experiments, through coding agents that assisted in writing our code. We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, design our research methodology or experiments, clean and reformat datasets, or interpret results; formulating or proving mathematical claims, translation, and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft, rewrite, and proofread parts of the paper, to create analysis figures, and to search for references. We have reviewed all AI-assisted work: the authors reviewed the code and proofread the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL `https://openreview.net/forum?id=ryF7rTqgl`.

Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5826–5846, 2022. doi: 10.1109/TPAMI.2021.3067464.

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

Georgios Chatzichristodoulou, Despoina Kosmopoulou, Antonios Kritikos, Anastasia Poulopoulou, Efthymios Georgiou, Athanasios Katsamanis, Vassilis Katsouros, and Alexandros Potamianos. MEDUSA: A multimodal deep fusion multi-stage training framework for speech emotion recognition in naturalistic conditions. In *Proc. Interspeech 2025*, pp. 4683–4687, 2025. doi: 10.21437/Interspeech.2025-2636.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024. doi: 10.52202/079017-0850.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5395–5405, 2020. doi: 10.1109/CVPR42600.2020.00544.

Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. Spot the conversation: Speaker diarisation in the wild. In *Proc. Interspeech 2020*, pp. 299–303, 2020. doi: 10.21437/Interspeech.2020-2337.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL `https://arxiv.org/abs/2604.27393`.

Soumya Dutta, Smruthi Balaji, and Sriram Ganapathy. A mixture-of-experts model for multimodal emotion recognition in conversations. *Computer Speech & Language*, 100:101965, 2026. doi: 10.1016/j.csl.2026.101965.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6460–6468, 2018. doi: 10.1109/CVPR.2018.00676.

Liangyu Fu, Xuecheng Wu, Danlei Huang, and Xinyi Yin. FAMNet: Integrating 2D and 3D features for micro-expression recognition via multi-task learning and hierarchical attention. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025a. doi: 10.1109/IJCNN64981.2025.11228699.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling*, 2025b. URL `https://openreview.net/forum?id=qQb1JLrwol`.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, pp. 148–166. Springer, 2024. doi: 10.1007/978-3-031-73337-6_9.

Manning Gao, Leheng Zhang, Shiqin Han, Haifeng Hu, Yuncheng Jiang, and Sijie Mai. GRCF: Two-stage groupwise ranking and calibration framework for multimodal sentiment analysis, 2026. URL https://arxiv.org/abs/2601.09606.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Mandar Gogate, Ahsan Adeel, and Amir Hussain. Deep learning driven multimodal fusion for automated deception detection. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6. IEEE, 2017. doi: 10.1109/SSCI.2017.8285382.

Yuan Gong, Jin Yu, and James Glass. VocalSound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155, 2022. doi: 10.1109/ICASSP43922.2022.9746828.

Diandian Guo, Fangfang Yuan, Cong Cao, Xixun Lin, Chuan Zhou, Hao Peng, Yanan Cao, and Yanbing Liu. MuVaC: A variational causal framework for multimodal sarcasm understanding in dialogues. In *Proceedings of the ACM Web Conference 2026*, pp. 7318–7329, 2026. doi: 10.1145/3774904.3792491.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2046–2056, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1211.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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

Dejiao Huang, Menglei Xia, Ruyi Chang, Xiaohan Kong, and Shuai Guo. GT-SRR: A structured method for social relation recognition with GGNN-based transformer. *Sensors*, 25(10):2992, 2025. doi: 10.3390/s25102992.

Petr Husak, Jan Cech, and Jiri Matas. Spotting facial micro-expressions "in the wild". In *Proc. Computer Vision Winter Workshop (CVWW)*, 2017. https://cmp.felk.cvut.cz/~cechj/ME/.

Isabel Jiménez-Velasco, Jorge Zafra-Palma, Rafael Muñoz-Salinas, and Manuel J. Marín-Jiménez. Proxemics-Net++: Classification of human interactions in still images. *Pattern Analysis and Applications*, 27(2):49, 2024. doi: 10.1007/s10044-024-01270-3.

Dimitrios Kollias and Stefanos Zafeiriou. Aff-Wild2: Extending the Aff-Wild database for affect recognition, 2019. URL https://arxiv.org/abs/1811.07770.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using EMOTIC dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2020. doi: 10.1109/TPAMI.2019.2916866.

Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, and James M. Rehg. Towards social AI: A survey on understanding social interactions, 2024. URL https://arxiv.org/abs/2409.15316.

Thanathai Lertpetchpun, Tiantian Feng, Dani Byrd, and Shrikanth Narayanan. Developing a high-performance framework for speech emotion recognition in naturalistic conditions challenge for emotional attribute prediction. In *Proc. Interspeech 2025*, pp. 4648–4652, 2025. doi: 10.21437/Interspeech.2025-1082.

Hengzhi Li, Megan Tjandrasuwita, Yi R. Fung, Armando Solar-Lezama, and Paul Pu Liang. MimeQA: Towards socially-intelligent nonverbal foundation models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 157031–157059, 2025a. doi: 10.52202/085713-4733.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Dual-glance model for deciphering social relationships. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2669–2678, 2017. doi: 10.1109/ICCV.2017.289.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Ruibo Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards the future of universal omni-language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 34822–34851, 2025b. doi: 10.52202/085713-1036.

Gen-Bing Liong, John See, and Chee-Seng Chan. Spot-then-recognize: A micro-expression analysis network for seamless evaluation of long videos. *Signal Processing: Image Communication*, 110: 116875, 2023. doi: 10.1016/j.image.2022.116875.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, pp. 216–233. Springer, 2024. doi: 10.1007/978-3-031-72658-3_13.

Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2019. doi: 10.1109/TAFFC.2017.2736999.

Meng Luo, Bobo Li, Shanqing Xu, Shize Zhang, Qiuchan Chen, Menglu Han, Wenhao Chen, Yanxiang Huang, Hao Fei, Mong-Li Lee, and Wynne Hsu. Unveiling the cognitive compass: Theory-of-mind–guided multimodal emotion reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8VSrk2CaBr.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in AI agents: Technical challenges and open questions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20541–20560, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1143. URL https://aclanthology.org/2024.emnlp-main.1143/.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI)*, pp. 59–66. Association for Computing Machinery, 2015. doi: 10.1145/2818346.2820758.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. Interspeech 2023*, pp. 3222–3226, 2023. doi: 10.21437/Interspeech.2023-205.

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 (ACL)*, pp. 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1050.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020. doi: 10.1109/SC41405.2020.00024.

Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/ec8956637a99787bd197eacd77acce5e-Paper.pdf.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. Gaze-LLE: Gaze target estimation via large-scale learned encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28874–28884, 2025. doi: 10.1109/CVPR52734.2025.02689.

Elena Ryumina, Maxim Markitantov, Alexandr Axyonov, Dmitry Ryumin, Mikhail Dolgushin, Denis Dresvyanskiy, and Alexey Karpov. Team RAS in 10th ABAW competition: Multimodal valence and arousal estimation approach, 2026. URL https://arxiv.org/abs/2603.13056.

Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3316–3325. IEEE, 2020. doi: 10.1109/WACV45572.2020.9093515.

Jiajun Sun and Zhe Gao. A two-stage dual-modality model for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 10th ABAW Workshop*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/ABAW/papers/Sun_A_Two-Stage_Dual-Modality_Model_for_Facial_Expression_Recognition_CVPRW_2026_paper.pdf.

Hamza Tahboub, Weiyan Shi, Gang Hua, and Huaizu Jiang. SocialFusion: Addressing social degradation in pre-trained vision-language models. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=ofYhEoKIEx.

Nikos Theodoridis, Reenu Mohandas, Ganesh Sistu, Anthony Scanlan, Ciarán Eising, and Tim Brophy. Probing visual concepts in lightweight vision-language models for automated driving. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=HlBBy19ojC.

Emilio Villa-Cueva, S M Masrur Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. MoMentS: A comprehensive multimodal benchmark for theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22591–22611, 2025. doi: 10.18653/v1/2025.findings-emnlp.1230.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-IQ 2.0 challenge: Benchmarking multimodal social understanding. GitHub repository, https://github.com/abwilf/Social-IQ-2.0-Challenge, 2023.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706, 2023. doi: 10.18653/v1/2023.findings-emnlp.717.

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, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12447–12457, 2024. doi: 10.1109/CVPR52733.2024.01183.

Yi Yang, Simon Baker, Anitha Kannan, and Deva Ramanan. Recognizing proxemics in personal photos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3522–3529, 2012. doi: 10.1109/CVPR.2012.6248095.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM-V: A GPT-4V level MLLM on your phone, 2024. URL https://arxiv.org/abs/2408.01800.

Jun Yu, Yunxiang Zhang, Naixiang Zheng, Lingsi Zhu, and Guoyuan Wang. Hierarchical granularity alignment and state space modeling for robust multimodal AU detection in the wild, 2026. URL https://arxiv.org/abs/2603.11306.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025. doi: 10.1609/aaai.v39i28.35383.

A Appendix Outline

This appendix provides supplementary material organized into five sections.

Appendix B — Dataset Descriptions and Prompt Templates details the evaluation setup for each of the 16 datasets in SocialBench, including dataset statistics (Section B.1 and Figures 5, 6), output conventions and representative prompt examples with their expected JSON output formats (Section B.2), and a summary of the input and task variants per dataset (Table 2).

Appendix C — Zero-Shot Performance: Extended Results reports the full per-task, per-model zero-shot point estimates across all three cognitive levels (Tables 3, 4, 5), complementing the radar chart in Figure 2. The corresponding 95% bootstrap confidence intervals are provided in the separate supplementary material. This appendix also includes JSON format compliance rates (Table 6), an ablation of the output-format instruction and parser (Section C.1), and an ablation of the number of sampled video frames (Section C.2).

Appendix D — Learning Social Tasks: Extended Results provides the full numerical comparison of zero-shot, single-task fine-tuning, and joint fine-tuning against SOTA (Table 10), a per-task visualization of joint vs. single-task delta gains (Figure 9), and detailed training hyperparameters, compute budget definition, and sampling strategy ablation (Section D.2, Table 13). A LoRA hyperparameter sensitivity analysis is reported in Section D.3.

Appendix E — Understanding Cross-task Relationships at Inference Time complements the joint-training analysis of Section 5 with an inference-time study: predictions of the joint model on a source task are injected as textual hints into the prompt of a target task. It describes the protocol, summarises the relative-gain matrix (Figure 10), discusses the results, and presents the full hint transfer matrices broken down by cognitive level (Figures 11, 12, 13).

Appendix F — Probing Results by Task shows the zero-shot and fine-tuned probe–output scores as separate panels (Figure 14), reports the best probe type and score for each task alongside the corresponding model output and their difference (Tables 15, 16), and gives the absolute probe and output scores by cognitive level (Section F.1).

B Dataset Descriptions and Prompt Templates

B.1 Dataset Statistics

Figure 5 shows the distribution of evaluation samples across tasks within each cognitive level, capped at 3,000 samples per task. Figure 6 jointly reports training set sizes, average clip durations for temporal tasks, and the overall sample distribution across cognitive levels. The pronounced imbalance in training set sizes (Figure 6a, from 96 samples for RLDD to 216k for VideoCoAtt; MEVIEW has zero training examples) directly motivates the sampling strategy ablation in Appendix D.2.

