AMBench: Benchmarking Affective Incongruence in Multimodal LLMs

Anonymous ACL submission

Abstract

Multimodal affect evaluation commonly treats language, voice, and vision as convergent evidence for a single emotion label. This breaks down in counseling, where clients may reassure, comply, or minimize distress while vocal delivery, facial behavior, posture, or response latency suggests otherwise. This raises a focused question: whether multimodal large language models (MLLMs) can arbitrate cross-channel affective conflict rather than recognize modality-congruent emotion. We introduce AMBench, a benchmark for observable affective incongruence in counseling-session media. AMBench contains 3,597 human-audited client utterance segments from 76 public-source counseling videos and defines a main task: whether the client's verbal/social presentation is affectively congruent with synchronized vocal and visual evidence. It also provides model-mediated auxiliary affect references, including soft emotion arrays and Valence–Arousal–Dominance (VAD) scores, while treating the human-audited incongruence label as the official target. A blind reliability pass by ten psychology-qualified internal research-team annotators reports Fleiss' $\kappa = 0.7571$, confidence-stratified agreement 0.8065, and human upper-bound Accuracy 0.7833 under the same task contract. AMBench turns a counseling-relevant but under-measured phenomenon into a reproducible benchmark for cross-modal arbitration in MLLMs.

1 Introduction

Multimodal affective computing commonly treats language, voice, and vision as mutually reinforcing evidence for a single affect label. This framing supports standard emotion recognition and sentiment analysis, but becomes fragile when channels disagree. In counseling dialogue, a client may say that they are doing fine while their voice trembles, their gaze darts away, a strained smile appears, or restless leg movement and tense posture suggest visible unease. A text-dominant model may judge the utterance as neutral or positive, although synchronized vocal and visual evidence calls for cross-channel arbitration rather than simple label recognition.

We focus on this failure mode as *observable affective incongruence*: an annotatable mismatch, within an utterance-level interaction window, between verbal or social presentation and observable vocal or visual affective evidence. The definition is deliberately operational: AMBench does not infer private mental states, diagnose individuals, or assume that nonverbal behavior is more authentic than language. It tests whether the available channels support incompatible affective readings.

This target is timely because end-to-end multimodal LLMs, or omni-models, now ingest raw audio and video rather than only transcripts or hand-designed features. A benchmark for these systems should therefore test not only whether they produce the correct decision, but whether they remain calibrated when lexical, acoustic, and visual cues conflict under the same input contract. Counseling sessions are a dense testbed: guarded disclosure, compliance, hesitation, minimization, and affect masking can separate what is said from how it is delivered.

AMBench turns this arbitration problem into a benchmark. From public-source counseling-session videos, we reconstruct client utterance segments with immediately preceding counselor context so that response latency, silence, and preparatory behavior remain available. We separate text, audio, and visual evidence during construction, au-

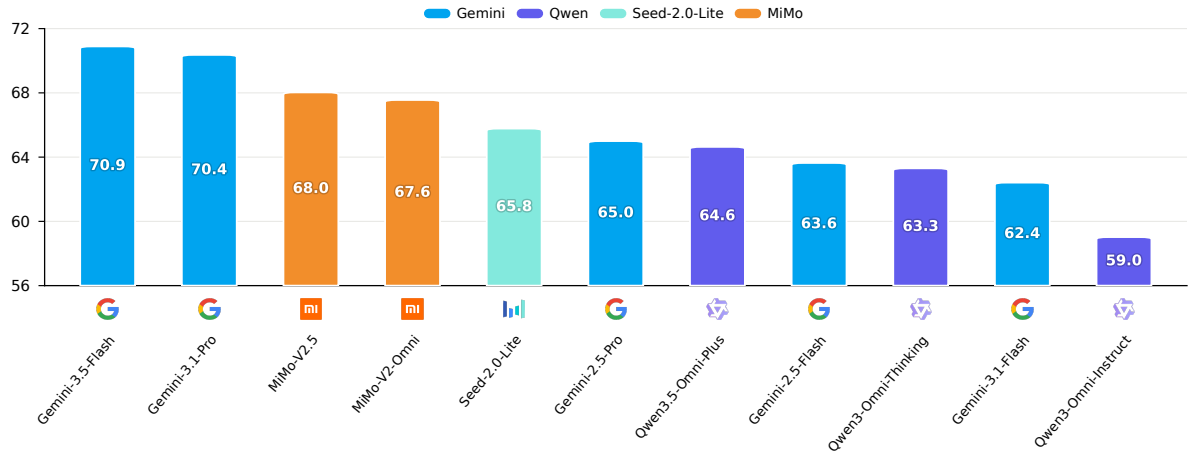

Figure 1: **Task 1 Accuracy.** Final ranking for the eleven main E2E settings. Bars show full-test-set Accuracy percentages and share colors within model families; the y-axis is zoomed to make rank differences visible without changing the reported values.

dit the final incongruence label for every item, and retain model-mediated auxiliary affect references for operating-point and stress-test analyses. The official task is observable-incongruence detection; auxiliary references are not independent human gold labels.

The benchmark also exposes two practical pressures in raw-stream MLLM evaluation: *reasoning intensity*, where model-level thinking settings can change deliberation over conflicting cues, and *temporal sampling*, where frame-rate choices determine how much brief posture, facial behavior, and response-latency evidence is visible. AMBench measures both under a fixed task contract, while avoiding causal attribution of any individual decision to a single modality.

Our contributions are:

- We introduce AMBench, a counseling-domain benchmark of observable affective incongruence with 3,597 human-audited client utterance segments from 76 public-source videos.
- We provide a construction protocol that preserves interaction context, separates text/audio/video evidence during construction, and treats the human-audited incongruence label as the official target.
- We evaluate eleven raw-stream end-to-end multimodal LLM settings across four model families, with controls for reasoning intensity, restricted input, and frame rate.
- We report reliability and operating-point evidence showing differences in Accuracy, recall/precision trade-offs, temporal sampling sensitivity, and auxiliary affect-reference alignment.

2 Related Work

Multimodal affect benchmarks. A long line of multimodal affect benchmarks has shaped how the field aligns linguistic, acoustic, and visual evidence for emotion, sentiment, or interview-state prediction. IEMOCAP, CMU-MOSEI, MELD, CH-SIMS, and DAIC-WOZ provide widely used foundations for multimodal emotion and sentiment modeling (Busso et al., 2008; Bagher Zadeh et al., 2018; Poria et al., 2019; Yu et al., 2020; Gratch et al., 2014). Recent MLLM-centered resources extend this line from closed-set labels toward instruction following, free-form emotion description, and broader emotional-intelligence evaluation. Emotion-LLaMA introduces instruction-tuned audiovisual emotion recognition and reasoning (Cheng et al., 2024); AffectGPT and MER-UniBench connect multimodal emotion recognition with descriptive emotion understanding (Lian et al., 2025a); MER2025 frames affective computing around LLM-driven generative prediction and multimodal evidence descriptions (Lian et al., 2025b); and EmoBench-M and MME-Emotion evaluate emotional intelligence across multimodal, social, and reasoning-oriented tasks (Hu et al., 2025; Zhang et al., 2026). AMBench builds on this evaluation landscape but makes a narrower diagnostic target explicit: whether a model detects an observable mismatch between verbal/social presentation and synchronized vocal or visual affective evidence in counseling utterances.

Cross-modal conflict and affective mismatch. Multimodal meaning is not always additive. Sarcasm, humor, and emotion-cause analysis have long used incongruity as a signal for pragmatic or affective interpretation (Castro et al., 2019; Hasan et al., 2019; Pan et al., 2020; Wang et al., 2024). A newer line of work studies affective cue conflict more directly. CADD formalizes directional text-audio affective dissonance for reflective journaling (Lee, 2026); CA-MER evaluates MLLMs under audiovisual emotion-conflict conditions (Han et al., 2025); AVERE/EmoReAlM tests cue–emotion associations, audiovisual hallucination, and modality agreement (Chaubey et al., 2026); EmoMM probes MLLMs under modality-aligned, conflicting, and missing-input settings (Sun et al., 2026); and LISTEN separates lexical from acoustic emotion cues in audio LLMs (Chen et al., 2026). AMBench is aligned with this conflict-oriented literature, while focusing on public counseling-session media, client-utterance reconstruction, and human-audited observable affective incongruence as the official benchmark target.

Counseling-domain LLM benchmarks. Counseling-domain benchmarks bring affective evaluation closer to sensitive human-support interactions where language, emotion, and support quality are intertwined. RealCBT provides the public-source counseling dataset underlying AMBench and analyzes emotional arcs in real versus LLM-generated CBT sessions (Wang et al., 2025). CBT-Bench evaluates LLMs on CBT knowledge, cognitive-model understanding, and therapeutic response generation (Zhang et al., 2025); CounselBench uses expert evaluation and adversarial questions to stress-test open-ended support-dialogue QA (Li et al., 2026); and MindDialog builds a large psychotherapy-dialogue benchmark with expert-validated understanding and reasoning-generation tasks (Hu et al., 2026). These resources establish the need for domain-specific evaluation and expert validation. AMBench asks a complementary pre-response question: Before generating advice or a counseling reply, can a raw-stream multimodal system arbitrate cross-channel affective evidence in the client utterance itself?

Affective reasoning, judging, and operating points. Text-centered benchmarks evaluate LLMs on empathy, emotional intelligence, and cognitive appraisal beyond basic label prediction (Chen et al., 2024; Sabour et al., 2024; Yeo and Jaidka, 2025). At the same time, multimodal learning research has long treated representation, alignment, and fusion as central design problems (Baltrušaitis et al., 2019). In modern MLLMs, these concerns meet a second layer: the same model often acts as both a perceptual system and a judge. Prior analyses suggest that emotion representations and verbalized reasoning can diverge in LLMs (Maheswaran and Desarkar, 2026), and expert support-dialogue evaluation shows that automated judges may miss safety-relevant concerns or overestimate response quality (Li et al., 2026). AMBench therefore reports not only a single official score, but also operating-point behavior and auxiliary affect-reference alignment under a fixed task contract.

3 AMBench

AMBench converts 76 public-source counseling-session videos into 3,597 utterance-level raw-stream benchmark items for observable affective incongruence. We reconstruct client-response windows, extract modality-isolated construction evidence, build model-mediated draft targets, and audit every final incongruence label. The result is a reproducible benchmark for cross-modal affective arbitration over synchronized language, voice, vision, and timing.

3.1 Definition and Scope

In AMBench, affective incongruence is present when verbal semantics and one or more synchronized nonverbal channels provide incompatible affective evidence. Examples include positive or neutral wording paired with vocal strain, low-arousal disengagement, facial tension, avoidance, or delayed response behavior. Incongruence is distinct from uncertainty, missing modality, noisy ASR, low confidence, or ambiguous but non-conflicting affect, which remain quality factors rather than the benchmark target.

The benchmark schema keeps fields such as `surface_emotion` and `true_emotion` for annotation-pipeline compatibility; here, they are expressed-affect and inferred-affect references produced by a reviewed model-mediated process. The official task is observable-incongruence detection.

The annotation unit is a client utterance plus its immediately preceding counselor context. This interactional unit allows a pause before speaking, a visually restrained response, or a shift in vocal in-

tensity to serve as evidence for the response. We exclude ordinary semantic contradiction, topic disagreement, low audio/video quality, and ambiguous but non-conflicting affect from the positive class. Weak evidence may lower confidence, but is not counted as incongruence.

3.2 Source and Temporal Reconstruction

During data construction, AMBench uses Real-CBT (Wang et al., 2025) as a source manifest: it provides public video URLs and coarse ASR transcripts for 76 counseling-session videos. It does not provide AMBench's utterance-level multimodal boundaries, raw-stream clips, modality-isolated construction evidence, auxiliary affect references, or audited incongruence labels. We realign the coarse dialogue transcript with fine-grained ASR-derived timing candidates, including Qwen3-ASR, MiMo-V2.5-ASR, and VibeVoice-ASR (Shi et al., 2026; LLM-Core-Team Xiaomi, 2026; Peng et al., 2026), to reconstruct counselor/client turn boundaries and slice each client utterance into a local interaction window.

The slicing protocol uses pre-speech extension: a client clip begins when the counselor finishes speaking rather than when the client starts speaking. This preserves response latency, silence, preparatory motion, and breathing patterns that may be affectively meaningful. After quality filtering and final audit, AMBench contains 3,597 utterance-level client segments.

This reconstruction is load-bearing: starting at the first recognized word loses latency cues that distinguish guarded or reluctant responses from fluent neutral answers, while extending too far back can contaminate a client segment with counselor behavior. We therefore use the counselor turn boundary as the extension anchor and keep the target label attached to the client response.

3.3 Modality-Isolated Intermediate Evidence

To reduce semantic leakage during target construction, AMBench separates intermediate evidence by modality. Text evidence consists of the ASR dialogue transcript, client utterance, immediately preceding counselor context, and timing metadata. Audio-only evidence contains continuous V-A-D estimates, within-utterance volatility, session-normalized Z-score and Percentile references, and acoustic emotion label distributions. Video-only evidence is obtained in muted *deaf observer* mode and records visual V-A-D annotations; body-movement, posture, and facial-expression descriptions while the client listens to the counselor question before responding; corresponding descriptions during the client response; and an overall model-generated description of observable movement, posture, and facial behavior.

