CONTACT: A HUMAN-GROUNDED BENCHMARK AND SURPRISAL-BASED PREDICTIVE SCORER FOR CONVERSATIONAL NATURALNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spoken dialogue evaluation has advanced from response correctness to turn-taking, emotional responsiveness, and conversational quality. Yet a central challenge remains: determining whether automatic scores reflect the naturalness people experience across both targeted interaction failures and spontaneous conversation. We introduce CONTACT, a human-grounded benchmark and a predictive scorer to address this challenge. The benchmark combines live human–human conversations with elicited turn-taking and affective disruptions, and free-form human–AI dialogues. Human ratings anchor the assessment of turn-taking, affective, and overall naturalness across the benchmark's subsets, connecting diagnostic control with naturalistic evaluation. Alongside this resource, we propose a surprisal-based scorer trained on approximately 1,600 hours of natural dyadic speech to anticipate future turn-taking and affective states. Surprisal of the observed speech given a previous context yields dimension-specific scores and an overall composite without requiring a matched reference conversation. Results suggest that the proposed scorer provides a higher correlation with the perceived naturalness provided by the benchmark than current full-duplex evaluation systems. Together, the benchmark and scorer connect targeted diagnosis, human-perceived interaction quality, and natural-dialogue predictive learning, providing a reusable basis for evaluating conversational naturalness.

1 INTRODUCTION

Audio large language models (LLMs) are capable of having real-time spoken interactions with humans. However, this interaction is still not as natural as human-human (H-H) interactions (Lin et al., 2025a). New tools are needed to help developers identify where the sources of turn-taking and affective disfluencies are (Zhao et al., 2026). A fluent spoken dialogue system must do more than produce plausible responses: an otherwise relevant response can disrupt a conversation by entering before the speaker has finished, or expressing an emotional reaction that does not fit the exchange.

In this context, we define interactional *naturalness* as the degree to which a spoken exchange is perceived by the speakers or human listeners as resembling everyday human conversation in its turn-taking and affective responses. Lower naturalness reflects disruptive timing or contextually inappropriate affective expression, whereas higher naturalness reflects contextually appropriate turn-taking and affective responses (Sections 3.1 and 3.2 describe how the evaluators rated naturalness in this study). Our ultimate goal is to provide a benchmark and its associated automatic measurements on turn-taking and affective appropriateness, which would inform developers and scientists about designing new conversational agents that provide more natural and seamless conversations, from a human perception perspective.

There is a growing interest in benchmarking and evaluating spoken-dialogue naturalness beyond the analysis of response content (Arora et al., 2025; Lin et al., 2025a; 2026a; Zhao et al., 2026; Wang et al., 2026; Xu et al., 2026). However, these approaches do not jointly evaluate turn-taking and affective responses, despite evidence that these dimensions are interrelated (Koudenburg et al., 2011). Moreover, the reviewed benchmarks lack recordings of H-H conversations featuring contextually inappropriate affective responses or unnatural turn-taking, assessed by both participants and

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

external evaluators, to validate algorithms and metrics under these conditions. Building on previous advances and these missing pieces in the state of the art, we propose the following claims:

- *an automatic naturalness metric should respond to targeted, human-perceived interaction naturalness and remain informative in spontaneous H–AI conversation.*
- *live conversations between humans with a varied range of perceived naturalness in terms of timing and affective appropriateness are needed to validate naturalness algorithms and metrics*
- *naturalistic H–AI interaction recordings with a wide range of naturalness annotations complement and reinforce such validation.*

To address these claims, we introduce CONTACT[1], a human-grounded benchmark with perceived human naturalness ratings that comprises two main parts: a set of H-H conversations with a variety of turn-taking naturalness and affective perceived appropriateness, and a set of H-AI conversations involving several speech AI agents. In the H-H conversations set, speaker pairs engage in semi-structured conversations, comprising two subsets: the *awkward interaction* subset, with elicited disruptions such as premature entry, competitive floor conflict, and blocked invited responses; the *affective mismatch* subset, featuring elicited insufficient engagement, incompatible emotional valence, and disproportionate emotional intensity. The human–AI conversation set complements this diagnostic setting with topic-guided conversations between human participants and speech agents that have a wide range of perceived naturalness annotated.

Alongside the CONTACT benchmark, and to show its usefulness, we introduce a predictive naturalness scorer. Our scorer predicts future joint voice activity and emotion using a shared causal speech backbone and two Transformer branches. Appendix A depicts an overview of the benchmark.