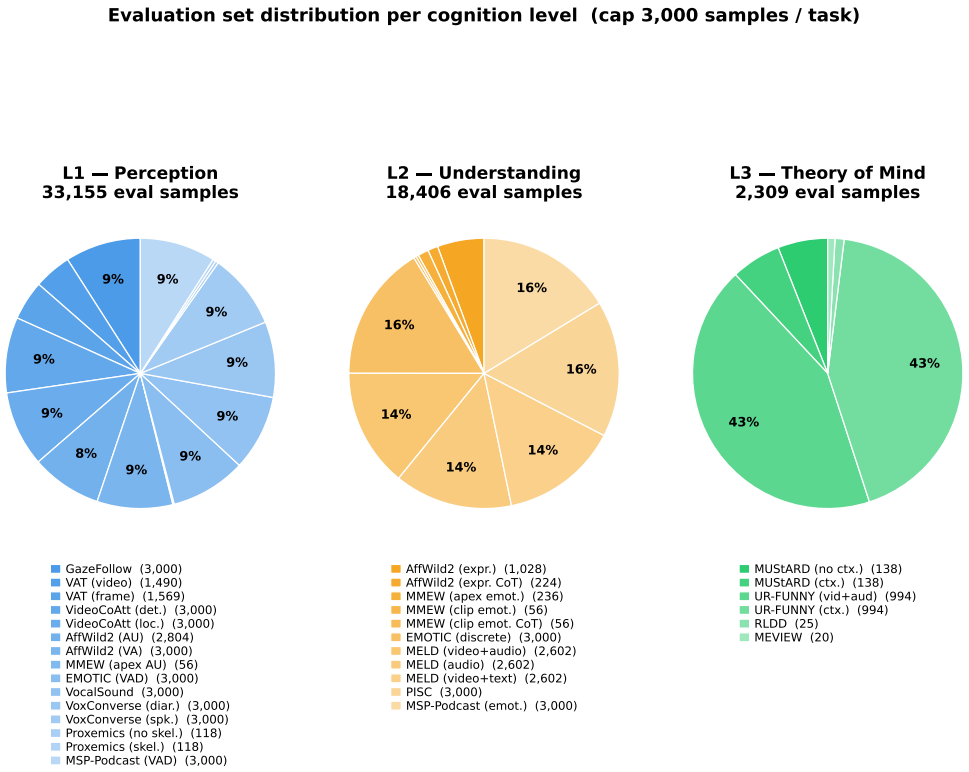


Figure 5: **Evaluation set distribution per cognitive level.** Each pie shows the share of evaluation samples (capped at 3,000 per task) contributed by each task within the level. All task variants are shown individually.

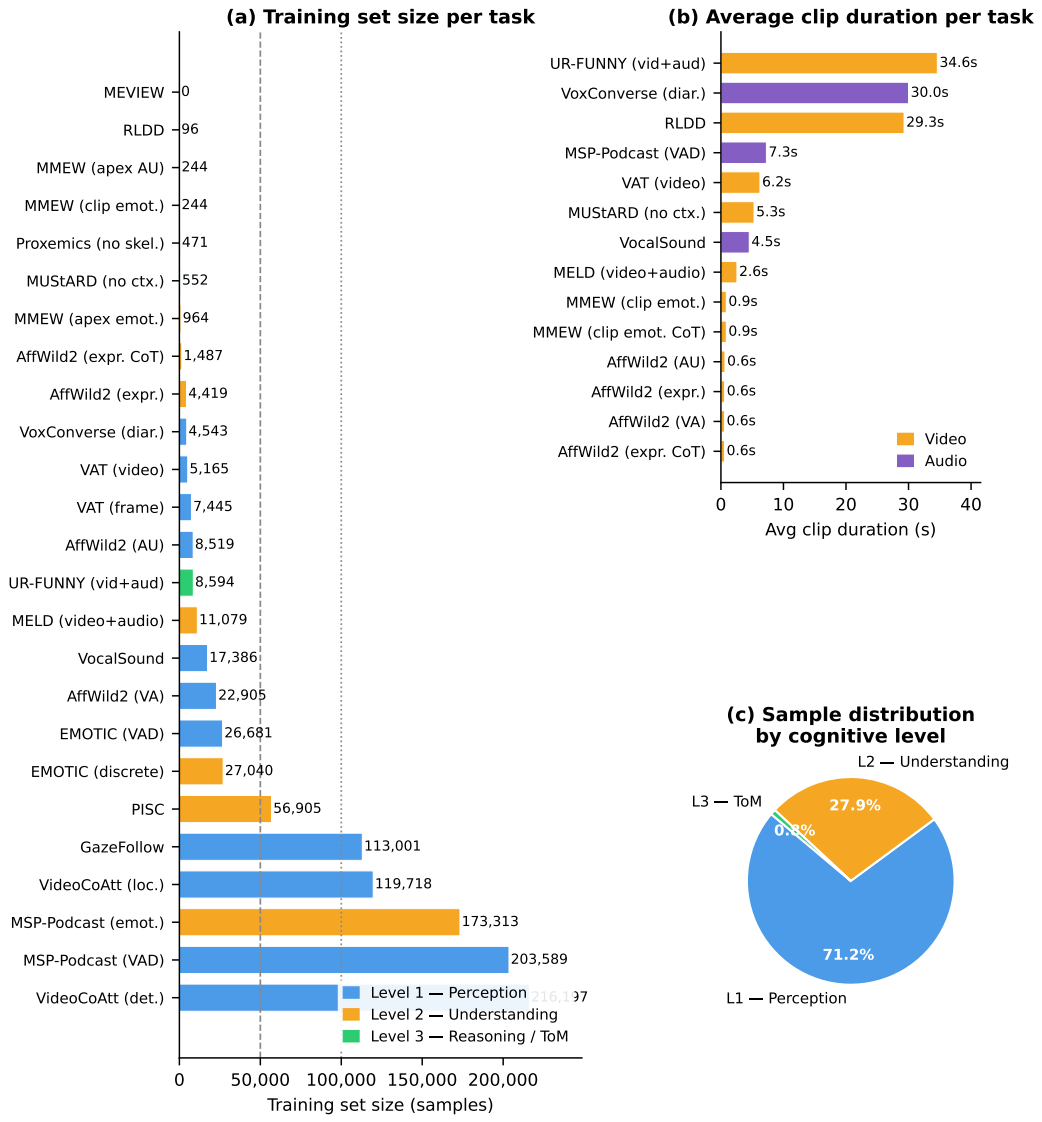


Figure 6: **Dataset statistics.** **(a)** Training set size per task; dashed lines at 50k and 100k samples). **(b)** Average clip duration per task for temporal (video/audio) datasets, spanning from 0.6 s (AffWild2) to 34.6 s (UR-FUNNY). **(c)** Total sample distribution across the three cognitive levels (within-level duplicates deduplicated).

B.2 Output Conventions and Prompt Examples

All tasks use task-specific JSON schemas covering five output types: single-label classification, multi-label classification, continuous regression, spatial localization, and temporal sequence labeling. Each input consists of a *system* message specifying the task and required output schema, and a *user* message containing the media, any additional context, and the question. Below, we describe the output conventions and illustrate the unified prompt-response format through examples from four of these types.

Spatial localisation — GazeFollow (L1). The model receives an image and the bounding box of a person's head; it must predict the 2-D gaze target. All spatial coordinates are normalized to $[0, 1000]$ following the native representation of Qwen vision encoders. The out-of-frame case is handled via a dedicated boolean key.

> **[System]** You are an expert in gaze estimation. Given an image and the bounding box of a person's head, predict where they are looking. Bounding box coordinates are normalized to [0, 1000] in (x1, y1, x2, y2) format. Provide your answer as a valid JSON object: {"gaze_point": [x, y]} or {"out_of_frame": true} if the target lies outside the image.
>
> **[User]** <image>
> Person's head bounding box: [312, 87, 398, 195]
> Where is this person looking?
>
> **[Expected output]** {"gaze_point": [541, 302]}

Binary classification — MuStARD sarcasm detection (L3). The model receives a video clip and its audio track, optionally preceded by the dialogue context; it must output a boolean judgment.

> **[System]** You are an expert at detecting sarcasm in multimodal communication. Given a video clip and audio recording of a speaker, along with the preceding dialogue context, determine whether their utterance is sarcastic. Provide your answer as a valid JSON object: {"sarcasm": true} or {"sarcasm": false}.
>
> **[User]** <video>
> Dialogue context: "Did you enjoy the meeting?" / "Oh, it was the highlight of my week."
> Is this utterance sarcastic?
>
> **[Expected output]** {"sarcasm": true}

Continuous regression — AffWild2 valence-arousal (L1). The model receives a 16-frame video clip and predicts two continuous affective dimensions at the center frame.

> **[System]** You are an expert in affective computing. Given a 16-frame video clip, predict the valence and arousal at the center frame. Both values are continuous in [-1, 1]. Provide your answer as a valid JSON object: {"valence": x.xxx, "arousal": x.xxx}.
>
> **[User]** <video>
> What are the valence and arousal values at the center frame of this clip?
>
> **[Expected output]** {"valence": 0.412, "arousal": -0.183}

Multi-label classification — MMEW action unit detection (L1). The model receives a single apex-frame image and must identify which facial action units are active; the output list may be empty.

> **[System]** You are an expert in facial action unit analysis. Given an apex frame of a micro-expression, identify the

```
active facial action units.  Provide your answer as a
valid JSON object:  {"action_units": ["AU6", "AU12"]} or
{"action_units": []} if none are active.

[User] <image>
Which facial action units are active in this frame?

[Expected output] {"action_units": ["AU1", "AU4", "AU15"]}
```

B.3 TASK VARIANTS

We distinguish input variants, which change the evidence available to the model, from task variants, which change the prediction target. Table 2 summarizes the main configurations.

Table 2: **Dataset Variants Summary by Cognitive Level**

Level	Dataset	Task	Variants
L1	VAT (Chong et al., 2020)	Att. target localization	*image, video*
L1	Proxemics (Yang et al., 2012)	Physical distance/touch	*w/ & w/o skeleton*
L2	MMEW (Ben et al., 2022)	Emotion recognition	*apex, clip*
L2	MELD (Poria et al., 2019)	Emotion in dialogue	*video+audio, audio only, video+trans.*
L2	EMOTIC (Kosti et al., 2020)	Contextual emotion (cat/VAD)	*discrete, VAD*
L3	MuStARD (Castro et al., 2019)	Sarcasm detection	*w/ & w/o context*
L3	UR-Funny (Hasan et al., 2019)	Humor detection	*w/ & w/o context*

C ZERO-SHOT PERFORMANCE OF EXISTING MLLMS: EXTENDED RESULTS

Tables 3, 4, and 5 report the full per-task, per-model zero-shot point estimates for L1 (Perception), L2 (Understanding), and L3 (Theory of Mind) respectively, complementing the radar chart in Figure 2. Each table includes a naive statistical baseline (majority class for classification, training-set mean for regression, center point for localization) and the published SOTA as reference points. Entries marked "–" indicate tasks for which a model produced no parseable output or does not natively support the required input modality. The corresponding 95% bootstrap confidence intervals appear in the separate supplementary material.