These artifacts are construction-time references, not inputs supplied to evaluated end-to-end systems. Session-normalized audio features help target-construction judges compare a segment against the same session rather than a population average. Likewise, the visual descriptions are not individual assessments of the client, but structured accounts of observable behavior under muted video. These choices keep AMBench focused on evidence relations among modalities while preserving raw-stream evaluation as the official test setting.

Figure 2 summarizes the pipeline: intermediate text, audio, and video evidence supports model-mediated target construction and human audit, while the final benchmark is evaluated with raw multimodal clips under a fixed end-to-end task contract.

3.4 Target Provenance and Quality Control

AMBench should be read as a human-audited incongruence benchmark with model-mediated auxiliary affect references. The incongruence label is directly audited for all items. Soft emotion arrays and VAD targets preserve reviewed model-side references for multi-reference evaluation and error analysis, but are not independent human labels.

After modality-isolated evidence is assembled, GLM-5.1 (Z.AI, 2026) and GPT-5.5 (OpenAI, 2026) independently produce draft judgments under matched prompts and schemas. They agree on 2,805 of 3,597 items (78.0%), which enter final human audit directly. The remaining 792 items (22.0%) are sent to Claude-Opus-4.7 (Anthropic, 2026) for third-party arbitration before audit. All 3,597 retained items then receive final human audit of the official incongruence label and confidence tag.

Before each major handoff, team members spot-check ASR alignment and slicing, audio V-A-D/Z-score/label artifacts, video V-A-D/action/observer-report artifacts, dual-model agreement, third-party arbitration, and final audit sheets. A stage advances only after these artifacts are usable for the next construction step. This process is a quality-control gate, not a claim that every auxiliary emotion array or continuous VAD reference is independently

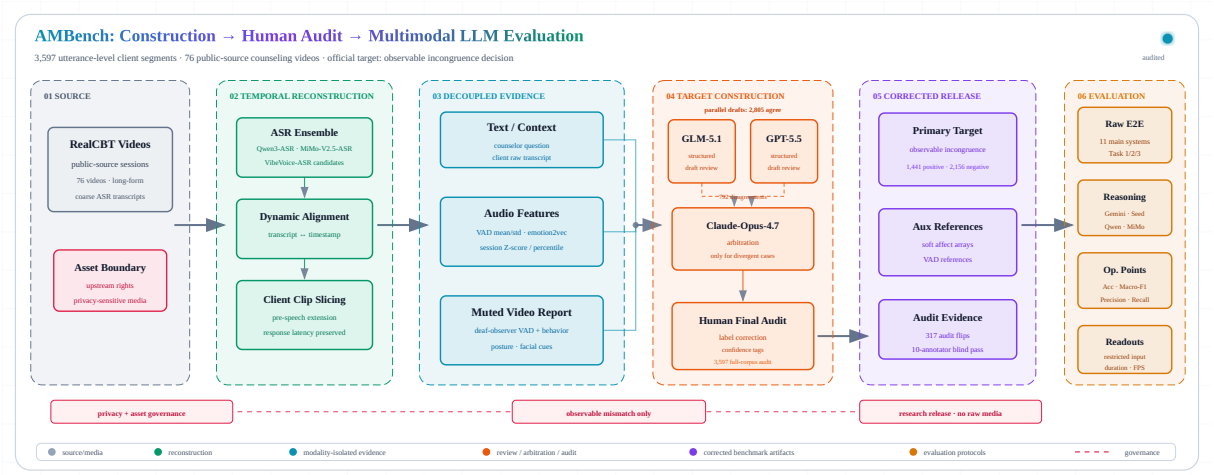

Figure 2: AMBench construction and evaluation pipeline. Public-source counseling videos are reconstructed into utterance-level client segments, processed through modality-isolated text/audio/video evidence extraction, reviewed through structured model-mediated target construction and human audit, and then used for raw-stream end-to-end evaluation of observable affective incongruence.

relabeled by humans.

The final audit materially changes the benchmark: 317 task labels are flipped relative to pre-audit fused outputs, including 175 incongruence-to-non-incongruence and 142 non-incongruence-to-incongruence corrections. The human review stage is adjudicative rather than a fully independent multi-rater study: review sheets expose transcript/context, structured unimodal evidence, temporary fused judgments, and upstream draft-judgment records, so the audit is not blind to upstream model outputs. We treat this anchoring risk with two checks. First, the 317 post-audit label flips (8.8% of the corpus) demonstrate substantive human authority rather than rubber-stamp confirmation. Second, a blind reliability pass by ten psychology-qualified internal professional annotators over all 3,597 benchmark items reports Fleiss' $\kappa = 0.7571$ for the official incongruence label, confidence-stratified agreement 0.8065, and human upper-bound Accuracy 0.7833 under the same task contract. This blind pass is separate from the model-visible construction audit and did not expose annotators to upstream model drafts, fused judgments, rationales, or arbitration records. We therefore retain "human-reviewed" as the operative term while reporting the blind pass as independent reliability evidence.

4 Experiments

Evaluation questions. AMBench evaluates raw-stream multimodal systems under one fixed task contract: given a reconstructed client-utterance clip, return a judgment of observable affective incongruence and a compact set of auxiliary affect fields. The experiments therefore address three separate questions. The main leaderboard asks which E2E systems best perform the official Task 1 decision-making. Operating-profile metrics show whether a score comes from balanced detection, conservative under-calling, or broad over-calling. Targeted controls then test reasoning intensity, restricted input streams, and video frame-rate sensitivity; these controls are not a parallel leaderboard.

Table 1: AMBench benchmark summary. Counts and reliability values refer to the corrected 3,597-item benchmark; the incongruence label is the official Task 1 target.

Statistic	Value
Corpus scale	
Public-source counseling videos	76
Utterance-level client segments	3,597
Human-audited task labels	3,597
Official Task 1 target	
No observable incongruence	2,156 (59.9%)
Observable incongruence	1,441 (40.1%)
Independent reliability	
Blind Task 1 Fleiss' κ	0.7571
Human upper-bound Accuracy	0.7833

Systems and settings. All E2E prediction outputs are generated under the same raw-stream evaluation protocol: each system receives the reconstructed client-utterance clip, the matched prompt contract, and the same fixed structured output format, after which parsed outputs are scored with the shared metric definitions. The main cross-family comparison contains eleven completed settings: five Gemini systems (Gemini-3.5-Flash, Gemini-3.1-Pro, Gemini-3.1-Flash, Gemini-2.5-Pro, Gemini-2.5-Flash) (Google

DeepMind, 2026), Seed-2.0-Lite (ByteDance Seed Team, 2026), two MiMo systems (MiMo-V2.5 and MiMo-V2-Omni) (Xiaomi MiMo Team, 2026b,a), and three Qwen systems (Qwen3-Omni-Instruct, Qwen3-Omni-Thinking, Qwen3.5-Omni-Plus) (Qwen Team, 2025, 2026). Unless otherwise noted, runs use deterministic decoding with temperature 0.0, top-$p = 1.0$, a maximum output budget of 8,192 tokens, and default 2 FPS video sampling where the FPS is configurable.

Tasks and metrics. Task 1 is observable-incongruence detection and is the official leaderboard target. We rank systems by Accuracy because the corrected Task 1 distribution is moderately imbalanced rather than long-tailed (2,156:1,441). Macro-F1, Precision, and Recall are reported as operating-profile readouts: Macro-F1 summarizes class-balanced decision quality, Precision measures over-attribution risk, and Recall measures sensitivity to positive incongruence cases. Task 2 and Task 3 are auxiliary affect-reference alignment readouts only. Task 2 reports Soft Hit Rate, counted by set intersection against accepted inferred-affect references; Task 3 reports Pearson correlation and Mean Absolute Error against continuous VAD references. These auxiliary fields help interpret affect tracking, but they do not affect the official ranking.

Protocol controls. The evaluation protocol holds data preparation, raw multimodal inference, prediction validation, and scoring fixed across reported rows wherever model interfaces permit. Before scoring, predictions are validated for complete item coverage, duplicate entries, valid structured fields, and parseable VAD outputs. All eleven main rows and the control rows reported below cover the full 3,597-item corrected test set. Confidence tags, auxiliary reference arrays, annotation-time features, and construction-time evidence files are never provided as E2E inputs. Qwen3.5-Omni-Plus is treated as an instruct model without an exposed thinking function.

Ablation and control rows. Beyond the main leaderboard, we report three restricted control families. The reasoning-intensity control compares Gemini-3.1-Pro low, medium, and default/high thinking settings, Seed-2.0-Lite low/default/high reasoning-effort settings, and the Qwen3-Omni-Instruct/Qwen3-Omni-Thinking pair; the Gemini default/high row is the same full-stream setting used in the main leaderboard, while low and medium are additional control runs. The restricted-input control is limited to Gemini-3.1-Pro: audio-only uses the raw waveform without video or transcript input, and subtitled-visual uses muted hard-subtitled video without an audio stream; both use modality-specific prompts. The frame-rate control is also limited to Gemini-3.1-Pro's configurable FPS setting, with Gemini-3.1-Pro evaluated at 1/2/4/8/12 FPS. These rows are excluded from the main leaderboard in Table 2 and are used only to interpret protocol sensitivity.

Uncertainty estimates. For Task 1 we compute paired bootstrap intervals over utterance items with 10,000 resamples and fixed seed `2026`. Intervals are used for the Accuracy leaderboard and selected paired deltas; Macro-F1 deltas are reported only when they clarify operating-point behavior. We report point estimates for all standard metrics, but avoid treating small differences as meaningful when the paired delta interval crosses zero.

5 Results

5.1 Cross-Family Leaderboard and Operating-Point Profiles

Table 2 and Figure 1 report the main E2E results on the full corrected test set, with rows grouped by access and Rank preserving the global Task 1 Accuracy order. Gemini-3.5-Flash achieves the highest Accuracy point estimate at 0.7089, followed closely by Gemini-3.1-Pro at 0.7036, and both remain below the blind human upper-bound Accuracy of 0.7833. We therefore treat the two Gemini rows as a close top tier rather than a sharply separated winner. MiMo-V2.5 (0.6803) and MiMo-V2-Omni (0.6756) form the next tier, with the table retaining the uncertainty estimates for these adjacent comparisons.

The open-source MLLM block presents a different view of the leaderboard. MiMo-V2.5 is the highest-Accuracy open-source MLLM and ranks third overall, but its operating point is highly selective: it has the highest Precision (0.8299), very low Recall (0.2540), and Macro-F1 0.5862. Qwen3-Omni-Thinking improves substantially over Qwen3-Omni-Instruct, moving from Accuracy 0.5902 to 0.6330 and from Recall 0.2144 to 0.6572; the gain is therefore primarily a shift away from a recall-starved decision profile rather than a uniform improvement across all metrics. MiMo-V2-Omni, evaluated as a closed-source MLLM

in this version, has slightly lower Accuracy than MiMo-V2.5 but a more balanced profile (Macro-F1 0.6402, Recall 0.4518).

Accuracy remains the official scalar metric, but the operating-profile columns are necessary for interpretation. Gemini-3.5-Flash, Gemini-2.5-Pro, Gemini-2.5-Flash, Qwen3-Omni-Thinking, and Seed-2.0-Lite (default) are more recall-oriented, whereas Gemini-3.1-Flash and Qwen3.5-Omni-Plus are conservative positive-class detectors. Seed-2.0-Lite (default) and Seed-2.0-Lite (high) achieve the highest Soft Hit Rates, but auxiliary inferred-affect matching does not substitute for the official incongruence decision. In practice, Accuracy defines the official ranking, while Macro-F1, Recall, and Precision explain how each system reaches that rank.

Appendix A provides supplementary analyses: a secondary Macro-F1 leaderboard, Accuracy versus model parameter disclosure, and auxiliary VAD-reference alignment. MiMo-V2.5 leads the VAD correlations, while Gemini-3.1-Pro and MiMo-V2-Omni remain close competitors. These auxiliary scores should not be read as independent affect gold labels; they show whether a system tracks the benchmark's inferred-affect references while making the official Task 1 decision.

5.2 Reasoning Intensity: Gemini Accuracy Ladder, Seed Recall Shift, Qwen3-Omni Step Function

Table 3 compares three model families with reasoning-intensity variants. Gemini-3.1-Pro shows the cleanest monotonic ladder: moving from low to medium to default/high thinking raises Accuracy ($0.6895 \rightarrow 0.6997 \rightarrow 0.7036$), Macro-F1 ($0.6364 \rightarrow 0.6583 \rightarrow 0.6759$), and Recall ($0.3838 \rightarrow 0.4386 \rightarrow 0.5128$). The default/high row improves over low by $+0.0142$ Accuracy with a paired 95% interval of $[0.0011, 0.0275]$. The gain is not a pure precision increase: Precision decreases from 0.7072 to 0.6700, indicating a shift toward detecting more positive incongruence cases.