The results complement Section 4: across all three levels, MLLMs consistently fall short of task-specific SOTA methods, with failures concentrated on tasks requiring fine-grained temporal or spatial perception (e.g., gaze detection, micro-expression recognition). Stronger zero-shot results on VocalSound, PISC, and MELD video+transcript involve broad semantic or acoustic patterns that large pretrained models already capture well.

Table 3: **L1 Zero-shot results — Perception.** Model cells report point estimates; confidence intervals appear in the separate supplementary material. F1, Acc., and mAP values are in %; Dist. and CCC are raw scores. **Baseline**: naive predictor (majority class / mean prediction). **SOTA**: best published task-specific model. When a model produces no parseable output for a sample, the baseline prediction is substituted for metric computation. †Small test set (< 100 samples).

Dataset	Task / Variant	Metric	*Baseline*	*Qwen3-Omni*	*Qwen2.5-Omni*	*MiniCPM-o 4.5*	*MiniCPM-o 2.6*	*Gemma-4 E4B*	*SOTA*
GazeFollow	Gaze target detection	Dist. ↓	0.313	0.230	0.416	0.333	0.352	0.342	**0.099**
VAT	Att. target (frame)	Dist. ↓	0.290	0.248	0.453	0.319	0.379	0.318	**0.103**
	Att. target (video)	Dist. ↓	0.300	0.281	0.425	0.341	0.366	0.323	**0.103**
VocalSound	Vocal sound classification	Acc. ↑	16.7	88.3	92.7	18.9	71.0	29.2	**98.0**
VoxConverse	Speaker count estimation	Exact Acc.	36.6	64.2	48.9	61.5	50.9	50.1	—
Proxemics†	Touch detection	mAP	23.0	31.9	25.1	23.7	22.7	24.4	**73.8**
	Touch detection (+skel.)	mAP	23.0	30.1	25.0	23.2	22.5	21.7	**73.8**
MMEW†	AU detection (apex)	macro F1	0.0	2.2	2.3	6.4	1.8	18.5	—
AffWild2	Action unit detection	macro F1	18.9	39.3	41.9	23.5	25.1	22.2	**59.5**
	Valence-arousal estimation	CCC ↑	0.000	0.388	0.367	0.213	0.111	0.259	**0.658**
MSP-Podcast	VAD estimation	CCC ↑	0.000	0.318	0.411	0.438	0.147	0.096	0.685

Table 6 reports format compliance rates across all tasks and models. Qwen3-Omni and Gemma-4 E4B achieve near-perfect compliance (97% and 96% respectively), while Qwen2.5-Omni is also reliable at 92%. By contrast, MiniCPM-o 4.5 and MiniCPM-o 2.6 show substantially lower average compliance (68% and 60%), with critical failures on several tasks: MiniCPM-o 4.5 produces parseable output on fewer than 5% of Proxemics touch-detection and PISC samples, and MiniCPM-o 2.6 produces no valid output at all on MELD video+audio. Qwen2.5-Omni, despite its high average, produces no valid output on VoxConverse diarization. The compliance issues are not uniform across task types: spatial localization (VAT), touch detection (Proxemics), a multi-label classification task grounded in spatial understanding, and tasks requiring sequence output (diarization) show the lowest compliance rates for most models, reflecting the difficulty of generating structured coordinate, label-list, or temporal-sequence JSON compared to single-label outputs. Scores from low-compliance model–task combinations should therefore be interpreted with caution.

Table 4: **L2 Zero-shot results — Understanding.** Model cells report point estimates; confidence intervals appear in the separate supplementary material. F1, Acc., UAR, and mAP values are in %; AAE, DER, and L^2 are raw scores. **Baseline**: naive predictor. **SOTA**: best published task-specific model. When a model produces no parseable output for a sample, the baseline prediction is substituted for metric computation. AffWild2 expression recognition uses the label-only prompt on the same 1,028 validation clips for all models. VideoCoAtt localization SOTA is 62.84 pixels on a 1,000-unit coordinate scale; we divide by 1,000 and round 0.06284 to 0.063.

Dataset	Task / Variant	Metric	*Baseline*	*Qwen3-Omni*	*Qwen2.5-Omni*	*MiniCPM-o 4.5*	*MiniCPM-o 2.6*	*Gemma-4 E4B*	*SOTA*
VoxConverse	Speaker diarization	DER ↓	31.4	31.0	–	36.2	44.6	35.9	**8.2**
MMEW	Emotion recog. (apex)	UAR ↑	14.3	30.1	27.8	13.3	28.0	26.1	**83.6**
	Emotion recog. (clip)	UAR ↑	14.3	25.2	28.2	12.8	14.3	13.3	**83.6**
	Emotion recog. (AU-guided CoT)	UAR ↑	14.3	16.1	30.4	11.3	10.0	21.8	**83.6**
AffWild2	Expression recognition	macro F1	5.1	34.6	33.3	22.3	24.8	26.7	**53.7**
	Expr. recog. (VA/AU-guided CoT)	macro F1	5.1	40.9	28.5	33.3	17.3	29.4	—
MELD	Emotion in dialogue (V+A)	wtd. F1	31.3	53.4	52.0	49.9	31.3	41.9	**71.8**
	Emotion in dialogue (Audio)	wtd. F1	31.3	53.7	52.7	50.7	36.7	46.2	**60.1**
	Emotion in dialogue (V+T)	wtd. F1	31.3	62.1	60.9	35.4	44.1	60.1	**71.8**
EMOTIC	Contextual emotion (cat.)	mAP	17.0	24.0	20.9	19.3	18.0	20.8	**38.1**
	Valence-arousal-dom. (VAD)	AAE ↓	1.03	1.37	1.82	1.15	1.51	1.832	**0.93**
PISC	Social relation classif.	mAP	3.6	49.2	33.0	5.0	4.8	22.7	**80.5**
VideoCoAtt	Co-attention detection	Acc2	50.9	66.9	51.4	58.8	51.8	48.8	**78.1**
	Co-attention localization	L^2 ↓	0.285	0.246	0.310	0.287	0.290	0.385	**0.063**
MSP-Podcast	Emotion recognition	macro F1	6.8	19.3	17.3	9.3	11.2	16.2	**47.2**

Table 5: **L3 Zero-shot results — Theory of Mind & Reasoning.** Model cells report point estimates; confidence intervals appear in the separate supplementary material. Acc2, Acc., and UAR values are in %. **Baseline**: naive predictor. **SOTA**: best published task-specific model. When a model produces no parseable output for a sample, the baseline prediction is substituted for metric computation. †Small test set (< 100 samples).

Dataset	Task / Variant	Metric	*Baseline*	*Qwen3-Omni*	*Qwen2.5-Omni*	*MiniCPM-o 4.5*	*MiniCPM-o 2.6*	*Gemma-4 E4B*	*SOTA*
MUStARD	Sarcasm detection (w/ ctx.)	Acc2	47.1	55.1	56.5	51.4	46.5	47.1	**77.4**
	Sarcasm detection (w/o ctx.)	Acc2	47.1	52.2	47.1	55.1	43.4	47.1	**77.4**
UR-Funny	Humor detection (V+A+ctx.)	Acc2	49.3	59.2	57.9	52.8	52.1	55.2	**73.6**
	Humor detection (V+A)	Acc2	49.3	53.3	49.9	50.1	53.5	51.1	**73.6**
RLDD†	Deception detection	Acc. ↑	48.0	52.0	52.0	55.8	51.7	52.0	**96.4**
MEVIEW†	Concealed emotion recog.	UAR ↑	16.7	22.8	16.1	11.2	16.7	26.1	**68.5**

Table 6: **Compliance with the requested JSON output format (%) per model and task.** Fraction of test samples for which the model produced a parseable, correctly typed JSON response under the original JSON-format prompt, usable for metric computation. Orange: <80 %. Red: <20 %.

Dataset	**Task / Variant**	*Qwen3-Omni*	*Qwen2.5-Omni*	*MiniCPM-o 4.5*	*MiniCPM-o 2.6*	*Gemma-4 E4B*
GazeFollow	Gaze target detection	100	100	49	72	100
VAT	Att. target (frame)	86	53	79	57	83
	Att. target (video)	62	55	53	39	62
VocalSound	Vocal sound classif.	100	100	95	77	100
VoxConverse	Speaker count	100	100	90	81	100
Proxemics	Touch detection	100	74	1	30	21
	Touch detection (+skel.)	100	73	40	28	35
MMEW	AU detection (apex)	100	100	100	93	100
AffWild2	AU detection	100	100	96	93	100
	Valence-arousal	100	100	64	89	100
VoxConverse	Speaker diarization	97	0	57	66	100
MMEW	Emotion recog. (apex)	100	100	14	81	100
	Emotion recog. (clip)	100	100	96	80	100
	Emotion recog. (clip+hint)	100	100	100	70	100
AffWild2	Expr. recognition	100	100	74	98	100
	Expr. recognition (+ VA/AU hint)	100	100	89	100	100
MELD	Emotion in dial. (V+A)	100	100	100	0	100
	Emotion in dial. (Audio)	100	100	100	38	100
	Emotion in dial. (V+T)	100	100	18	46	100
EMOTIC	Context. emotion (cat.)	100	100	70	74	100
	VAD	100	100	62	71	100
PISC	Social relation classif.	100	97	3	2	82
VideoCoAtt	Co-attn. detection	100	100	73	39	100
	Co-attn. localization	100	100	81	18	100
MUStARD	Sarcasm (w/ ctx.)	100	100	58	92	100
	Sarcasm (w/o ctx.)	100	100	59	34	100
UR-Funny	Humor (V+A+ctx.)	100	100	49	39	100
	Humor (V+A)	100	100	33	44	100
RLDD	Deception detection	100	100	92	84	100
Average		97	92	68	60	96

Table 7: **Effect of the output-format instruction.** Median change (points) relative to the JSON-request prompt over model–task cells. *Adjusted*: all responses, unrecoverable ones charged at the naive baseline. *Raw*: recoverable responses only. *Rec.*: median share of responses with a recoverable answer.