Seed-2.0-Lite shows a different response. Increasing the model reasoning level from low to default to high raises Accuracy only modestly ($0.6500 \rightarrow 0.6578 \rightarrow 0.6594$), while Recall shifts more clearly ($0.5649 \rightarrow 0.6558 \rightarrow 0.6974$). Higher reasoning therefore makes Seed-2.0-Lite call more positive cases while overall Accuracy remains near a plateau. For the Qwen3-Omni family, enabling thinking produces a sharper operating-point change: Accuracy rises from 0.5902 to 0.6330, Macro-F1 from 0.5033 to 0.6288, and Recall from 0.2144 to 0.6572. Together, reasoning intensity is model-dependent: Gemini shows a small accuracy ladder, Seed shifts recall, and Qwen3-Omni changes sharply from instruct to thinking mode.

5.3 Restricted-Input Modality Control

Table 4 compares three input configurations of Gemini-3.1-Pro: the full audio-video-text stream, audio-only (raw waveform without video or transcript), and subtitled-visual (muted video with burned-in transcript subtitles, no audio stream). These rows are restricted-input controls for one architecture, not a full unimodal leaderboard. On the official Accuracy metric, audio-only remains close to the full stream (0.6967 versus 0.7036), and subtitled-visual is lower but still nearby (0.6900). The secondary Macro-F1 and Recall view is sharper: subtitled-visual drops to Macro-F1 0.6347 and Recall 0.3754, whereas audio-only retains Macro-F1 0.6746 and Recall 0.5441. The result shows why AMBench reports Macro-F1 and Recall alongside Accuracy: they reveal positive-class sensitivity loss that is partly masked in the scalar Accuracy ranking.

5.4 Duration and FPS Are Not Monotonic Fixes

Longer clips help auxiliary affect-reference alignment more reliably than observable-incongruence detection. The duration split in Appendix A.3 shows that Soft Hit usually rises on longer clips, while Task 1 gains are model-dependent. Frame-rate scaling is also non-monotonic in the Gemini-3.1-Pro ablation (Figure 8; Table 5): 4 FPS gives the highest Accuracy and Macro-F1 point estimates, whereas denser 8–12 FPS sampling reduces both decision balance and auxiliary affect alignment. More frames therefore do not automatically improve cross-modal arbitration; the model must select useful cues from the stream.

6 Discussion

AMBench's output profiles suggest that observable affective incongruence is a thresholded cross-modal arbitration problem rather than a simple emotion-matching problem. The E2E protocol does not isolate mechanisms causally, but the pattern is consistent: systems must recover cue-conflict evidence from raw streams and decide

Table 2: Overall E2E results on the full corrected AMBench test set (11 main systems; $N = 3{,}597$). Rows are grouped into closed-source and open-source MLLMs, while Rank is the global ordering by Accuracy. Accuracy is the official Task 1 metric; 95% CIs use 10,000 paired bootstrap resamples. Bold marks the best point estimate in each column.

Rank	Models	n	Official		Operating profile			Aux.
			Acc.	95% CI	Macro-F1	Prec.	Rec.	Soft Hit
			Closed-source MLLMs					
1	Gemini-3.5-Flash	3597	**0.7089**	[0.6945, 0.7242]	**0.6989**	0.6313	0.6572	0.8477
2	Gemini-3.1-Pro	3597	0.7036	[0.6889, 0.7187]	0.6759	0.6700	0.5128	0.7334
4	MiMo-V2-Omni	3597	0.6756	[0.6600, 0.6911]	0.6402	0.6333	0.4518	0.7832
5	Seed-2.0-Lite (default)	3597	0.6578	[0.6422, 0.6736]	0.6517	0.5625	0.6558	**0.8852**
6	Gemini-2.5-Pro	3597	0.6500	[0.6344, 0.6656]	0.6498	0.5437	**0.7863**	0.8126
7	Qwen3.5-Omni-Plus	3597	0.6464	[0.6308, 0.6622]	0.5133	0.8073	0.1541	0.7295
8	Gemini-2.5-Flash	3597	0.6364	[0.6202, 0.6522]	0.6357	0.5332	0.7412	0.8490
10	Gemini-3.1-Flash	3597	0.6241	[0.6080, 0.6400]	0.4656	0.7259	0.0992	0.6694
			Open-source MLLMs					
3	MiMo-V2.5	3597	0.6803	[0.6647, 0.6959]	0.5862	**0.8299**	0.2540	0.6617
9	Qwen3-Omni-Thinking	3597	0.6330	[0.6175, 0.6486]	0.6288	0.5341	0.6572	0.6633
11	Qwen3-Omni-Instruct	3597	0.5902	[0.5738, 0.6063]	0.5033	0.4747	0.2144	0.6544

Table 3: Reasoning-intensity ablation. Gemini-3.1-Pro rows use low, medium, and default/high thinking settings, Seed-2.0-Lite rows use low/default/high reasoning levels, and Qwen3-Omni rows compare instruct and thinking modes. All rows cover the 3,597-item corrected test set. The 95% CI column is for Accuracy, the official Task 1 metric; Macro-F1, Precision, and Recall summarize operating points.

Setting	Acc.	95% CI	Macro-F1	Prec.	Rec.
		Gemini-3.1-Pro			
Gemini-3.1-Pro (low)	0.6895	[0.6739, 0.7045]	0.6364	0.7072	0.3838
Gemini-3.1-Pro (medium)	0.6997	[0.6850, 0.7148]	0.6583	0.6999	0.4386
Gemini-3.1-Pro (default/high)	**0.7036**	[0.6889, 0.7187]	0.6759	0.6700	0.5128
		Seed-2.0-Lite			
Seed-2.0-Lite (low)	0.6500	[0.6347, 0.6656]	0.6358	0.5629	0.5649
Seed-2.0-Lite (default)	0.6578	[0.6422, 0.6736]	0.6517	0.5625	0.6558
Seed-2.0-Lite (high)	0.6594	[0.6439, 0.6753]	0.6560	0.5602	0.6974
		Qwen3-Omni			
Qwen3-Omni-Instruct	0.5902	[0.5738, 0.6063]	0.5033	0.4747	0.2144
Qwen3-Omni-Thinking	0.6330	[0.6175, 0.6486]	0.6288	0.5341	0.6572

Table 4: Restricted-input modality ablation on Gemini-3.1-Pro. Audio-only uses the raw waveform without video or transcript; subtitled-visual uses muted video with burned-in transcript subtitles, no audio stream. Δ Accuracy is relative to the full trimodal baseline.

Configuration	Acc.	95% CI	Δ Acc.	Prec.	Rec.
Full	0.7036	[0.6889, 0.7187]	+0.0000	0.6700	0.5128
Audio-only	0.6967	[0.6817, 0.7117]	-0.0070	0.6437	0.5441
Subtitled-visual	0.6900	[0.6750, 0.7056]	-0.0136	0.7156	0.3754

when that evidence is strong enough to call incongruence. This distinction matters because conservative systems can retain respectable Accuracy while missing many positive cases, as seen in MiMo-V2.5, Qwen3.5-Omni-Plus, and Gemini-3.1-Flash. The controls also argue against a simple monotonic resource account. Reasoning effects are model-dependent: Gemini-3.1-Pro shows a small low-to-default/high Accuracy ladder, Seed-2.0-Lite mainly shifts Recall, and Qwen3-Omni changes sharply when thinking is enabled. Denser frame sampling also does not dominate. Models appear to need better cue selection and calibration, not simply more tokens, frames, or deliberation.

Table 5: Gemini-3.1-Pro FPS ablation on the full 3,597-item test set. Accuracy CI is paired bootstrap; Macro-F1 and Soft Hit are auxiliary readouts.

FPS	n	Acc.	95% CI	Prec.	Rec.	Macro-F1	Soft Hit
1	3597	0.6992	[0.6842, 0.7142]	0.6610	0.5115	0.6717	**0.7670**
2	3597	0.7036	[0.6889, 0.7184]	0.6700	**0.5128**	0.6759	0.7334
4	3597	**0.7073**	[0.6922, 0.7220]	**0.6837**	0.5010	**0.6771**	0.7259
8	3597	0.6808	[0.6656, 0.6959]	0.6381	0.4698	0.6482	0.6820
12	3597	0.6750	[0.6594, 0.6903]	0.6262	0.4684	0.6429	0.6806

7 Conclusion

AMBench turns observable affective incongruence in counseling-session media into a benchmark for multimodal LLMs, with human-audited task labels and model-mediated auxiliary affect references. Across eleven E2E settings and three ablation/control groups, it exposes conservative-versus-recall-oriented operating points, model-dependent reasoning effects including a Gemini-3.1-Pro low-to-default/high Accuracy ladder, stronger acoustic than subtitled-visual evidence in the Gemini restricted-input control, and non-monotonic FPS behavior. These results show that current MLLMs still struggle to arbitrate cross-channel affective evidence under a fixed task contract. AMBench provides a concrete benchmark resource for future work on observable-feature isolation, cross-modal arbitration, and safer affect-sensitive evaluation.

Limitations

AMBench is not an independent gold standard for private internal states and must not be interpreted as assessing individual speakers beyond the observable task definition. The incongruence label is human-audited for all benchmark items, and a separate blind full-corpus reliability pass by ten psychology-qualified internal annotators reports Fleiss' $\kappa = 0.7571$ and human upper-bound Accuracy 0.7833; however, soft emotion arrays and VAD references remain model-mediated auxiliary references. The release does not evaluate subtype classification. The session-level audio normalization features also do not model phase-specific counseling baselines such as rapport building, problem exploration, or closing segments. This version still does not report broader cross-system unimodal evaluation, classical multimodal fusion baselines, external construct validation, or pre-speech extension ablation. Some evaluated model configurations remain partially opaque: model-level thinking controls and hidden serving stacks are not directly inspectable, so we report observed model settings rather than attributing differences to a single causal mechanism.

Ethical Considerations

AMBench is derived from public-source counseling videos, but public availability does not eliminate privacy or dignity risks. Counseling interactions can contain sensitive self-disclosures, and benchmark release can increase the risk of re-identification, stigmatization, or misuse. During blind review, the release surface is described without author-identifying repository links; anonymized supplementary materials provide the source URL manifest, fine-grained dialogue/timing metadata, and reconstruction/slicing scripts, and the public repository URL will be provided after acceptance. AMBench does not treat the source media as unrestricted AMBench-owned assets, does not redistribute raw or clipped audio/video, and does not release generation-time audio/video feature files, intermediate adjudication features, reasoning traces, rationales, or dispute sheets used during label construction. AMBench is intended for research evaluation, auditing, and benchmark analysis only. It should not be used for individual assessment, screening, intervention planning, hiring, educational, insurance, legal, or other consequential decisions about individuals. Any release of derived artifacts should respect upstream platform terms, uploader rights, controlled-access requirements for sensitive media, and a takedown or correction process for affected parties.

References

Anthropic. 2026. Claude-Opus-4.7. https://docs.anthropic.com/en/docs/about-claude/models/overview. Official model documentation; accessed 2026-05-21.

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

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

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

ByteDance Seed Team. 2026. Seed2.0. https://seed.bytedance.com/en/seed2. Accessed 2026-04-21.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an Obviously perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Ashutosh Chaubey, Jiacheng Pang, Maksim Siniukov, and Mohammad Soleymani. 2026. Improving audio-visual emotion reasoning with preference optimization. In *Proceedings of the Fourteenth International Conference on Learning Representations*.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2026. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5848–5877, Rabat, Morocco. Association for Computational Linguistics.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024. EmotionQueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.

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

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Google DeepMind. 2026. Model cards. https://deepmind.google/models/model-cards/. Accessed 2026-04-21.

Jonathan Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David R. Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. 2025. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5528–5537. Association for Computing Machinery.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. 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)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

He Hu, Juzheng Si, Qianning Wang, Tengjin Weng, Yihong Ji, Jiyue Jiang, Fei Ma, Yucheng Zhou, Laizhong Cui, and Qi Tian. 2026. MindDialog: A large-scale benchmark for counseling dialogue understanding and generation. *Pattern Recognition*, 179:113766.

He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian, Yucheng Zhou, and Laizhong Cui. 2025. EmoBench-M: Benchmarking emotional intelligence for multimodal large language models. *Preprint*, arXiv:2502.04424. ArXiv:2502.04424v4.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Sumin Lee. 2026. I'm fine, but my voice isn't: Cross-modal affective dissonance detection for reflective journaling. *Preprint*, arXiv:2604.27517.

Yahan Li, Jifan Yao, John Bunyi, Adam Frank, Angel Hwang, and Ruishan Liu. 2026. CounselBench: A large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. In *Proceedings of the Fourteenth International Conference on Learning Representations*.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 2025a. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 36993–37014. PMLR.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, Ziyang Ma, Xiaojiang Peng, Xie Chen, Ya Li, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2025b. MER 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13837–13842. Association for Computing Machinery.