Prompt condition	Cells	Δ adj.	Δ raw	Rec. (%)
JSON request	30	—	—	100
No JSON (prose)	30	−0.39	−0.38	100
Free generation	30	−1.59	−0.50	99

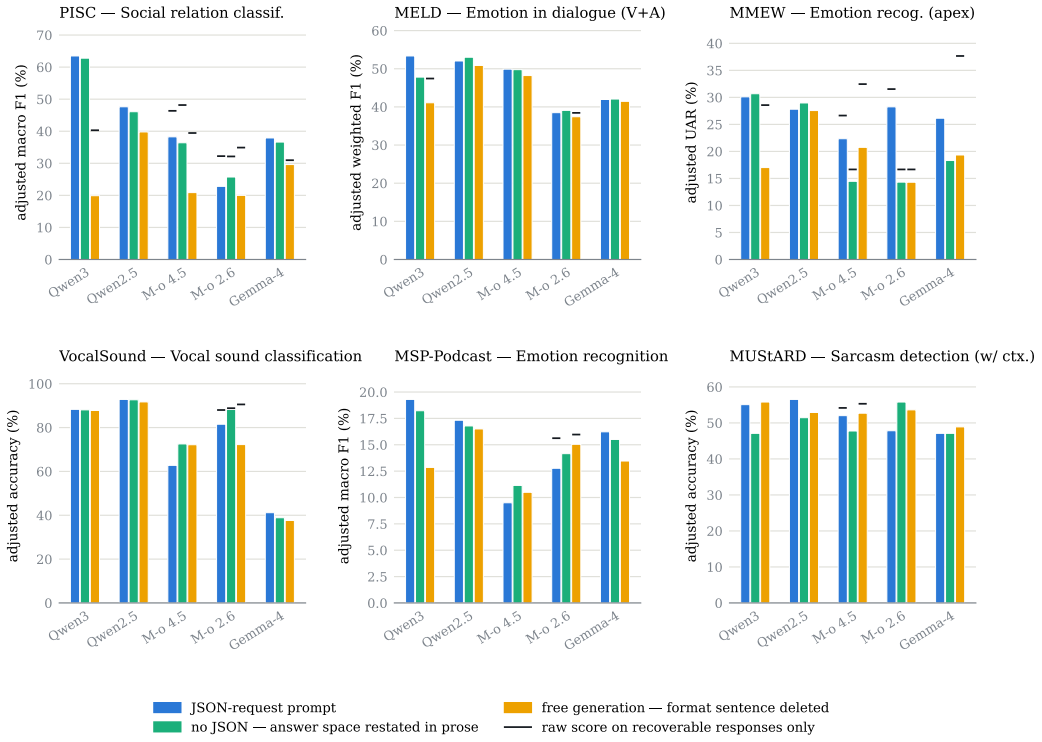


Figure 7: **Adjusted scores under the three output-format conditions** (lenient parser; unrecoverable responses charged at the naive baseline). Black ticks give the raw score on recoverable responses where it differs.

C.1 Output-format and parser ablation

To test whether the requested JSON format drives the zero-shot scores, we reran six task variants (PISC, MELD video+audio, MMEW apex, VocalSound, MSP-Podcast emotion labels, MUStARD) for five models under two alternative prompts: *no JSON*, where the answer space is described in prose, and *free generation*, where the format request is simply removed. All conditions are scored with a lenient parser that also recovers answers from free text; unrecoverable responses are charged at the naive baseline, as in the main results.

Table 7, Figure 7 and Table 8 (per-cell scores) show that the JSON request is not what limits performance. Replacing it with a prose description of the answer space has little effect (median −0.39 points), and removing it altogether slightly lowers scores (median −1.59 points), mostly because fewer responses contain a recoverable answer rather than because answers get worse (median raw change −0.50).

Table 8: **Per-cell scores behind Figure 7 and Table 7.** For each prompt condition: *Adj.*, the adjusted score (lenient parser; unrecoverable responses charged at the naive baseline, $(\mathrm{raw} \cdot n_{\mathrm{rec}} + \mathrm{baseline} \cdot n_{\mathrm{unrec}})/n$); *Raw*, the score on recoverable responses only; *Rec.*, the percentage of responses with a recoverable answer. n is the number of evaluated items, identical across the three conditions of a row. PISC is scored with macro F1, unlike the mAP reported in Table 4. Because the main results use the strict parser, the JSON-request column can differ from Tables 3–5; the largest gaps (VocalSound for MiniCPM-o 4.5 and Gemma-4) come from valid JSON answers stored under a key the strict metric does not read.

			JSON-request prompt			**No JSON**			**Free generation**		
Task (metric)	**Model**	n	Adj.	Raw	Rec.	Adj.	Raw	Rec.	Adj.	Raw	Rec.
PISC (macro F1)	Qwen3-Omni	3000	63.5	63.5	100	62.8	62.8	100	19.9	40.3	41
	Qwen2.5-Omni	3000	47.6	47.6	100	46.1	46.1	100	39.8	40.2	99
	MiniCPM-o 4.5	3000	38.2	46.4	80	36.4	48.2	72	20.9	39.4	45
	MiniCPM-o 2.6	3000	22.8	32.2	64	25.7	32.1	76	20.0	34.9	49
	Gemma-4	3000	37.9	37.9	100	36.6	36.6	100	29.6	31.0	95
MELD (weighted F1)	Qwen3-Omni	2602	53.3	53.3	100	47.8	47.8	100	41.1	47.4	61
	Qwen2.5-Omni	2602	52.0	52.0	100	53.0	53.1	100	50.9	51.0	99
	MiniCPM-o 4.5	2602	49.9	49.9	100	49.8	49.8	100	48.2	48.4	99
	MiniCPM-o 2.6	2602	38.5	38.8	95	39.1	39.2	98	37.4	38.4	86
	Gemma-4	2602	41.9	41.9	100	42.1	42.1	100	41.5	41.5	100
MMEW (UAR)	Qwen3-Omni	236	30.1	30.1	100	30.7	30.7	100	17.0	28.6	19
	Qwen2.5-Omni	236	27.8	27.8	100	29.0	29.0	100	27.6	27.6	100
	MiniCPM-o 4.5	236	22.3	26.6	65	14.5	16.7	8	20.8	32.5	36
	MiniCPM-o 2.6	236	28.2	31.5	81	14.3	16.7	1	14.3	16.7	0
	Gemma-4	236	26.1	26.1	100	18.3	18.3	100	19.3	37.7	22
VocalSound (accuracy)	Qwen3-Omni	3000	88.3	88.3	100	88.1	88.1	100	87.8	87.8	100
	Qwen2.5-Omni	3000	92.8	92.8	100	92.7	92.7	100	91.7	91.7	100
	MiniCPM-o 4.5	3000	62.8	63.2	99	72.5	72.5	100	72.2	72.2	100
	MiniCPM-o 2.6	3000	81.5	88.0	91	88.2	88.8	99	72.2	90.6	75
	Gemma-4	3000	41.2	41.2	100	38.9	38.9	100	37.6	37.6	100
MSP-Podcast (macro F1)	Qwen3-Omni	3000	19.3	19.3	100	18.2	18.2	100	12.9	12.9	100
	Qwen2.5-Omni	3000	17.3	17.3	100	16.8	16.8	100	16.5	16.5	100
	MiniCPM-o 4.5	3000	9.5	9.5	100	11.1	11.1	100	10.5	10.5	100
	MiniCPM-o 2.6	3000	12.8	15.6	68	14.2	14.2	100	15.0	16.0	90
	Gemma-4	3000	16.2	16.2	100	15.5	15.5	100	13.5	13.5	100
MUStARD (accuracy)	Qwen3-Omni	138	55.1	55.1	100	47.1	47.1	100	55.8	55.8	100
	Qwen2.5-Omni	138	56.5	56.5	100	51.4	51.4	100	52.9	52.9	100
	MiniCPM-o 4.5	138	52.0	54.2	70	47.8	48.0	72	52.7	55.3	68
	MiniCPM-o 2.6	138	47.8	47.8	100	55.8	55.8	100	53.6	53.6	100
	Gemma-4	138	47.1	47.1	100	47.1	47.1	100	48.9	48.9	99

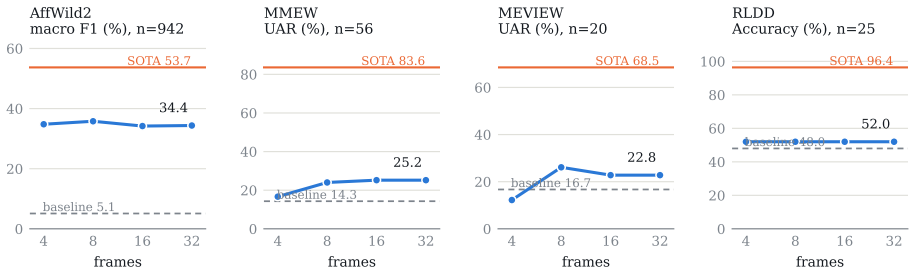


Figure 8: **Qwen3-Omni zero-shot score by number of sampled frames.** Each task has its own metric and vertical scale. Dashed lines show the naive baseline and solid lines the task-specific reference result. The horizontal axis shows 4, 8, 16, and 32 frames.

Table 9: **Number-of-frames ablation for Qwen3-Omni zero-shot.** Metrics are in percentage points; $\Delta_{8\to32}$ is the change from 8 to 32 frames. All four tasks had zero parse failures at every frame count. MEVIEW and RLDD have small test sets, so modest differences should not be overinterpreted.

Dataset	Task	Metric	n	Baseline	Sampled frames				$\Delta_{8\to32}$	SOTA
					4	8	16	32		
AffWild2	Expression recognition	macro F1	942	5.1	34.79	**35.78**	34.19	34.37	−1.42	53.7
MMEW	Emotion recognition (clip)	UAR	56	14.3	16.67	24.01	**25.20**	**25.20**	+1.19	83.6
MEVIEW	Concealed emotion recognition	UAR	20	16.7	12.22	**26.11**	22.78	22.78	−3.33	68.5
RLDD	Deception detection	Accuracy	25	48.0	**52.00**	**52.00**	**52.00**	**52.00**	0.00	96.4

C.2 Number-of-frames ablation

For four video tasks, we forced Qwen3-Omni zero-shot inference to use 4, 8, 16, or 32 sampled frames from the same 32-frame source clips, without retraining. Short clips can yield fewer frames than the nominal sampling maximum in the standard evaluation. Table 9 and Figure 8 report the resulting scores. The AffWild2 expression subset here contains 942 clips and uses the corrected label-only prompt; the main AffWild2 expression result uses 1,028 clips, so the two scores should not be compared as if they used identical test sets.

From 8 to 32 frames, Qwen3-Omni changes by $+1.19$ points on MMEW, -1.42 on AffWild2 expression, -3.33 on MEVIEW, and zero on RLDD. These results do not show a consistent improvement with more frames. The small MMEW, MEVIEW, and RLDD subsets limit how precisely those differences can be interpreted; RLDD's constant majority-class prediction in particular does not establish robustness to frame sampling.

D LEARNING SOCIAL TASKS: EXTENDED RESULTS

D.1 DETAILED EXPERIMENTAL RESULTS

While the main part of the paper evaluates the impact of adaptation on SOCIALBENCH using a spider plot, we complement that analysis here by presenting the full task-by-task breakdown. Table 10 provides the detailed results of Qwen3-Omni across all benchmark tasks, contrasting zero-shot inference against one epoch of supervised fine-tuning (SFT). Consistent with the positive delta (Δ) values, a single epoch of tuning yields improvements across many tasks. However, multiple tasks exhibit non-negligible regressions, and there are no clear, overarching trends correlating these performance shifts with specific task categories or difficulty levels.