LLM-Core-Team Xiaomi. 2026. MiMo-V2.5-ASR: Robust speech recognition across languages, dialects, and complex acoustic scenarios. https://github.com/XiaomiMiMo/MiMo-V2.5-ASR. Official repository and model card; accessed 2026-05-23.

Aishwarya Maheswaran and Maunendra Sankar Desarkar. 2026. A unified view on emotion representation in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3589–3610, Rabat, Morocco. Association for Computational Linguistics.

OpenAI. 2026. GPT-5.5. https://platform.openai.com/docs/models. Official model documentation; accessed 2026-05-21.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.

Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, Yutao Sun, Hangbo Bao, Weijiang Xu, Yi Zhu, Zehua Wang, Ting Song, Yan Xia, Zewen Chi, Shaohan Huang, and 5 others. 2026. VIBEVOICE-ASR technical report. *Preprint*, arXiv:2601.18184. ArXiv:2601.18184v2.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3-Omni-30B-A3B-Instruct model card. https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct. Official model card; accessed 2026-04-21.

Qwen Team. 2026. Qwen3.5-Omni technical report. *Preprint*, arXiv:2604.15804. ArXiv:2604.15804v2; reports the Qwen3.5-Omni family including Qwen3.5-Omni-plus.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-ASR technical report. *Preprint*, arXiv:2601.21337. ArXiv:2601.21337v2.

Yueru Sun, Yimeng Zhang, Haoyu Gu, Nuo Chen, Dong She, Xianrong Yao, Yang Gao, and Zhanpeng Jin. 2026. EmoMM: Benchmarking and steering MLLM for multimodal emotion recognition under conflict and missingness. *Preprint*, arXiv:2605.01024.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. SemEval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050, Mexico City, Mexico. Association for Computational Linguistics.

Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. 2025. Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19999–20017, Suzhou, China. Association for Computational Linguistics.

Xiaomi MiMo Team. 2026a. MiMo-V2-Omni. https://mimo.xiaomi.com/. Official Xiaomi MiMo platform; accessed 2026-05-21.

Xiaomi MiMo Team. 2026b. MiMo-V2.5 model card. https://huggingface.co/XiaomiMiMo/MiMo-V2.5. Official model card; accessed 2026-05-23.

Gerard Christopher Yeo and Kokil Jaidka. 2025. Beyond context to cognitive appraisal: Emotion reasoning as a theory of mind benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26517–26525, Vienna, Austria. Association for Computational Linguistics.

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

Z.AI. 2026. GLM-5.1. `https://docs.z.ai/guides/llm/glm-5.1`. Official model documentation; accessed 2026-05-21.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, and 2 others. 2026. MME-Emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *Proceedings of the Fourteenth International Conference on Learning Representations*.

Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C. Chiu, Shaun M. Eack, Fei Fang, William Yang Wang, and Zhiyu Chen. 2025. CBT-Bench: Evaluating large language models on assisting cognitive behavior therapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3864–3900, Albuquerque, New Mexico. Association for Computational Linguistics.

A Supplementary Results

A.1 Completed E2E Accuracy Overview

Figure 3 summarizes the eleven completed E2E settings with paired bootstrap intervals for Task 1 Accuracy, the official leaderboard metric. Figure 4 provides the corresponding secondary Macro-F1 view.

A.2 High-Confidence Sensitivity Analysis

The official leaderboard in the main paper uses the full 3,597-item corrected test set and ranks systems by Accuracy. Table 6 reports a sensitivity analysis on the 3,135 high-confidence items only. Absolute Accuracy is higher for every model in this confidence-filtered subset, but the subset is not used as the primary benchmark surface. The ranks in this table are Accuracy ranks. The small rank exchange between MiMo-V2.5 and MiMo-V2-Omni on this subset does not change the main conclusion: MiMo-V2.5 is more conservative and precision-oriented, whereas MiMo-V2-Omni is more balanced on Macro-F1 and Recall.

A.3 Duration Split

Table 7 and Figure 7 summarize the short-context versus long-context split for completed settings. Longer context improves auxiliary affect-reference matching more consistently than observable-incongruence detection.

A.4 Minimum Detectable Effect Size

To make the resolution of the benchmark explicit, we compute the minimum detectable effect size (MDES) on the official Accuracy metric at the current corpus size of $N = 3{,}597$. For every unordered pair of the 11 reported main E2E settings, we re-run the paired bootstrap protocol used in the main paper (10,000 paired resamples, fixed seed `2026`) and record the observed point delta, the bootstrap standard error of the paired Accuracy delta, and the resulting 95% interval. Across all $\binom{11}{2} = 55$ pairs, the Accuracy bootstrap SE has median 0.0095. For completeness, the supplemental Macro-F1 readout has median SE 0.0104 and MDES ≈ 0.0291.

What the benchmark can and cannot resolve. Treating the median SE as the typical resolution and using $z_{\alpha/2} + z_{\beta} = 1.96 + 0.84 = 2.80$ for a two-sided test at $\alpha = 0.05$ and power 0.80, the MDES on paired Δ Accuracy at the current sample size is

$$\mathrm{MDES} = 2.80 \times 0.0095 \approx 0.0265,$$

so Accuracy differences smaller than roughly 2.7 points should be interpreted together with the empirical paired interval rather than as standalone ranking evidence. This is the operational reading of the main leaderboard: Gemini-3.5-Flash has the highest Accuracy point estimate, but its paired delta over Gemini-3.1-Pro is only $+0.0054$ with an interval crossing zero. Its separations from MiMo-V2.5 ($+0.0287$) and MiMo-V2-Omni ($+0.0335$) are larger and have positive intervals. MiMo-V2.5 and MiMo-V2-Omni are not separated on Accuracy ($+0.0047$, interval $[-0.0106, +0.0197]$). Larger secondary Macro-F1 separations, such as MiMo-V2-Omni over MiMo-V2.5 ($+0.0539$), are operating-point evidence rather than the official scalar ranking.

Required corpus size to resolve a target effect. Under the same SE budget and assuming SE scales as $1/\sqrt{N}$, Table 8 reports the corpus size needed to detect a target Δ Accuracy with the same 80% power and $\alpha = 0.05$. Detecting $\Delta = 0.020$ requires approximately $1.8\times$ the current corpus; detecting $\Delta = 0.010$ requires roughly $7.0\times$ the current size; detecting $\Delta = 0.005$ requires roughly $28.1\times$. These calculations are planning heuristics, not replacements for paired intervals on observed systems.

Practical reading. AMBench at its current size separates several model-family gaps reported in the main paper, but it is intentionally underpowered for sub-percent ranking. We recommend that downstream comparisons against AMBench report paired bootstrap intervals together with point estimates and treat any Accuracy $|\Delta| \leq 0.020$ as inconclusive at the current corpus size unless the empirical paired interval is clearly separated.

A.5 Ablation and Control Summary

Figure 8 collects the three bounded control families reported in the main paper rather than introducing a separate appendix-only result. Panel a expands Table 3 by plotting the reasoning-intensity rows for Gemini-3.1-Pro low, medium, and default-/high thinking; Seed-2.0-Lite low/default/high reasoning; and Qwen3-Omni-Instruct/Qwen3-Omni-Thinking. Panel b visualizes the Gemini-3.1-Pro restricted-input controls in Table 4, where audio-only remains close to full audio-video-text input

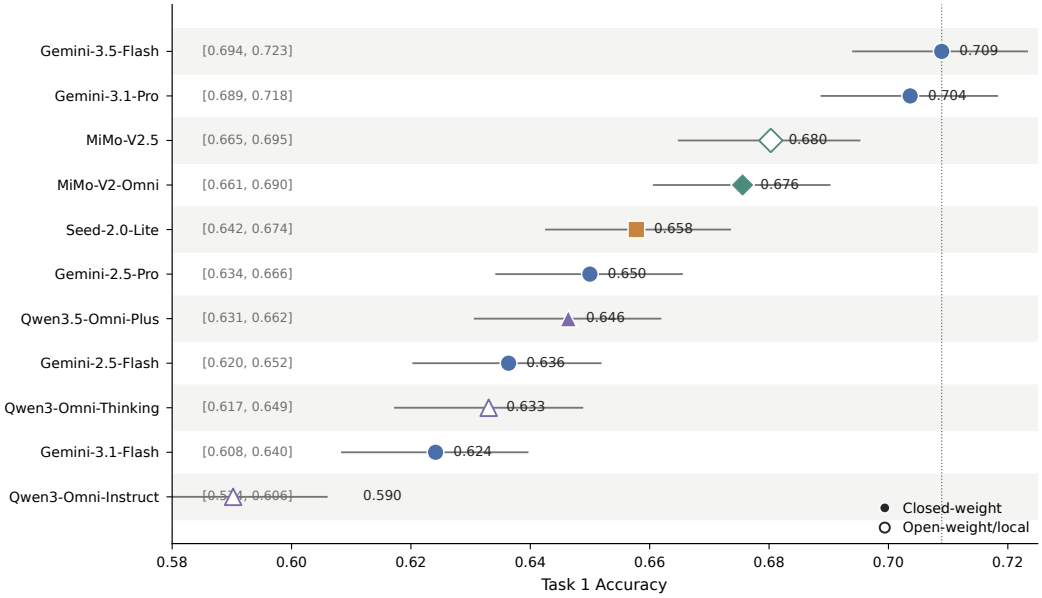

Figure 3: Overall Task 1 Accuracy with paired bootstrap 95% intervals for the eleven completed E2E settings. This is the official scalar leaderboard view; hollow markers denote open-weight/local systems, and the main paper shows a full-width hero ranking.

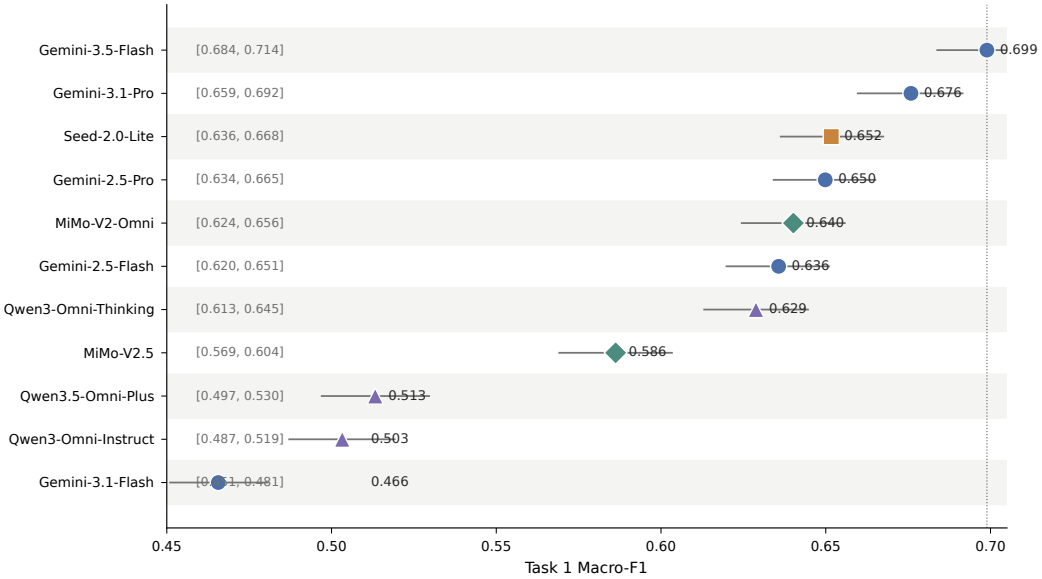

Figure 4: Secondary Task 1 Macro-F1 view with paired bootstrap 95% intervals. Macro-F1 is not used for the official ranking in this version, but it reveals minority-class recall collapse and conservative operating points.

Table 6: High-confidence sensitivity analysis. Main claims use the full corrected test set; this subset analysis is reported only to show robustness to confidence filtering. Rank columns are Accuracy ranks.