Table 10: **Zero-shot vs. One-Epoch Fine-Tuning vs. Joint Training on SOCIALBENCH.** Performance of Qwen3-Omni across all benchmark tasks. The corresponding 95% bootstrap confidence intervals are provided in the separate supplementary material. **Joint**: joint fine-tuning with square-root sampling (25% budget). **Bold**: best result overall; underline: second best. Δ_{1ep}: 1-epoch − 0-shot; Δ_{Jt}: Joint − 1-epoch; **green** / **red**: improvement / regression on the task metric. AffWild2 expression recognition uses the corrected label-only evaluation on 1,028 clips. †Small test set.

Level	Dataset	Task / Variant	0-shot	1-ep	Joint	SOTA	Δ_{1ep}	Δ_{Jt}	Metric
L1 · Perception	GazeFollow	Gaze target detection	0.230	0.155	0.167	**0.099**	−0.075	+0.012	Dist. ↓
	VAT	Att. target localization (*frame*)	0.248	0.176	0.169	**0.103**	−0.072	−0.007	Dist. ↓
		+ *video*	0.281	0.197	0.195	**0.103**	−0.084	−0.002	
	VocalSound	Vocal sound classification	0.882	0.944	0.934	**0.980**	+0.062	−0.010	Acc. ↑
	VoxConverse	Speaker count estimation	0.642	**0.763**	0.726	—	+0.121	−0.037	Exact Acc. ↑
	Proxemics	Physical touch	0.319	0.353	0.594	**0.738**	+0.034	+0.241	mAP ↑
		+ *skeleton*	0.301	0.328	0.586	**0.738**	+0.026	+0.258	
	MMEW	Action unit detection	0.022	0.207	**0.282**	—	+0.185	+0.075	macro F1 ↑
	AffWild2	Action unit detection	0.393	0.521	0.470	**0.595**	+0.128	−0.051	macro F1 ↑
		Valence-arousal estimation	0.388	0.373	0.409	**0.658**	−0.015	+0.036	CCC ↑
	MSP-Podcast	VAD estimation	0.318	**0.734**	0.720	0.685	+0.416	−0.014	CCC ↑
L2 · Understanding	VoxConverse	Speaker diarization	0.310	0.252	0.268	**0.082**	−0.058	+0.016	DER ↓
	MMEW	Emotion recognition (*apex*)	0.301	0.331	0.412	**0.836**	+0.030	+0.081	UAR ↑
		Emotion recognition (*clip*)	0.252	0.253	0.349	**0.836**	+0.001	+0.096	
	AffWild2	Expression recognition	0.346	0.455	0.445	**0.537**	+0.109	−0.010	macro F1 ↑
	MELD	Emotion in dialogue (*video* + *audio*)	0.534	0.570	0.578	**0.718**	+0.036	+0.008	wtd. F1 ↑
		audio only	0.537	0.578	0.578	**0.601**	+0.041	±0.000	
		video + *transcript*	0.621	0.652	0.685	**0.718**	+0.031	+0.033	
	EMOTIC	Contextual emotion (categorical)	0.240	0.235	0.238	**0.381**	−0.005	+0.003	mAP ↑
		VAD	1.366	1.082	1.003	**0.926**	−0.284	−0.079	AAE ↓
	PISC	Social relation classification	0.492	0.693	0.693	**0.805**	+0.201	±0.000	mAP ↑
	VideoCoAtt	Co-attention detection	0.669	**0.867**	0.818	0.781	+0.198	−0.049	Acc2 ↑
		Co-attention localization	0.246	0.156	0.131	**0.063**	−0.090	−0.025	L^2 ↓
	MSP-Podcast	Emotion recognition	0.193	0.195	0.203	**0.472**	+0.002	+0.008	macro F1 ↑
L3 · ToM	MuStARD	Sarcasm detection (*w/ context*)	0.551	0.725	0.681	**0.774**	+0.174	−0.044	Acc2 ↑
		w/o context	0.522	0.710	0.645	**0.774**	+0.188	−0.065	
	UR-Funny	Humor detection (*video* + *context*)	0.592	0.677	0.643	**0.736**	+0.085	−0.034	Acc2 ↑
		+ *video-audio*	0.533	0.602	0.572	**0.736**	+0.069	−0.030	
	RLDD†	Deception detection	0.520	0.480	0.800	**0.964**	−0.040	+0.320	Acc. ↑
	MEVIEW†	Concealed emotion recognition	0.228	0.300‡	0.333	**0.685**	+0.072	+0.033	UAR ↑

‡The 0.300 MEVIEW result uses the adapter trained on RLDD and evaluated on MEVIEW without MEVIEW training examples. This transfer protocol was fixed before consulting MEVIEW results.

Figure 9 shows the gains and losses recalculated from the current table. The largest positive changes are VideoCoAtt co-attention localization (+0.397 on the SOTA-scaled axis), Proxemics with skeleton (+0.350), and RLDD (+0.332). The largest regressions are VoxConverse diarization (−0.195), GazeFollow (−0.121), and AffWild2 action-unit detection (−0.086). AffWild2 expression recognition changes by −0.019, while MELD video+transcript improves by +0.046 on the SOTA-scaled axis. Across the 30 tasks, 16 improve, 12 regress, and two are unchanged at the precision reported in Table 10; these heterogeneous effects do not establish a uniform advantage for joint training.

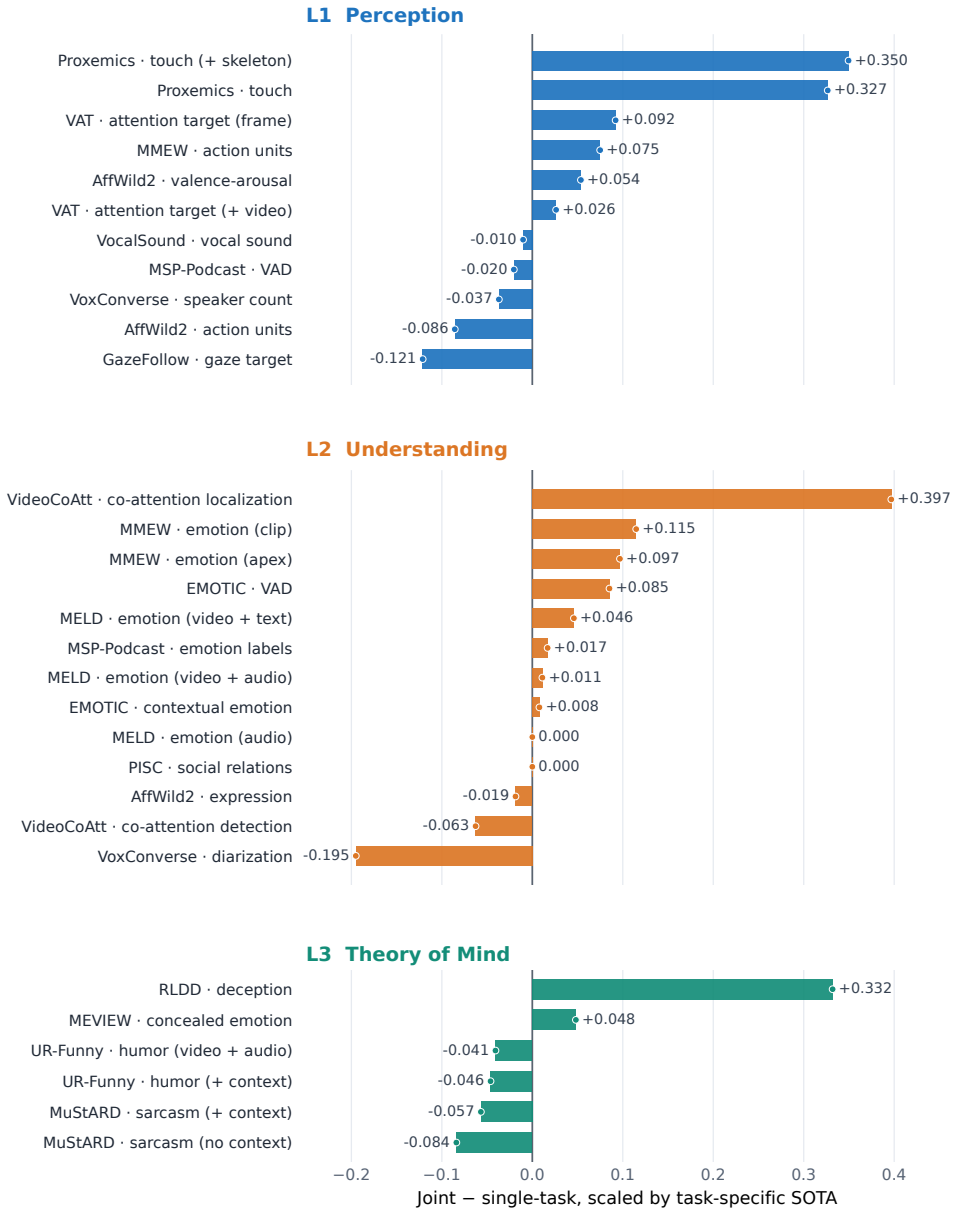


Figure 9: **Joint vs. single-task gain by task.** Recomputed from the 1-epoch and joint columns of Table 10. Each bar is $s(\text{Joint} - \text{1ep})/\text{SOTA}$, with $s = +1$ for higher-is-better metrics and $s = -1$ for lower-is-better metrics. For VoxConverse speaker count and MMEW action units, which lack a published SOTA in the table, the scale is the theoretical maximum of 1. Positive bars favor joint training. Colors identify the three cognitive levels.

D.2 Detailed Training Protocol and Sampling Strategies

Hardware and Hyperparameters. Our fine-tuning experiments are executed using DeepSpeed ZeRO-3 (Rajbhandari et al., 2020). We apply Low-Rank Adaptation (LoRA) specifically to the self-attention projections and cross-modal modality mergers. This targeted adaptation yields a highly parameter-efficient setup, training only $\approx 0.046\%$ of the total parameters. The complete hardware setup and LoRA hyperparameters are summarized in Table 11.

Table 11: Training hardware and LoRA hyperparameters.

Setting / Hyperparameter	Value
Base Model	Qwen3-Omni-30B-A3B-Thinking
Hardware	4× NVIDIA H100 (80 GB)
Training Framework	MS-Swift (Zhao et al., 2025) + DeepSpeed ZeRO-3 (Rajbhandari et al., 2020)
Precision	bfloat16
Effective Batch Size (b)	16 (1/GPU × 4 accumulation steps)
Max Sequence Length	8,192 tokens
Video Sampling	16 frames for inference
Random Seed	42
Learning Rate	2×10^{-4}
LR Scheduler	Cosine decay with 5% linear warmup
Weight Decay	0.01
Gradient Clipping	max norm 1.0
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
LoRA Target Modules	$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$, cross-modal projectors
Trainable Parameters	≈14.5M (≈0.046%)

Compute Estimates. Joint fine-tuning runs for ≈5,982 gradient steps at an average of ≈43 s/step, yielding approximately 71 hours of wall-clock time on 4×H100 (≈286 H100-GPU-hours). Single-task fine-tuning trains for one epoch on the full task-specific training set wherever one is available; MEVIEW has zero training examples, and its reported score is transfer from the RLDD adapter. Across these single-task experiments, training consumes an additional ≈180 H100-GPU-hours, for a total experimental budget of ≈466 H100-GPU-hours. Preliminary experiments (sampling ablation at 1,000 steps, LoRA rank sensitivity) add ≈30 H100-GPU-hours. Sequence packing is disabled throughout due to heterogeneous video-length inputs, which reduces throughput relative to text-only workloads.