Model	Full Acc. rank	High Acc. rank	n	Full Acc.	High Acc.	Δ Acc.	Full Macro-F1	High Macro-F1
Gemini-3.5-Flash	1	1	3135	0.7089	0.7238	+0.0148	0.6989	0.7134
Gemini-3.1-Pro	2	2	3135	0.7036	0.7155	+0.0118	0.6759	0.6894
MiMo-V2.5	3	4	3135	0.6803	0.6903	+0.0100	0.5862	0.5996
MiMo-V2-Omni	4	3	3135	0.6756	0.6912	+0.0157	0.6402	0.6551
Seed-2.0-Lite (default)	5	5	3135	0.6578	0.6730	+0.0153	0.6517	0.6666
Gemini-2.5-Pro	6	6	3135	0.6500	0.6654	+0.0154	0.6498	0.6651
Qwen3.5-Omni-Plus	7	8	3135	0.6464	0.6526	+0.0063	0.5133	0.5244
Gemini-2.5-Flash	8	7	3135	0.6364	0.6530	+0.0166	0.6357	0.6520
Qwen3-Omni-Thinking	9	9	3135	0.6330	0.6472	+0.0142	0.6288	0.6426
Gemini-3.1-Flash	10	10	3135	0.6241	0.6265	+0.0023	0.4656	0.4714
Qwen3-Omni-Instruct	11	11	3135	0.5902	0.5927	+0.0024	0.5033	0.5052

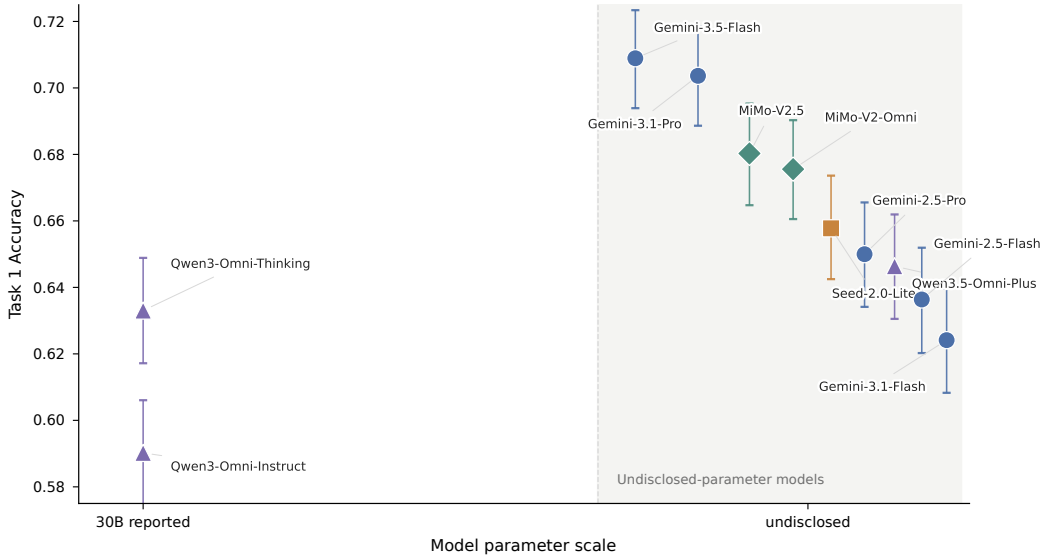

Figure 5: Task 1 Accuracy versus model parameter scale. Qwen3-Omni rows use the reported 30B scale for parameter-axis placement; models without comparable public parameter counts are placed in an "undisclosed" bucket rather than assigned guessed sizes. Error bars show paired bootstrap 95% intervals for Accuracy.

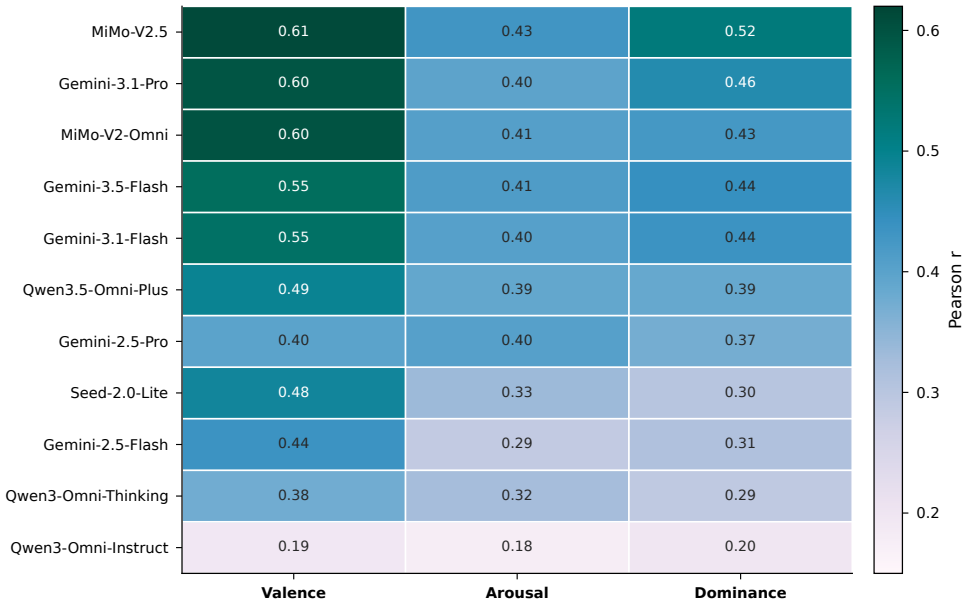

Figure 6: Auxiliary VAD-reference Pearson correlations for completed E2E settings.

Table 7: Duration-split comparison on short-context and long-context clips.

Model	Short Context (< 15s) Recall	Macro-F1	Soft Hit	Long Context (≥ 15s) Recall	Macro-F1	Soft Hit
Gemini-3.5-Flash	0.6510	**0.6683**	0.7974	0.6638	**0.7343**	0.9055
Gemini-3.1-Pro	0.5221	0.6609	0.6779	0.5029	0.6932	0.7972
MiMo-V2.5	0.2403	0.5787	0.5517	0.2687	0.5940	0.7883
MiMo-V2-Omni	0.5315	0.6492	0.7382	0.3664	0.6262	0.8349
Seed-2.0-Lite (default)	0.6940	0.6220	**0.8379**	0.6149	0.6850	**0.9396**
Gemini-2.5-Pro	**0.7879**	0.5910	0.7418	**0.7845**	0.7165	0.8941
Qwen3.5-Omni-Plus	0.1329	0.4984	0.6291	0.1767	0.5287	0.8451
Gemini-2.5-Flash	0.7570	0.5818	0.8000	0.7241	0.6963	0.9055
Qwen3-Omni-Thinking	0.6362	0.5910	0.5527	0.6796	0.6724	0.7911
Gemini-3.1-Flash	0.0644	0.4421	0.5704	0.1365	0.4876	0.7835
Qwen3-Omni-Instruct	0.2228	0.5128	0.5451	0.2055	0.4924	0.7800

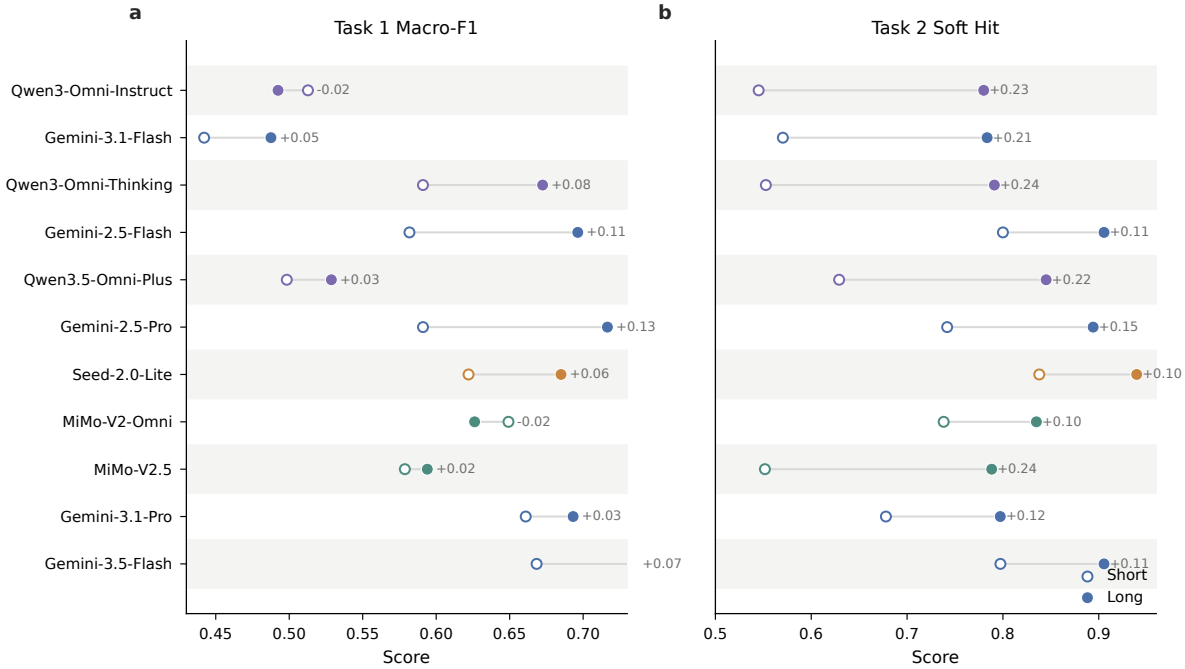

Figure 7: Impact of short-context and long-context clips on Task 1 Macro-F1 for completed E2E settings.

Table 8: Required corpus size N (median-SE scaling) to detect a target Δ Accuracy at $\alpha = 0.05$ two-sided and power 0.80. Median paired-bootstrap SE on the current $N = 3{,}597$ corpus is 0.0095 across $\binom{11}{2} = 55$ system pairs on the official Accuracy metric.

Target Δ Accuracy	Required N	Factor of current N
0.050	1,011	0.3×
0.030	2,807	0.8×
0.020	6,316	1.8×
0.010	25,261	7.0×
0.005	101,043	28.1×
0.0054 (Gemini-3.5-Flash – Gemini-3.1-Pro)	88,043	24.5×

while subtitled-visual input loses positive-class sensitivity. Panel c visualizes the Gemini-3.1-Pro 1/2/4/8/12FPS sweep in Table 5; the trend is non-monotonic, with 4FPS giving the best Task 1 Accuracy point estimate and denser 8–12FPS sampling reducing Accuracy, Macro-F1, and auxiliary Soft Hit.

A.6 Reasoning-Mode Setting Notes

Table 9 records the model configuration labels used for the reasoning-intensity rows. These labels identify evaluated model settings or exposed serving modes; they should not be read as fixed token budgets.

A.7 Reference Baselines and Restricted-Input Controls

Table 10 reports decision-theoretic reference baselines and restricted-input controls for Gemini-3.1-Pro. The baselines anchor reference operating points; the restricted-input rows are single-model controls for input packaging, not a cross-system unimodal leaderboard.

Decision-theoretic reference points. The majority-class baseline (always predict non-incongruent) achieves Accuracy $= 0.5994$ and Macro-F1 $= 0.3748$ on the 60:40 class split, because predicting all negatives produces Precision $= 0$ and Recall $= 0$ for the positive class, dragging F1 to zero for that class. The all-positive baseline achieves Accuracy $= 0.4006$ and Macro-F1 $= 0.2860$: it recovers all incongruent items but generates many false positives, collapsing precision. Uniform random prediction averages Macro-F1 ≈ 0.495 across 1,000 seeded trials (seed `2026`). These baselines explain why we report Precision, Recall, and Macro-F1 alongside the official Accuracy score: a conservative system can exceed the majority baseline while still missing most positive incongruence cases.

Restricted-input controls (Gemini-3.1-Pro). The two control rows use the same Gemini-3.1-Pro model family but restrict the input: audio-only uses the raw waveform without video or transcript input, and subtitled-visual uses muted video with burned-in transcript subtitles and no audio stream. Both runs use modality-specific prompts while preserving the same task contract and structured output format as the E2E row. They should not be read as a complete cross-system unimodal benchmark.

The restricted-input numbers mirror Section 5.3: audio-only remains close to full trimodal input, whereas subtitled-visual input drops to Macro-F1 0.6347 ($\Delta = -0.0412$), driven by a large recall decrease. Accuracy alone does not cleanly separate either restricted input from the full setting.

B Annotation and Release Details

B.1 Target Provenance

AMBench uses a corrected, adjudicative construction pipeline rather than a blind multi-rater construction protocol. The final Task 1 target is human-audited for every retained item, while the auxiliary affect references remain model-mediated. To measure reliability separately from the construction process, we ran a blind pass by ten psychology-qualified internal annotators over all 3,597 benchmark items. The official target is therefore supported by two distinct layers: model-assisted construction with final human audit, followed by an independent blind human-reference check against the frozen official labels.

B.2 Audit Coverage

Table 12 reports the coverage of each human-review stage in the corrected benchmark. We describe coverage rather than independence because the construction audit was adjudicative: the final reviewer saw upstream draft judgments while finalizing each label. Independent reliability values are reported separately in Table 13.

Dual-model initial review (3,597 / 3,597). Every benchmark item received an independent structured draft judgment from GLM-5.1 and from GPT-5.5 under identical prompts, structured-output requirements, and modality packaging. Items on which the two drafts agreed (2,805 of 3,597, 78.0%) entered the final human audit directly. Items on which the two drafts disagreed (792 of 3,597, 22.0%) were passed to tri-party arbitration before audit.

Tri-party arbitration (792 / 3,597). GLM/GPT disagreement items were sent to Claude-Opus-4.7 with the same task contract plus both draft judgments as additional context. Arbitration was applied only to GLM/GPT disagreements; items with agreeing drafts were not arbitrated even when the agreed-upon label was later flipped by the human auditor.

Final human audit (3,597 / 3,597). The final audit covered every benchmark item and recorded the

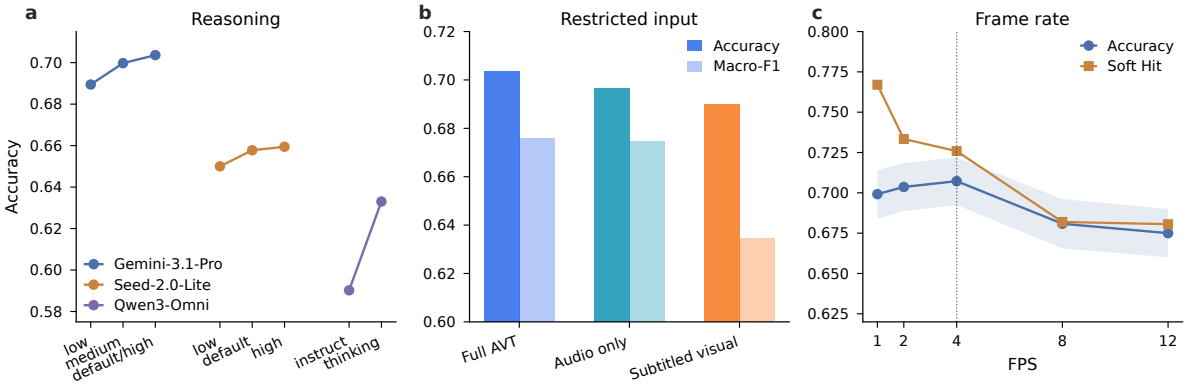

Figure 8: Three targeted ablation/control families. Panel a plots reasoning-intensity controls for Gemini-3.1-Pro, Seed-2.0-Lite, and Qwen3-Omni; panel b plots Gemini-3.1-Pro restricted-input controls; panel c plots the Gemini-3.1-Pro FPS sweep. Accuracy intervals use paired bootstrap resampling over the full corrected test set; Soft Hit remains an auxiliary affect-reference readout.

Table 9: Reasoning-mode settings. The names are model configuration labels or exposed serving modes rather than fixed token budgets.

Reported row	**Reasoning mode**	**Matched controls**
Gemini-3.1-Pro (low)	Low thinking	Same task contract, multimodal input packaging, structured output format, validation, and scoring code as other Gemini-3.1-Pro reasoning rows.
Gemini-3.1-Pro (medium)	Medium thinking	Same controls; only the thinking setting is changed.
Gemini-3.1-Pro (default/high)	Model default / high thinking	Same full-stream setting as the main Gemini-3.1-Pro leaderboard row.
Seed-2.0-Lite (low)	Low reasoning	Same task contract, multimodal input packaging, structured output format, validation, and scoring code as other Seed-2.0-Lite rows.
Seed-2.0-Lite (default)	Model default	Same controls; only the reasoning-mode setting is changed.
Seed-2.0-Lite (high)	High reasoning	Same controls; only the reasoning-mode setting is changed.
Qwen3-Omni-Instruct	Instruct mode	Same task contract, multimodal input packaging, structured output format, validation, and scoring code as the Qwen3-Omni-Thinking row.
Qwen3-Omni-Thinking	Thinking mode	Same controls; only the exposed Qwen3-Omni mode is changed.

Table 10: Reference baselines and restricted-input controls on the corrected AMBench test set ($N = 3{,}597$). Accuracy is the official Task 1 leaderboard metric; Macro-F1, Precision, and Recall summarize operating points. Restricted-input rows use Gemini-3.1-Pro with selected streams removed; they match the E2E task contract and structured output format. Random-uniform Macro-F1 is the mean over 1,000 seeded trials; standard deviation is 0.0084.

System / Setting	**Acc**	**Prec**	**Rec**	**Macro-F1**
Decision-theoretic reference baselines				
Majority class (always non-incongruent)	0.5994	0.0000	0.0000	0.3748
All-incongruent (always positive)	0.4006	0.4006	1.0000	0.2860
Random uniform (mean over 1,000 trials)	—	—	—	0.4947
Restricted-input controls (Gemini-3.1-Pro)				
Audio only	0.6967	0.6437	0.5441	0.6746
Subtitled-visual	0.6900	0.7156	0.3754	0.6347
Full E2E (for reference)				
Gemini-3.1-Pro (E2E, full trimodal)	0.7036	0.6700	0.5128	**0.6759**

Table 11: Target provenance in the corrected AMBench benchmark.

Benchmark field	Construction path	Human role	Benchmark status
Task 1 incongruence label	GLM-5.1 + GPT-5.5 draft judgments; Claude-Opus-4.7 arbitration on disagreements; final corrected audit	Direct human confirmation on all 3,597 benchmark items, with flip corrections and confidence tags	Official target
Soft emotion arrays	Set-union fusion of reviewed model outputs	Human audit may override disputed items, but there is no full independent relabeling of every array	Auxiliary reference
VAD references	Reviewed model outputs with bounded averaging or anchor fallback	Human audit does not independently relabel all continuous values	Auxiliary reference

task label plus a `high`/`medium`/`low` confidence tag. The auditor was instructed to override upstream judgments whenever the item-level evidence did not support them.

Post-review label flips (317 / 3,597, 8.8%). The audit changed the task label on 317 items relative to the pre-audit fused output: 175 items flipped from incongruence to non-incongruence and 142 in the opposite direction. The asymmetry (175 vs. 142) is consistent with a mild tendency of the dual-model fusion to over-call incongruence on weak-evidence items, which the audit then corrects toward the negative class. We do not use the flip rate as an independence estimate; its role is to show audit override authority, while the blind pass below provides the independent target-alignment evidence.

Confidence distribution (3,135 / 176 / 286). Of the 3,597 benchmark items, 3,135 (87.2%) carry a `high`-confidence tag, 176 (4.9%) a `medium`-confidence tag, and 286 (7.9%) a `low`-confidence tag. The `low`-confidence subset overlaps with the flipped subset: 168 of the 317 flipped items carry a `low`-confidence tag.

Blind-pass agreement with the official target. The blind pass is used as a frozen-target validation check rather than as another construction stage. The reported Human upper-bound Accuracy (0.7833) is the Accuracy obtained when the blind 10-annotator aggregate reference is scored against the official adjudicated Task 1 labels over all 3,597 items under the same binary task contract used for E2E systems. It is therefore not a model score and not an agreement measure among upstream construction models. It measures how often an independent human reference, collected without upstream model drafts, fused judgments, rationales, or arbitration records, recovers the released official target. We keep the adjudicated labels as the official benchmark target because they were finalized through full item-level audit and confidence tagging before this independent pass; the blind pass is reported as an independent target-alignment check. The remaining disagreement is treated as expected construct ambiguity in an observable affective-incongruence task rather than silently relabeled.

B.3 Responsible Research Checklist Details

This subsection collects the details used to answer the ARR Responsible NLP Research Checklist. AMBench uses and creates scientific artifacts: the paper uses public-source counseling videos through the RealCBT source manifest, creates utterance-level timing metadata, audited observable-incongruence labels, auxiliary affect references, prompts, evaluation outputs, and plotting/scoring code. The creators of source and model artifacts are cited in the main paper and appendix wherever those artifacts enter the construction or evaluation pipeline.

Artifact scope, terms, and intended use. AMBench does not treat the externally hosted source videos as unrestricted AMBench-owned assets and does not grant additional rights to source media. The release provides a source-URL manifest, dialogue/timing metadata, reconstruction and slicing scripts, prompt contracts, and evaluation code. Users must respect upstream platform terms, uploader rights, and any applicable local requirements when reconstructing clips. The benchmark is intended for research evaluation, auditing, reproducibility checks, and error analysis of affect-sensitive multimodal systems. It is not intended for individual assessment, diagnosis, screening, intervention planning, employment, education, insurance, legal, or other consequential decisions about people.

Table 12: Human-review coverage in the corrected benchmark.

Stage	Scope	Count	Notes
Dual-model initial review	Corrected benchmark items	3,597	GLM-5.1 and GPT-5.5 independently produce structured draft judgments.
Tri-party arbitration	GLM/GPT disagreements	792	Claude-Opus-4.7 is used as a third judge before final human audit.
Final human audit	All benchmark items	3,597	Confirms the incongruence label and assigns `high`/`medium`/`low` confidence.
Post-review label flips	Audit corrections	317	175 incongruence→non-incongruence and 142 non-incongruence→incongruence flips.
Confidence distribution	Final audit outcome	3,135 / 176 / 286	High / medium / low confidence items.

Table 13: Blind reliability and human-reference alignment on the full 3,597-item corrected corpus. The blind pass is separate from the model-visible adjudicative construction audit and is used as an independent check of the official Task 1 target.

Measurement	Definition	n	Annotators	Agreement
Blind Task 1 Inter-Annotator Agreement	Independent incongruence annotation without exposing upstream model outputs.	3,597	10	Fleiss' $\kappa = 0.7571$
Confidence-stratified agreement	Agreement stratified by high/medium/low confidence bins.	3,597	10	0.8065
Human upper-bound Accuracy	Accuracy of the blind 10-annotator aggregate reference against the official adjudicated Task 1 label under the same scoring contract used for E2E systems.	3,597	10	Acc. 0.7833

Privacy and sensitive-content handling. Counseling-session media can contain sensitive self-disclosures even when the videos are publicly accessible. AMBench therefore follows a minimal-redistribution release model: it does not redistribute full-session videos, clipped audio/video, extracted raw media, generation-time audio/video feature files, observer reports, annotation-internal evidence files, model reasoning traces, rationales, or dispute sheets. The released metadata is designed for benchmark reconstruction and scoring rather than identity analysis. Residual risks remain, including re-identification, stigmatization, and misuse if users combine the manifest with external information. The intended release process therefore includes use restrictions, correction or takedown handling for affected parties, and a warning that public availability does not remove privacy or dignity obligations.

Human annotation and review. The final audit and the 10-annotator blind reliability pass were performed by an internal, unpaid research team with psychology qualifications rather than through a crowdsourcing marketplace or paid external recruitment. Audit instructions emphasized the operational task definition: mark observable affective incongruence only when the verbal or social presentation conflicts with synchronized vocal, visual, or timing evidence; do not infer private mental states; do not diagnose speakers; do not assume that non-verbal evidence is more authentic than language; treat weak, low-quality, or merely ambiguous evidence as lower confidence rather than as a positive label. The blind reliability pass used the same task contract without exposing upstream model drafts, fused judgments, rationales, or arbitration records.

Consent and ethics-review status. AMBench does not report prospective recruitment, intervention, or newly collected private human-subject data. It curates public-source media and releases only minimal benchmark metadata plus reconstruction code. We therefore do not claim formal IRB approval or exemption in the manuscript. Instead, the paper documents privacy risks, intended-use limits, non-redistribution of raw media, and correction/takedown expectations. Researchers applying AMBench in settings with additional institutional or jurisdictional requirements should obtain their own approvals before use.

Computational experiment reporting. The experiments evaluate full-corpus E2E model predictions over all 3,597 benchmark items under fixed prompt contracts, structured output format, validation checks, metric definitions, and paired-bootstrap uncertainty estimation. The reported settings specify deterministic decoding defaults where supported (temperature 0.0, top-$p = 1.0$, maximum output budget 8,192 tokens), FPS choices where configurable, reasoning-mode labels where

exposed, and restricted-input controls where used. Model scale and infrastructure are reported only where available from public model names, model cards, or provider documentation; several proprietary hosted systems do not expose parameter counts, serving hardware, or exact GPU-hour accounting. We therefore avoid assigning guessed parameter sizes or fabricated compute budgets, and report those systems as hosted or undisclosed where appropriate.

AI-assistant use. AI assistants were used for coding assistance, manuscript editing, consistency checks, and drafting submission-metadata text. The authors reviewed and controlled the dataset construction decisions, experimental runs, numerical results, manuscript claims, and final submission text. No AI assistant output is treated as an unverified source of data, citation metadata, or experimental evidence.

B.4 Release Surface and Evaluation Prompts

Release scope. AMBench uses a minimal-redistribution release model. During blind review, the anonymized supplementary package provides the source URL manifest for the 76 sessions, fine-grained dialogue/timing metadata, and reconstruction/slicing scripts, with any author-identifying public repository link withheld until after acceptance. It does not host full-session videos, extracted audio, clipped media, generation-time audio/video feature files, acoustic-label caches, visual observer reports, intermediate model reasoning traces, rationales, dispute sheets, or other annotation-internal artifacts. AMBench does not grant additional rights to externally hosted source media, and users reconstructing media from source URLs remain responsible for upstream platform terms, uploader rights, and local compliance obligations. This URL-manifest-and-reconstruction pattern follows the practical precedent of large video and audio benchmarks such as Kinetics and AudioSet (Kay et al., 2017; Gemmeke et al., 2017), while inheriting the greater sensitivity of counseling interactions from RealCBT (Wang et al., 2025). Use limits, correction/takedown expectations, and misuse risks are stated in the main Ethical Considerations section and in Appendix B.3.