Evaluation uncertainty. Each configuration is evaluated once on a fixed set of examples. We estimate sampling uncertainty with 95% percentile confidence intervals from 1,000 bootstrap resamples (seed 42), recomputing the full task metric for each resample. The intervals are reported in the separate supplementary material for the zero-shot results (Tables 3–5) and the fine-tuning comparison (Table 10), and are particularly relevant for small evaluation sets (RLDD: 25 test samples; MMEW apex: 236 samples). Intervals are not available for a few Qwen3-Omni cells (MMEW clip-level emotion recognition and MEVIEW), which are marked "—" in the supplementary tables. They do not measure variation across training seeds or alternative frame samples. The fixed video sampling can affect long clips when the selected frames are not representative, particularly for micro-expression and deception tasks.

Joint Training Compute Budget. To ensure a fair comparison across sampling strategies during joint training, we defined a shared compute budget expressed as a fixed number of gradient steps $X = \lfloor K_{\text{budget}}/b \rfloor$. The budget is derived from capped dataset sizes:

$$K_{\text{budget}} = \alpha \cdot \sum_{i=1}^{N} \min(K_i,\ K_{\max}), \tag{1}$$

where K_i denotes the number of training samples for task i, $K_{\max}$=50 000 serves as a threshold to prevent the largest datasets from monopolizing the training budget, and α controls the total fraction of the capped pool utilized. In our main experiments, we set α=0.25, yielding X≈6 000 gradient steps.

Sampling Strategies. Because three of our methods strictly adhere to the same X-step budget, any observed variance in downstream performance can be attributed exclusively to the sampling distribution rather than differences in total compute. The simplified definitions of the four evaluated strategies are detailed in Table 12. These different strategies are quantitatively compared on a checkpoint after 1000 iterations in Table 13. Evaluation is performed on a subset of tasks to save compute. Square-root and proportional sampling are the two strongest strategies: proportional sampling is best on four of the nine rows (VoxConverse, AffWild2 valence-arousal, MELD video+audio, PISC) and square-root sampling on three (GazeFollow, MELD audio, UR-Funny), while first-exhausted sampling is best on VocalSound and EMOTIC and uniform sampling on none. No strategy dominates across tasks. We retain square-root sampling for the main experiments because it upweights smaller datasets (Table 12) while remaining close to the best strategy on most rows, but these results do not show it to be superior to proportional sampling. for running the full training experiments.

Table 12: Overview of joint-training sampling strategies evaluated on SOCIALBENCH.

Strategy	Probability Formulation	Description
Proportional	$p_i \propto \min(K_i, K_{\max})$	Samples proportionally to the capped dataset size. Large datasets dominate the training pool.
Square-root	$p_i \propto \sqrt{\min(K_i, K_{\max})}$	Smooths size disparities, upweighting smaller datasets to ensure better tail representation.
Uniform	$p_i = 1/N$	Samples all datasets equally. Small datasets cycle multiple times via online interleaving.
First-exhausted	$p_i \propto \sqrt{\min(K_i, K_{\max})}$	Identical to square-root sampling, but strictly halts as soon as the smallest dataset is exhausted.

Table 13: **Joint training — sampling strategy comparison (checkpoint-1000, $\approx$17% of budget).** Performance of Qwen3-Omni fine-tuned jointly on all SOCIALBENCH tasks under four sampling strategies at an intermediate checkpoint. Only tasks for which comparable results were available at this stage are shown. **Bold**: best joint result per row. Metric direction shown in column header.

				Reference		*Joint training (sampling strategy)*			
Lv.	**Dataset**	**Task / Variant**	**Metric**	**0-shot**	**1-ep**	**Sqrt**	**Uniform**	**Prop.**	**1st-exh.**
L1	GazeFollow	Gaze target detection	Dist. ↓	0.230	0.155	**0.182**	0.199	0.183	0.198
	VocalSound	Vocal sound classif.	Acc. ↑	0.883	0.944	0.921	0.896	0.912	**0.922**
	VoxConverse	Speaker count	Exact Acc. ↑	0.642	0.763	0.676	0.701	**0.720**	0.679
	AffWild2	Valence-arousal	CCC ↑	0.388	0.373	0.434	0.374	**0.468**	0.369
L2	MELD	Emotion (*V+A*)	wtd. F1 ↑	0.534	0.570	0.544	0.546	**0.550**	0.535
		Emotion (*audio*)	wtd. F1 ↑	0.537	0.578	**0.556**	0.524	0.553	0.511
	EMOTIC	Contextual emotion (cat.)	mAP ↑	0.240	0.235	0.220	0.185	0.214	**0.222**
	PISC	Social relation classif.	mAP ↑	0.492	0.693	0.671	0.662	**0.738**	0.686
L3	UR-Funny	Humor (*V+A+ctx.*)	Acc. ↑	0.592	0.677	**0.604**	0.603	0.542	0.538

D.3 LoRA HYPERPARAMETER SENSITIVITY ANALYSIS

We compare LoRA rank 16 (used in all main experiments) against rank 32 under the square-root sampling strategy at checkpoint-1000 ($\approx$17% of the full training budget). Table 14 shows that the two configurations perform comparably across all evaluated tasks, with absolute differences below 0.03 and no consistent advantage for either rank. Given this parity and the additional memory and compute overhead of rank 32 ($\approx$29M vs. $\approx$14.5M trainable parameters), we adopt rank 16 throughout and did not extend the rank-32 run to the full budget.

Table 14: **LoRA rank ablation** (sqrt sampling, checkpoint-1000). Differences are small and inconsistent, justifying rank 16 for all main experiments.

Dataset	Metric	Rank 16	Rank 32	Δ
GazeFollow	Dist. ↓	0.182	0.203	+0.021
AffWild2 VA	CCC ↑	0.434	0.431	−0.003
EMOTIC	mAP ↑	0.220	0.214	−0.006
PISC	mAP ↑	0.671	0.698	+0.027
MELD (audio)	wtd. F1 ↑	0.556	0.558	+0.003
MELD (V+A)	wtd. F1 ↑	0.544	0.560	+0.015
VocalSound	Acc. ↑	0.921	0.902	−0.019
UR-Funny	Acc. ↑	0.604	0.589	−0.015
VoxConverse	Exact Acc. ↑	0.676	0.657	−0.019
Avg. $\lvert\Delta\rvert$				0.013

E Understanding Cross-task Relationships at Inference Time

Section 5 focuses on cross-task synergies at training time by jointly fine-tuning on all tasks. A natural follow-up question is whether such interactions also emerge at inference time: can the model's own predictions on a source task serve as useful context for a different target task, without any additional training? Below, we provide an experimental answer to this question.

E.1 Protocol

Hint task selection. We define a hint task as any task whose predictions can be reliably textualized and injected into another task's prompt. A prerequisite for a meaningful hint is that the source predictions are *informative*: a hint drawn from a near-random predictor adds noise rather than signal, making it impossible to interpret any downstream effect as evidence of cross-task transfer. We therefore restrict hint tasks to those on which our joint-trained checkpoint achieves performance close to the state of the art at the respective cognitive level. This yields six hint tasks spanning L1 and L2: gaze target localization (GazeFollow), facial action unit detection (AffWild2 AU), vocal sound classification (VocalSound), co-attention detection and localization (VideoCoAtt), and social relation classification (PISC). These tasks cover the three primary input modalities (image, video, audio) and represent complementary social signals: spatial attention, facial dynamics, paralinguistic cues, group attention, and interpersonal context.

Experimental setup. The evaluation proceeds in two phases. In **Phase 1**, for each (hint task B_j, target dataset A_i) pair, we perform inference with the joint-trained checkpoint on the target dataset using the hint task's prompt (adapted to remove bounding-box requirements), producing a JSONL of hint predictions. In **Phase 2**, each hint prediction is prepended to the corresponding target sample's user message as a bracketed context line (`[Hint -- `B_j`: prediction]`), and standard inference is run on the target task. Pairs are excluded when the hint modality is incompatible with the target (e.g., an audio hint cannot be applied to an image-only dataset), yielding 144 valid pairs out of a theoretical $6 \times 30 = 180$.

Metrics For each valid pair (B_j, A_i), let $\mathcal{M}(A_i)$ denote the primary metric of task A_i without hint. We define the relative gain as $\Delta(B_j \to A_i) = \frac{\mathcal{M}(A_i|B_j) - \mathcal{M}(A_i)}{|\mathcal{M}(A_i)|} \times 100$, and the oracle-best hint as $B^*(A_i) = \text{argmax}_{B_j}\ \Delta(B_j \to A_i)$. Figure 10 summarises the full relative-gain matrix across all target levels; per-level detailed matrices are provided in Section E.3.

E.2 Results and Discussion

Hints help understanding, not perception. Cross-task transfer is strongly asymmetric across cognitive levels. Injecting hint predictions into L1 tasks, focused on extracting low-level physical signals, yields negligible or negative gains regardless of the hint source (Figure 10). In contrast, L2 tasks that integrate contextual cues to infer latent states (e.g., emotion, social relations) benefit

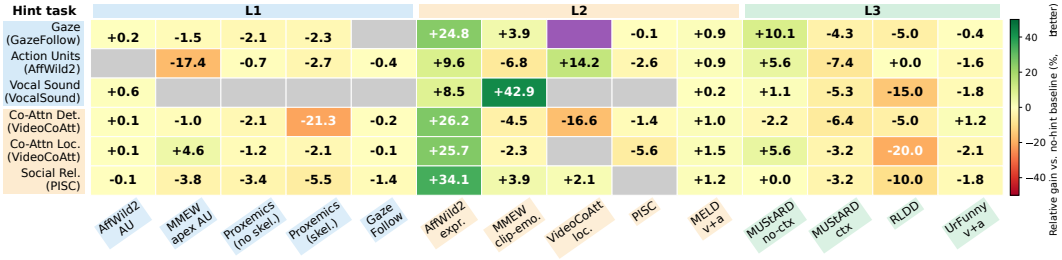


Figure 10: **Cross-task hint transfer across all SOCIALBENCH levels.** Each cell reports the relative gain (%) vs. the no-hint baseline on the target task's primary metric. Rows: hint tasks (L1–L2). Columns: target tasks grouped by cognitive level (L1, L2, L3). Grey: modality-incompatible pair. Purple: the requested JSON format is not respected. Full matrices for each level in Section E.3.

consistently from L1 hints. This asymmetry suggests that the bottleneck for L1 tasks lies in visual perception itself, which textual hints cannot bypass, whereas L2 tasks suffer primarily from a lack of grounding information that hints can supply. The phenomenon is less pronounced for task in the third cognition level (L3), as the improvements are less pronounced, and the losses can be a bit larger. However, in both L2 and L3 tasks, we observe a certain structure described next.