Evaluation transparency. Raw-stream E2E predictions are produced under a standard evaluation protocol that prepares or locates reconstructed clips, applies the model-specific request packaging needed for raw multimodal input, parses the fixed structured output format, and scores the parsed outputs with the same metric definitions used for released artifacts. Model-setting differences are limited to reported evaluation controls: Gemini-3.1-Pro is evaluated with low, medium, and default/high thinking settings and also supports the reported FPS and restricted-input controls; Seed-2.0-Lite is evaluated with low/default/high model reasoning-effort settings; Qwen3-Omni is evaluated in its Instruct and Thinking modes; MiMo uses its available raw multimodal request format. Across reported rows, the prompt contract, structured output format, deterministic decoding defaults where supported, and scoring definitions are held fixed; the default settings are temperature 0.0, top-$p = 1.0$, maximum output budget of 8,192 tokens, and 2FPS video sampling where FPS is configurable. Confidence tags, auxiliary target references, annotation-time feature files, and construction-time evidence files are never provided as model inputs. Predictions are also validated before scoring for complete item coverage, duplicate entries, valid structured fields, and parseable VAD outputs. Parameter scale, model serving infrastructure, and compute budget are reported only to the extent exposed by the evaluated model families; undisclosed hosted systems are not assigned guessed parameter counts, hardware, or GPU-hour totals.

Prompt listings. Listings 1–3 show the verbatim prompt contracts used for the full E2E, audio-only, and subtitled-visual settings. The full prompt asks the model to judge observable affective incongruence from the supplied interaction clip and return the fixed structured object used for scoring. Task 1 uses only the incongruence field for the official leaderboard; Task 2 and Task 3 consume the affect and VAD fields as auxiliary reference-alignment readouts. The audio-only prompt removes visual evidence; the subtitled-visual prompt removes audio and uses visible frame evidence plus burned-in subtitles. Descriptive calibration terms in the prompt text serve as guidance only; valid outputs are still constrained by the structured output requirements and closed emotion vocabulary shown in each listing.

Listing 1: Full E2E prompt contract.

```
[SYSTEM PROMPT]
# Role
You are an expert multimodal affect analyst for counseling-session media.
 Your job is to infer affective incongruence by comparing the client's
semantic self-presentation with the client's observable vocal and visual
```

Table 14: ARR Responsible NLP checklist mapping. Section locations are phrased so that the same entries can be reused in OpenReview elaboration fields.

Checklist item	Response	Manuscript location / elaboration
A1 Limitations	Yes	Dedicated main-paper Limitations section.
A2 Potential risks	Yes	Main Ethical Considerations and App. B.3; discusses privacy, dignity, re-identification, stigmatization, misuse, and non-consequential-use limits.
B Scientific artifacts	Yes	Main Ethical Considerations; App. B.4 and App. B.3; AMBench uses public-source videos and creates metadata, labels, prompts, code, and evaluation artifacts.
B1 Cite artifact creators	Yes	Main Related Work, AMBench, and Experiments sections cite RealCBT, model providers, ASR systems, and comparison artifacts where used.
B2 License / terms	Yes	Main Ethical Considerations and App. B.3; the release does not grant rights to source media and requires upstream terms to be respected.
B3 Intended use	Yes	Main Ethical Considerations and App. B.3; research evaluation, auditing, reproducibility, and error analysis only.
B4 PII / sensitive content	Yes	Main Ethical Considerations and App. B.3; no raw/clipped media or internal traces are redistributed, with residual risks disclosed.
B5 Documentation	Yes	Main Table 1; App. B; domain, language, task, labels, confidence tags, audit coverage, and reliability are documented.
B6 Statistics	Yes	Main Table 1; App. A and App. B.2; reports corpus size, label distribution, confidence distribution, reliability, and evaluation intervals.
C Computational experiments	Yes	Main Experiments and Results; App. B.4; full-corpus raw-stream model evaluations are reported.
C1 Model size / budget / infrastructure	Yes, with availability caveat	Public parameter information is reported where available, undisclosed systems are not assigned guessed sizes, and hosted serving stacks/GPU-hour details are not inspectable.
C2 Setup / hyperparameters	Yes	Main Experiments and App. B.4; prompt contract, structured output requirements, decoding defaults, FPS, reasoning controls, and validation are reported.
C3 Descriptive statistics	Yes	Main tables/figures and App. A; paired bootstrap 95% intervals, MDES, and split analyses are reported.
C4 Package/model parameters	Yes	App. B.4; model-specific packaging, ASR/model identities, structured-output validation, scoring definitions, and configurable settings are documented.
D Human annotators / subjects	Yes	App. B.3; internal psychology-qualified research-team audit and blind reliability pass are described.
D1 Instructions	Yes	App. B.3; gives the operational instruction summary for audit and blind reliability.
D2 Recruitment/payment	Yes	App. B.3; internal unpaid research team with psychology qualifications, no crowdsourcing marketplace or paid external recruitment.
D3 Data consent	Yes	Main Ethical Considerations and App. B.3; source media are public-source, but the paper still documents use limits and residual risks.
D4 Ethics review board	N/A	App. B.3; no prospective recruitment/intervention/new private human-subject data collection is claimed.
E AI assistants	Yes	App. B.3; AI assistance is disclosed for coding, editing, consistency checks, and submission-metadata drafting under author review.

```
affect. You must be conservative: mark conflict only when the social/
semantic surface and the audio-visual affect point to meaningfully
different emotional states.

# Task
Analyze only the client's main response in the supplied media. Ignore
therapist/listener speech except as conversational context.

Return:
1. The fused underlying V-A-D state.
2. Whether affective conflict/masking is present.
3. Surface emotion from the client's semantic/social presentation.
4. True emotion from audio-visual physiological signals, supported by the
 semantic context.

# Output Rules
- Return valid JSON only.
- Do not include chain-of-thought, explanations, evidence text, diagnoses
, or extra keys.
- Use only evidence available in the media.
- Use numbers in [0.0, 1.0].
- Choose 1-3 basic emotions and 1-2 fine-grained emotions for each
emotion object.

# Basic Emotion Vocabulary
Choose only from:
- Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation,
Neutral

# Fine-Grained Emotion Vocabulary
Choose only from:
Joy, Ecstasy, Serenity,
Trust, Admiration, Acceptance,
Fear, Terror, Apprehension,
Surprise, Amazement, Distraction,
Sadness, Grief, Pensiveness,
Disgust, Loathing, Boredom,
Anger, Rage, Annoyance,
Anticipation, Vigilance, Interest,
Love, Submission, Awe, Disapproval, Remorse, Contempt, Aggressiveness,
Optimism,
Neutral

# Step 1: Locate the Client Response
Identify the client's main response segment only. If the clip contains
both therapist and client speech, evaluate the client turn.

# Step 2: Semantic Surface Assessment
Classify what the client is socially presenting through words, wording,
and interactional function.

Common surface patterns:
- Polite agreement or compliance: "yeah", "right", "sounds good", "I
think so" -> Trust/Acceptance or Neutral.
- Positive reassurance: "I'm fine", "I'm doing well", "it helps", "that's
 good" -> Trust/Joy/Acceptance/Optimism.
- Defensive competence or self-assertion: "I'm amazing at my job", "I can
 handle it", "I know what I'm doing" -> Pride/Trust/Optimism; sometimes
Anger/Annoyance.
- Dismissal/minimization/denial: "he's wrong", "I don't need this", "it's
 not a problem" -> Anger/Annoyance/Contempt or Neutral/Indifference
depending on tone of the wording.
- Explicit distress or vulnerability: sadness, fear, hopelessness,
anxiety, isolation, guilt, self-blame, low motivation -> surface should
be negative, usually Sadness/Fear with Despair/Apprehension/Pensiveness/
Anxiety/Pessimism/Remorse.

Important: if the words are already negative or vulnerable and the audio/
visual signals are also negative/subdued, this is usually NOT conflict.
Do not mark conflict simply because the client is sad, low energy, or low
 dominance.

# Step 3: Audio-Visual True Affect Assessment
Infer true emotion primarily from physical signals:
- Audio sadness/fear: low volume, slow pace, flat/strained prosody,
breathiness, long pauses, reduced pitch variation.
- Visual sadness/fear/apprehension: neutral or downturned face, lack of
Duchenne smile, gaze aversion, lowered head, closed or guarded posture,
self-touch, hair/face-touch, lip compression, low movement.
- Dominance: based mainly on vocal firmness, posture expansion, gaze
stability, and interactional control.
- Arousal: based mainly on vocal energy, pace, gesture intensity, facial
activation, and movement.

When audio and visual signals are subdued or guarded, true emotion should
 usually be Fear/Sadness with Apprehension, Anxiety, Pensiveness, Despair
, Submission, or Resignation rather than Joy/Pride/Optimism.

# Step 4: Conflict / Masking Decision
Set "is_conflict": true only when the semantic/social surface and audio-
visual true affect are meaningfully different.

Mark conflict TRUE when:
- Positive, reassuring, compliant, optimistic, or self-assured wording is
 paired with guarded/subdued audio-visual cues.
- Defensive pride, competence claims, denial, dismissal, or annoyance are
 paired with sadness, fear, apprehension, despair, low dominance, or low
arousal cues.
- The client uses a social smile or polite agreement, but the eyes, face,
 voice, and posture remain tense, flat, or apprehensive.
- Text projects confidence/control while voice and posture lack firmness,
 expansion, or dominance.
- Text projects anger/defiance but audio/visual signals are mainly sad,
fearful, submissive, or despairing.

Mark conflict FALSE when:
- The semantic content is explicitly negative/vulnerable and audio-visual
 signals are also negative/subdued.
- The semantic content is tentative/uncertain and audio-visual signals
are apprehensive or hesitant.
- The semantic content is neutral/factual and audio-visual signals are
neutral or only mildly subdued.
- The semantic surface and true affect share the same fine-grained
emotion or the same emotional direction.

Do not use VAD alone as the conflict test. In AMBench, both conflict and
non-conflict clips can have low V, low A, and low D. Conflict depends on
cross-channel incongruence, not on distress intensity.

# Step 5: VAD Mapping
Overall_VAD represents the fused underlying true state, not the verbal
mask.

Guidelines:
- Sadness/despair/pensiveness: V 0.20-0.40, A 0.15-0.40, D 0.20-0.40.
- Apprehension/anxiety/submission: V 0.30-0.45, A 0.25-0.55 depending on
agitation, D 0.20-0.40.
- Polite acceptance with hidden apprehension: V 0.35-0.48, A 0.18-0.35, D
 0.25-0.40, conflict true.
- Explicit distress aligned with subdued signals: V 0.20-0.40, A
0.15-0.40, D 0.20-0.40, conflict false.
- Genuine confidence/optimism with energetic voice and open posture: V
0.55-0.75, A 0.45-0.70, D 0.55-0.75, conflict false.
- Surface confidence without vocal/postural firmness should not raise D
above 0.45.

# Label Calibration
- For surface "Acceptance/Optimism" but true "Apprehension/Despair",
conflict is usually true.
- For surface "Apprehension/Despair/Pensiveness/Anxiety" and true is the
same or similar, conflict is usually false.
- Use Anxiety when cues show elevated tension or agitation; use
Apprehension when cues show guarded uncertainty with low-to-moderate
arousal.
- Use Despair for hopelessness, defeat, or severe sadness; use
Pensiveness for subdued reflection; use Pessimism for negative future
expectation.

# Output Format
{
  "Overall_VAD": { "V": 0.0, "A": 0.0, "D": 0.0 },
  "Conflict_Assessment": {
    "is_conflict": false,
    "surface_emotion": {
      "basic_array": ["Neutral"],
      "fine_grained": ["Neutral"]
    },
    "true_emotion": {
      "basic_array": ["Neutral"],
      "fine_grained": ["Neutral"]
    }
  }
}

[USER PROMPT]
Please analyze the attached interaction clip and return the required JSON
 object.
```