Transfer depends more on the target than the source. Across all hint sources, the benefit of cross-task transfer is largely determined by the target task rather than by which hint is used. Expression recognition (AffWild2) is a strong beneficiary across all hint types: gaze (+25%), action units (+9.6%), social relations (+34%), suggesting it can readily exploit any additional social signal. By contrast, deception detection (RLDD) degrades under almost every hint (up to -20%), and MEVIEW drops sharply (-50%) under action-unit and social-relation hints. This points to a fundamental difference in how tasks can absorb external predictions: some targets benefit from any grounding, while others are disrupted by it, independently of dataset size.

Rich context can reduce hint effectiveness. In some cases, adding hints to an already information-rich input hurts rather than helps. In sarcasm detection (MuStARD), the context-free variant improves by +10% and +5.6% with gaze and action unit hints respectively, while the context-aware variant regresses by -4.3% and -7.4%. Similarly, for Proxemics, co-attention detection hints barely affect the skeleton-free variant (-2.1%) but degrade the skeleton-aware variant by -21%. This pattern is not universal: on UR-Funny, the context-aware variant responds slightly better to hints than the video+audio one (5 of 6 hints). Appended hints may therefore compete with existing textual context in some tasks, but the effect is task-dependent.

E.3 PER-LEVEL HINT TRANSFER MATRICES

Below, we provide the full cross-task experiment matrices for the three levels.

F PROBING RESULTS BY TASK

Tables 15 and 16 show the cross-validation-selected probe type for each task and its held-out score, alongside the model's output score. The reported difference is Δ = probe − model; its direction is interpreted according to whether the task metric is higher- or lower-is-better. Figure 14 shows the zero-shot and fine-tuned probe–output scores of Figure 4 as separate panels.

F.1 PROBE AND OUTPUT SCORES BY COGNITIVE LEVEL

Figures 15–17 report, for every task, the absolute scores behind Figure 4: the best linear probe and the model's output in each arm, next to the published SOTA. Unlike the SOTA-ratio scale used in the main text, these are raw task metrics, so bars are comparable only within a task. Arrows in the axis labels mark the metrics where lower is better. VoxConverse speaker count and MMEW action-unit detection have no published SOTA and therefore no black bar.

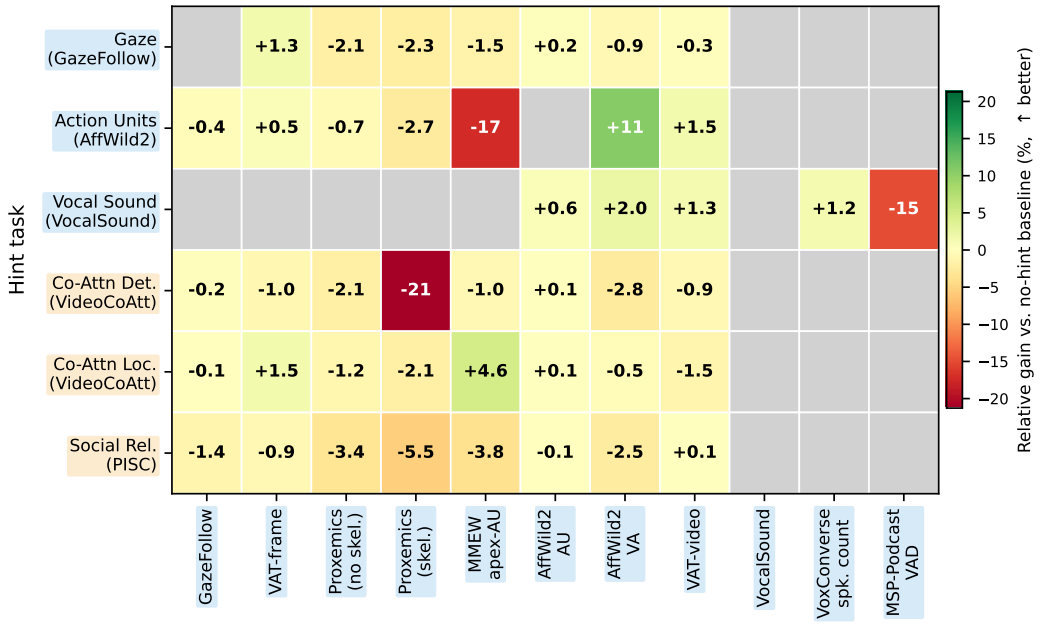


Figure 11: **Cross-task hint transfer — L1 · Perception targets.** Values are relative gain (%) vs. the no-hint baseline on each task's primary metric. y-axis labels indicate the hint task's cognitive level. Grey: modality-incompatible pair. Purple: format contamination (hint overrides output format). Baseline: joint-trained SQRT checkpoint (25% budget).

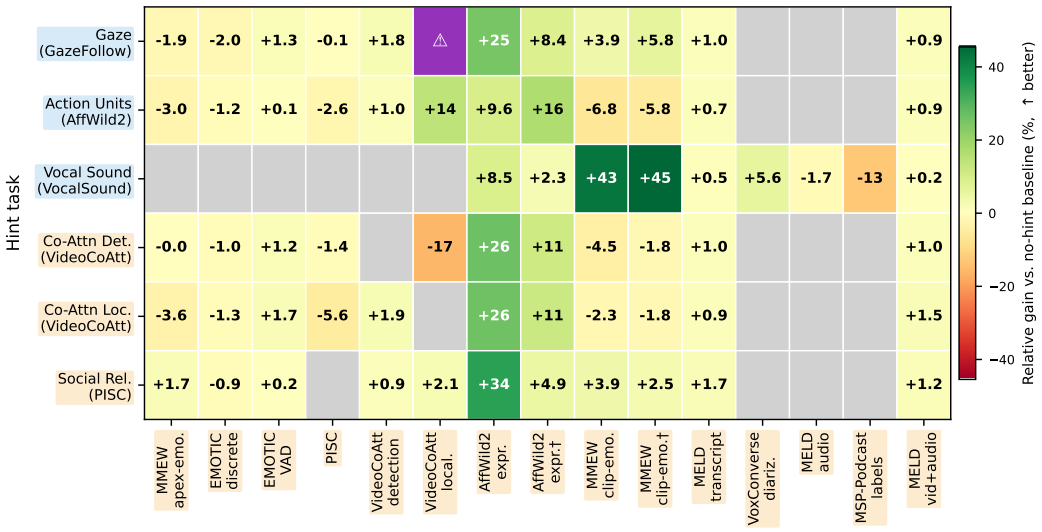


Figure 12: **Cross-task hint transfer — L2 · Understanding targets.** Same setup as Figure 11.

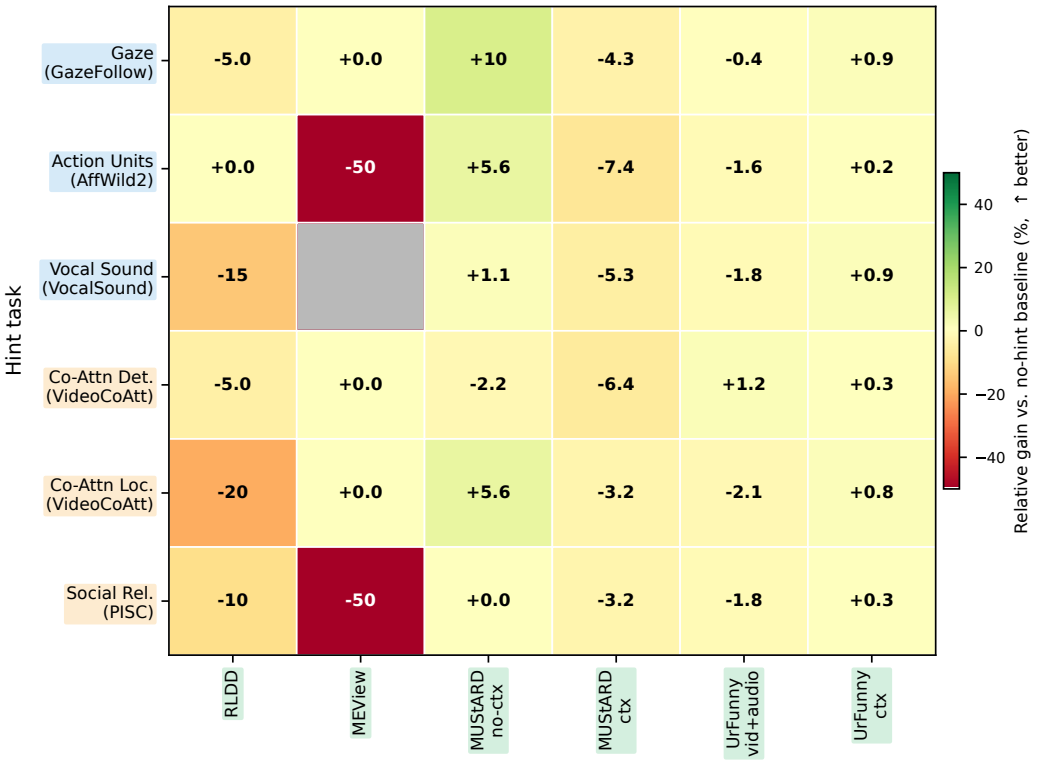


Figure 13: **Cross-task hint transfer — L3 · Theory of Mind targets.** Same setup as Figure 11.

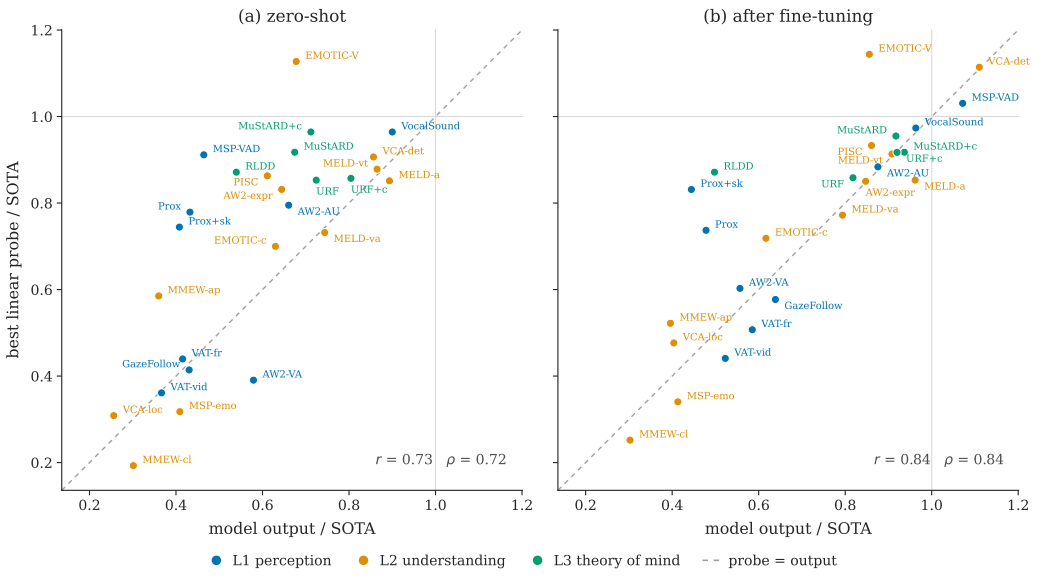


Figure 14: **Linear decodability versus generated-answer performance, zero-shot and after single-task fine-tuning.** Same axes, normalization, and tasks as Figure 4; each point is a task, and r and ρ are the Pearson and Spearman correlations between probe and output scores.

Table 15: **Best linear probe by task: zero-shot activations.** The probe type includes its selected layer. Δ = probe − model.

Task (metric)	Best probe (type; layer)	Probe	Model output	Δ
UR-FUNNY (video+audio) / Acc2	last token (L34)	0.628	0.533	+0.095
MELD (video+audio) / weighted F1	template (L27)	0.525	0.534	-0.009
VideoCoAtt (detection) / Acc2	per-sample text (L37)	0.708	0.669	+0.039
Aff-Wild2 (action units) / macro F1	question (L27)	0.473	0.393	+0.080
MSP-Podcast (VAD) / mean CCC	question (L17/L30/L17)	0.624	0.318	+0.306
GazeFollow / avg. distance ↓	template (L38)	0.239	0.230	+0.009
EMOTIC (discrete) / mAP	template (L40)	0.267	0.240	+0.027
Aff-Wild2 (VA) / mean CCC	video (L2/L21)	0.262	0.388	-0.126
MSP-Podcast (emotion) / macro F1	question (L46)	0.150	0.193	-0.043
VocalSound / accuracy	audio (L29)	0.945	0.882	+0.063
RLDD (n=25) / accuracy	last token (L27)	0.840	0.520	+0.320
MuStARD (context) / accuracy	last token (L15)	0.746	0.551	+0.195
MuStARD (no context) / accuracy	template (L14)	0.710	0.522	+0.188
PISC / mAP	question (L37)	0.695	0.492	+0.203
EMOTIC (VAD) / AAE ↓	template (L28/L38/L32)	0.822	1.366	-0.544
VoxConverse (count) / exact acc.	question (L35)	0.614	0.642	-0.028
VAT (frame) / distance ↓	template (L36)	0.234	0.248	-0.014
VAT (video) / distance ↓	template (L40)	0.285	0.281	+0.004
Proxemics (no skel.) / mAP	question (L34)	0.575	0.319	+0.256
Proxemics (skeleton) / mAP	template (L36)	0.549	0.301	+0.248
MMEW (apex AU) / macro F1	image (L13)	0.131	0.022	+0.109
MMEW (apex emo.) / UAR	question (L36)	0.489	0.301	+0.188
MMEW (clip emo.) / UAR	last token (L31)	0.161	0.252	-0.091
Aff-Wild2 (expression) / macro F1	template (L40)	0.447	0.346	+0.101
MELD (audio only) / wtd. F1	question (L30)	0.512	0.537	-0.025
MELD (video+transcript) / wtd. F1	question (L29)	0.631	0.621	+0.010
UR-FUNNY (video+context) / Acc2	template (L28)	0.631	0.592	+0.039
VideoCoAtt (localization) / L2 dist. ↓	per-sample text (L35)	0.204	0.246	-0.042

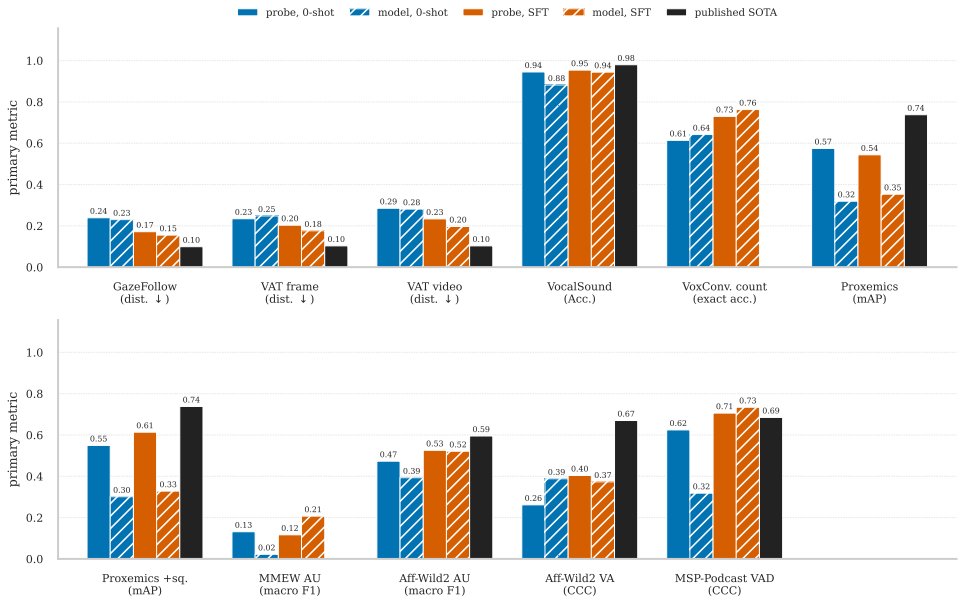


Figure 15: **L1 – Perception.** Best linear probe and model output, zero-shot and after single-task fine-tuning, against the published SOTA.

Table 16: **Best linear probe by task: fine-tuned activations.** The probe type includes its selected layer. Δ = probe − model.

Task (metric)	Best probe (type; layer)	Probe	Model output	Δ
UR-FUNNY (video+audio) / Acc2	last token (L38)	0.632	0.602	+0.030
MELD (video+audio) / weighted F1	question (L39)	0.554	0.570	-0.016
VideoCoAtt (detection) / Acc2	template (L47)	0.870	0.867	+0.003
Aff-Wild2 (action units) / macro F1	last token (L39)	0.526	0.521	+0.005
MSP-Podcast (VAD) / mean CCC	question (L34/L39/L39)	0.706	0.734	-0.028
GazeFollow / avg. distance ↓	template (L39)	0.172	0.155	+0.017
EMOTIC (discrete) / mAP	template (L44)	0.274	0.235	+0.039
Aff-Wild2 (VA) / mean CCC	question (L38/L36)	0.404	0.373	+0.031
MSP-Podcast (emotion) / macro F1	template (L46)	0.161	0.195	-0.034
VocalSound / accuracy	question (L41)	0.954	0.944	+0.010
RLDD (n=25) / accuracy	question (L23)	0.840	0.480	+0.360
MuStARD (context) / accuracy	last token (L36)	0.710	0.725	-0.015
MuStARD (no context) / accuracy	template (L41)	0.739	0.710	+0.029
PISC / mAP	question (L47)	0.751	0.693	+0.058
EMOTIC (VAD) / AAE ↓	template (L38/L33/L45)	0.810	1.082	-0.272
VoxConverse (count) / exact acc.	last token (L45)	0.730	0.763	-0.033
VAT (frame) / distance ↓	last token (L40)	0.203	0.176	+0.027
VAT (video) / distance ↓	template (L40)	0.234	0.197	+0.037
Proxemics (no skel.) / mAP	question (L35)	0.544	0.353	+0.191
Proxemics (skeleton) / mAP	question (L30)	0.614	0.328	+0.286
MMEW (apex AU) / macro F1	image (L10)	0.117	0.207	-0.090
MMEW (apex emo.) / UAR	question (L30)	0.436	0.331	+0.105
MMEW (clip emo.) / UAR	template (L33)	0.211	0.253	-0.042
Aff-Wild2 (expression) / macro F1	last token (L45)	0.457	0.455	+0.002
MELD (audio only) / wtd. F1	question (L45)	0.513	0.578	-0.065
MELD (video+transcript) / wtd. F1	template (L42)	0.656	0.652	+0.004
UR-FUNNY (video+context) / Acc2	question (L36)	0.675	0.677	-0.002
VideoCoAtt (localization) / L2 dist. ↓	template (L39)	0.132	0.156	-0.024

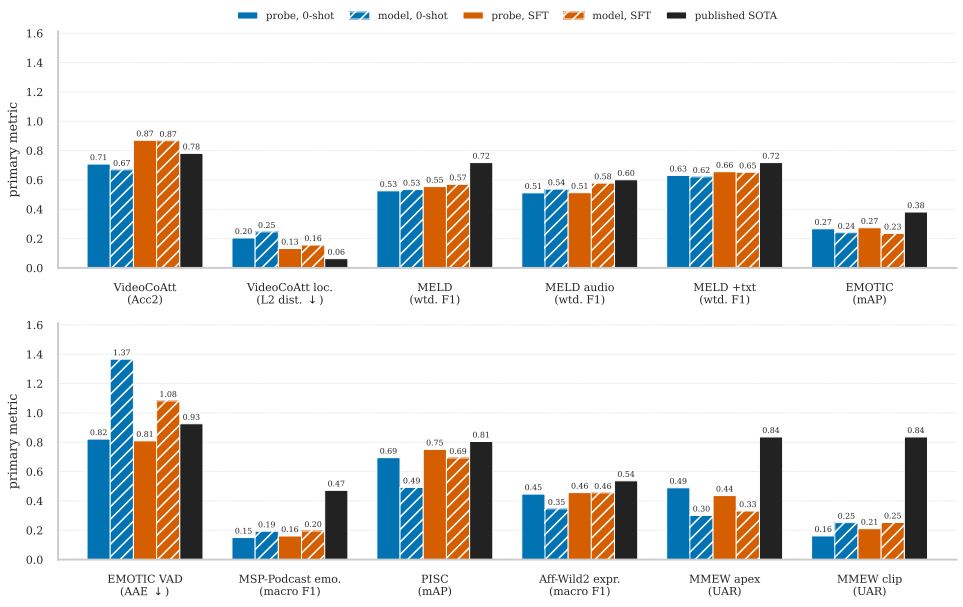


Figure 16: **L2 – Understanding.** Same conventions as Figure 15.

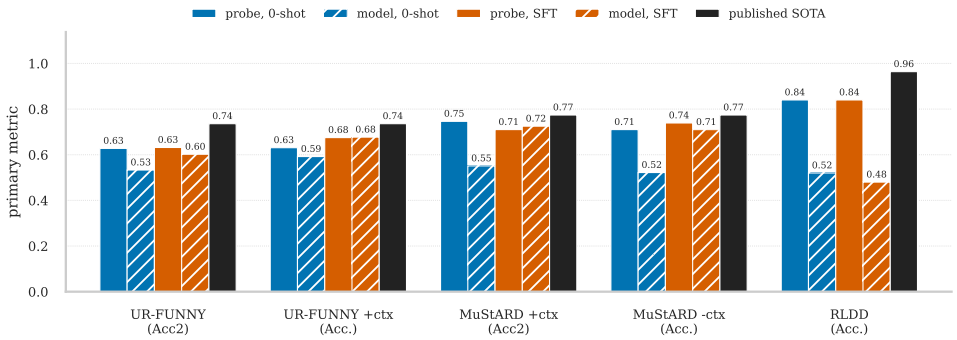


Figure 17: **L3 – Theory of Mind.** Same conventions as Figure 15. RLDD is shown for completeness; with 25 evaluation rows its probe–output difference is not interpretable.