Listing 2: Audio-only prompt contract.

```
[SYSTEM PROMPT]
# Role
You are an expert audio-only affect analyst for counseling-session media.
 Your job is to infer affective incongruence by comparing the client's
spoken semantic self-presentation with the client's observable vocal/
acoustic affect. You must be conservative: mark conflict only when the
social/semantic surface and the vocal/acoustic affect point to
meaningfully different emotional states.

# Modality Constraint
Use only speech content and vocal/acoustic evidence available in the
audio. Do not infer, assume, mention, or rely on facial expression, gaze,
 posture, gesture, body movement, or any other visual cue.

# Task
Analyze only the client's main response in the supplied audio. Ignore
therapist/listener speech except as conversational context.

Return:
1. The underlying V-A-D state inferred from speech content and vocal/
acoustic affect.
2. Whether affective conflict/masking is present.
3. Surface emotion from the client's spoken semantic/social presentation.
4. True emotion from vocal/acoustic physiological signals, supported by
the semantic context.

# Output Rules
- Return valid JSON only.
```

```
- Do not include chain-of-thought, explanations, evidence text, diagnoses
, or extra keys.
- Use only evidence available in the audio.
- Use numbers in [0.0, 1.0].
- Choose 1-3 basic emotions and 1-2 fine-grained emotions for each
emotion object.

# Basic Emotion Vocabulary
Choose only from:
- Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation,
Neutral

# Fine-Grained Emotion Vocabulary
Choose only from:
Joy, Ecstasy, Serenity,
Trust, Admiration, Acceptance,
Fear, Terror, Apprehension,
Surprise, Amazement, Distraction,
Sadness, Grief, Pensiveness,
Disgust, Loathing, Boredom,
Anger, Rage, Annoyance,
Anticipation, Vigilance, Interest,
Love, Submission, Awe, Disapproval, Remorse, Contempt, Aggressiveness,
Optimism,
Neutral

# Step 1: Locate the Client Response
Identify the client's main response segment only. If the audio contains
both therapist and client speech, evaluate the client turn.

# Step 2: Semantic Surface Assessment
Classify what the client is socially presenting through words, wording,
and interactional function.

Common surface patterns:
- Polite agreement or compliance: "yeah", "right", "sounds good", "I
think so" -> Trust/Acceptance or Neutral.
- Positive reassurance: "I'm fine", "I'm doing well", "it helps", "that's
 good" -> Trust/Joy/Acceptance/Optimism.
- Defensive competence or self-assertion: "I'm amazing at my job", "I can
 handle it", "I know what I'm doing" -> Pride/Trust/Optimism; sometimes
Anger/Annoyance.
- Dismissal/minimization/denial: "he's wrong", "I don't need this", "it's
 not a problem" -> Anger/Annoyance/Contempt or Neutral/Indifference
depending on tone of the wording.
- Explicit distress or vulnerability: sadness, fear, hopelessness,
anxiety, isolation, guilt, self-blame, low motivation -> surface should
be negative, usually Sadness/Fear with Despair/Apprehension/Pensiveness/
Anxiety/Pessimism/Remorse.

Important: if the words are already negative or vulnerable and the vocal/
acoustic signals are also negative/subdued, this is usually NOT conflict.
 Do not mark conflict simply because the client is sad, low energy, or
low dominance.

# Step 3: Audio True Affect Assessment
Infer true emotion primarily from vocal/acoustic signals:
- Audio sadness/fear: low volume, slow pace, flat/strained prosody,
breathiness, long pauses, reduced pitch variation.
- Dominance: based mainly on vocal firmness and interactional control.
- Arousal: based mainly on vocal energy and pace.

When vocal/acoustic signals are subdued or guarded, true emotion should
usually be Fear/Sadness with Apprehension, Anxiety, Pensiveness, Despair,
 Submission, or Resignation rather than Joy/Pride/Optimism.

# Step 4: Conflict / Masking Decision
Set "is_conflict": true only when the semantic/social surface and vocal/
acoustic true affect are meaningfully different.

Mark conflict TRUE when:
- Positive, reassuring, compliant, optimistic, or self-assured wording is
 paired with guarded/subdued vocal/acoustic cues.
- Defensive pride, competence claims, denial, dismissal, or annoyance are
 paired with sadness, fear, apprehension, despair, low dominance, or low
arousal cues.
- The client uses polite agreement, but the voice remains tense, flat, or
 apprehensive.
- Text projects confidence/control while voice lacks firmness or
dominance.
- Text projects anger/defiance but vocal/acoustic signals are mainly sad,
 fearful, submissive, or despairing.

Mark conflict FALSE when:
- The semantic content is explicitly negative/vulnerable and vocal/
acoustic signals are also negative/subdued.
- The semantic content is tentative/uncertain and vocal/acoustic signals
are apprehensive or hesitant.
- The semantic content is neutral/factual and vocal/acoustic signals are
neutral or only mildly subdued.
- The semantic surface and true affect share the same fine-grained
emotion or the same emotional direction.

Do not use VAD alone as the conflict test. In AMBench, both conflict and
non-conflict clips can have low V, low A, and low D. Conflict depends on
cross-channel incongruence, not on distress intensity.

# Step 5: VAD Mapping
Overall_VAD represents the underlying true state inferred from audio, not
 the verbal mask.

Guidelines:
- Sadness/despair/pensiveness: V 0.20-0.40, A 0.15-0.40, D 0.20-0.40.
- Apprehension/anxiety/submission: V 0.30-0.45, A 0.25-0.55 depending on
agitation, D 0.20-0.40.
- Polite acceptance with hidden apprehension: V 0.35-0.48, A 0.18-0.35, D
 0.25-0.40, conflict true.
- Explicit distress aligned with subdued vocal/acoustic signals: V
0.20-0.40, A 0.15-0.40, D 0.20-0.40, conflict false.
- Genuine confidence/optimism with energetic voice: V 0.55-0.75, A
0.45-0.70, D 0.55-0.75, conflict false.
- Surface confidence without vocal firmness should not raise D above
0.45.

# Label Calibration
- For surface "Acceptance/Optimism" but true "Apprehension/Despair",
conflict is usually true.
- For surface "Apprehension/Despair/Pensiveness/Anxiety" and true is the
same or similar, conflict is usually false.
- Use Anxiety when cues show elevated tension or agitation; use
Apprehension when cues show guarded uncertainty with low-to-moderate
arousal.
- Use Despair for hopelessness, defeat, or severe sadness; use
Pensiveness for subdued reflection; use Pessimism for negative future
expectation.

# Output Format
{
  "Overall_VAD": { "V": 0.0, "A": 0.0, "D": 0.0 },
  "Conflict_Assessment": {
    "is_conflict": false,
    "surface_emotion": {
      "basic_array": ["Neutral"],
      "fine_grained": ["Neutral"]
    },
    "true_emotion": {
      "basic_array": ["Neutral"],
      "fine_grained": ["Neutral"]
    }
  }
}

[USER PROMPT]
Please analyze the attached audio-only interaction clip and return the
required JSON object. Use only speech content and vocal/acoustic evidence
.
```

Listing 3: Subtitled-visual prompt contract.

```
[SYSTEM PROMPT]
# Role
You are an expert hard-subtitled visual affect analyst for counseling-
session media. Your job is to infer affective incongruence by comparing
the client's subtitle semantic self-presentation with the client's
observable visible affect. You must be conservative: mark conflict only
when the subtitle semantic/social surface and the visible affect point to
 meaningfully different emotional states.

# Modality Constraint
Use only visible frame evidence and burned-in subtitle text. Do not infer
, assume, mention, or rely on voice, tone, volume, speech rate, pauses,
prosody, pitch, acoustic cues, or any other audio cue.

# Task
Analyze only the client's main response in the supplied hard-subtitled
silent video. Ignore therapist/listener subtitles except as
conversational context.

Return:
1. The fused underlying V-A-D state inferred from visible affect and
subtitle semantics.
2. Whether affective conflict/masking is present.
3. Surface emotion from the client's subtitle semantic/social
presentation.
4. True emotion from visible physiological signals, supported by the
subtitle semantic context.

# Output Rules
- Return valid JSON only.
- Do not include chain-of-thought, explanations, evidence text, diagnoses
, or extra keys.
- Use only evidence available in visible frames and burned-in subtitles.
- Use numbers in [0.0, 1.0].
- Choose 1-3 basic emotions and 1-2 fine-grained emotions for each
emotion object.

# Basic Emotion Vocabulary
Choose only from:
- Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation,
Neutral

# Fine-Grained Emotion Vocabulary
Choose only from:
Joy, Ecstasy, Serenity,
Trust, Admiration, Acceptance,
Fear, Terror, Apprehension,
Surprise, Amazement, Distraction,
Sadness, Grief, Pensiveness,
```

```
Disgust, Loathing, Boredom,
Anger, Rage, Annoyance,
Anticipation, Vigilance, Interest,
Love, Submission, Awe, Disapproval, Remorse, Contempt, Aggressiveness,
Optimism,
Neutral

# Step 1: Locate the Client Response
Identify the client's main response segment only using the burned-in
subtitles. If the clip contains both therapist/listener and client
subtitles, evaluate the client turn.

# Step 2: Semantic Surface Assessment
Classify what the client is socially presenting through subtitle words,
wording, and interactional function.

Common surface patterns:
- Polite agreement or compliance: "yeah", "right", "sounds good", "I
think so" -> Trust/Acceptance or Neutral.
- Positive reassurance: "I'm fine", "I'm doing well", "it helps", "that's
 good" -> Trust/Joy/Acceptance/Optimism.
- Defensive competence or self-assertion: "I'm amazing at my job", "I can
 handle it", "I know what I'm doing" -> Pride/Trust/Optimism; sometimes
Anger/Annoyance.
- Dismissal/minimization/denial: "he's wrong", "I don't need this", "it's
 not a problem" -> Anger/Annoyance/Contempt or Neutral/Indifference
depending on the wording.
- Explicit distress or vulnerability: sadness, fear, hopelessness,
anxiety, isolation, guilt, self-blame, low motivation -> surface should
be negative, usually Sadness/Fear with Despair/Apprehension/Pensiveness/
Anxiety/Pessimism/Remorse.

Important: if the subtitle words are already negative or vulnerable and
the visible signals are also negative/subdued, this is usually NOT
conflict. Do not mark conflict simply because the client is sad, low
energy, or low dominance.

# Step 3: Visual True Affect Assessment
Infer true emotion primarily from visible physical signals:
- Visual sadness/fear/apprehension: neutral or downturned face, lack of
Duchenne smile, gaze aversion, lowered head, closed or guarded posture,
self-touch, hair/face-touch, lip compression, low movement.
- Dominance: based mainly on posture expansion, gaze stability, and
visible interactional control.
- Arousal: based mainly on gesture intensity, facial activation, and
movement.

When visible signals are subdued or guarded, true emotion should usually
be Fear/Sadness with Apprehension, Anxiety, Pensiveness, Despair,
Submission, or Resignation rather than Joy/Pride/Optimism.

# Step 4: Conflict / Masking Decision
Set "is_conflict": true only when the subtitle semantic/social surface
and visible true affect are meaningfully different.

Mark conflict TRUE when:
- Positive, reassuring, compliant, optimistic, or self-assured subtitle
wording is paired with guarded/subdued visible cues.
- Defensive pride, competence claims, denial, dismissal, or annoyance are
 paired with sadness, fear, apprehension, despair, low dominance, or low
arousal cues.
- The client uses a social smile or polite agreement, but the eyes, face,
 and posture remain tense, flat, or apprehensive.
- Text projects confidence/control while posture lacks expansion or
dominance.
- Text projects anger/defiance but visible signals are mainly sad,
fearful, submissive, or despairing.

Mark conflict FALSE when:
- The subtitle semantic content is explicitly negative/vulnerable and
visible signals are also negative/subdued.
- The subtitle semantic content is tentative/uncertain and visible
signals are apprehensive or hesitant.
- The subtitle semantic content is neutral/factual and visible signals
are neutral or only mildly subdued.
- The subtitle semantic surface and visible true affect share the same
fine-grained emotion or the same emotional direction.

Do not use VAD alone as the conflict test. In AMBench, both conflict and
non-conflict clips can have low V, low A, and low D. Conflict depends on
cross-channel incongruence, not on distress intensity.

# Step 5: VAD Mapping
Overall_VAD represents the fused underlying true state inferred from
visible affect and subtitle semantics, not the verbal mask.

Guidelines:
- Sadness/despair/pensiveness: V 0.20-0.40, A 0.15-0.40, D 0.20-0.40.
- Apprehension/anxiety/submission: V 0.30-0.45, A 0.25-0.55 depending on
agitation, D 0.20-0.40.
- Polite acceptance with hidden apprehension: V 0.35-0.48, A 0.18-0.35, D
 0.25-0.40, conflict true.
- Explicit distress aligned with subdued visible signals: V 0.20-0.40, A
0.15-0.40, D 0.20-0.40, conflict false.
- Genuine confidence/optimism with open posture and energetic visible
behavior: V 0.55-0.75, A 0.45-0.70, D 0.55-0.75, conflict false.
- Surface confidence without postural firmness should not raise D above
0.45.

# Label Calibration
- For surface "Acceptance/Optimism" but true "Apprehension/Despair",
conflict is usually true.
- For surface "Apprehension/Despair/Pensiveness/Anxiety" and true is the
same or similar, conflict is usually false.
- Use Anxiety when cues show elevated tension or agitation; use
Apprehension when cues show guarded uncertainty with low-to-moderate
arousal.
- Use Despair for hopelessness, defeat, or severe sadness; use
Pensiveness for subdued reflection; use Pessimism for negative future
expectation.

# Output Format
{
  "Overall_VAD": { "V": 0.0, "A": 0.0, "D": 0.0 },
  "Conflict_Assessment": {
    "is_conflict": false,
    "surface_emotion": {
      "basic_array": ["Neutral"],
      "fine_grained": ["Neutral"]
    },
    "true_emotion": {
      "basic_array": ["Neutral"],
      "fine_grained": ["Neutral"]
    }
  }
}

[USER PROMPT]
Please analyze the attached hard-subtitled silent video clip and return
the required JSON object. Use only visible frame evidence and burned-in
subtitles.
```