Consequently, our contributions are threefold:

- **A human-grounded interactional naturalness benchmark,** comprising live H-H and H-AI conversations, with perceived naturalness ratings.
- **A naturalness predictive scorer,** for turn-taking and affective appropriateness assessment.
- **Unified evaluation of naturalness metrics.** CONTACT supports metric comparisons based on sensitivity to elicited disruptions and agreement with graded human ratings.

2 Related Work

Table 1 contains a comparison of existing benchmarks and resources aimed at providing turn-taking and affective appropriateness evaluation tools. **Timing** denotes an explicit evaluation of perceived turn-taking or temporal coordination; **Affect.** denotes perceived affective appropriateness rather than emotion recognition alone. **Elicited H–H** refers to live human-produced interactional disruptions rather than post-hoc editing or resynthesis; **Live H–AI** indicates real humans interacting with a spoken dialogue system. **Human eval.** denotes subjective human ratings ground truth rather than event annotations or behavioral proxies.

Benchmark / resource	Timing	Affect.	Multi-turn	Elicited H–H	Live H–AI	Human eval.
Talking Turns (Arora et al., 2025)	✓	✗	✓	✗	✓	✗
Full-Duplex-Bench (Lin et al., 2025a)	✓	✗	✓✗	✗	✗	✗
MTalk-Bench (Du et al., 2025)	✓✗	✓✗	✓	✗	✗	✓
Full-Duplex-Bench-v2 (Lin et al., 2026a)	✓	✗	✓	✗	✗	✓
MULTI-Bench (Deng et al., 2025)	✗	✓	✓	✗	✗	✓
ParaS2SBench (Yang et al., 2026)	✗	✓	✗	✗	✗	✓
HumDial (Zhao et al., 2026)	✓	✓	✓	✓✗	✗	✓
ConvTTS / FDX-Conv (Xu et al., 2026)	✗	✗	✓	✗	✓	✓
UniSRM-Bench (Tasks 3–4) (Wang et al., 2026)	✗	✓	✓	✗	✗	✓✗
DyadEE / TRACE (Ugandhar et al., 2026)	✗	✓	✓	✗	✗	✓
TurnBench (Jiang et al., 2026)	✓	✗	✓	✗	✗	✗
CONTACT (Ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of spoken-interaction benchmarks and evaluation resources. ✓: covered; ✓✗: partially or relatedly covered; ✗: not covered under these definitions.

[1]Conversational data set: `https://github.com/review-anonymous-git/contact-samples` and code: `https://github.com/review-anonymous-git/contact-code`

Turn-Taking Naturalness. Conversational turn-taking coordinates speakers' participation through shifts, holds, pauses, overlaps, interruptions, and backchannels (Skantze, 2021). Listening tests show that perceived timing naturalness depends on conversational context (Roddy & Harte, 2020).

Some automatic benchmarks assess specific behaviors. Full-Duplex-Bench (Lin et al., 2025a) and subsequent extensions (Lin et al., 2026b) evaluate pause handling, backchanneling, turn transitions, interruptions, and overlapping and non-addressed speech. The recent HumDial Challenge adapts this protocol to evaluate full-duplex interaction (Zhao et al., 2026), while Talking Turns assesses event decisions using a supervised predictor and event-specific thresholds (Arora et al., 2025). Interpreting these diagnostics as naturalness measures requires perceptual validation.

Predictive models provide a foundation for automatic scoring. VAP learns to predict joint future voice activity (Ekstedt & Skantze, 2022). Related approaches combine acoustic and linguistic features (Lin et al., 2025b; Wang et al., 2024), while DualTurn uses dual-channel generative speech pretraining followed by turn-taking fine-tuning (Rajaa, 2026). Leveraging these models, in this study we use a causal predictor trained on natural conversations to test whether lower likelihoods of evaluated conversations from CONTACT correspond to lower human ratings of turn-taking naturalness.

Affective Appropriateness. Affective perception depends on the contextual interpretation of semantic and prosodic cues (Ben-David et al., 2016). Resources such as IEMOCAP provide categorical and dimensional annotations of expressed emotion (Busso et al., 2008), without directly assessing its conversational appropriateness. Recent benchmarks examine how spoken models respond to paralinguistic information. SD-Eval covers emotion, accent, age, and background sound (Ao et al., 2024), while ParaS2SBench assesses response content and speaking style under paralinguistic variation (Yang et al., 2026). EchoMind extends evaluation to vocal-cue perception, emotional reasoning, and empathetic generation (Zhou et al., 2026).

Other studies address emotional coherence and conflicting cues. EMO-Reasoning evaluates multi-turn emotional coherence using TTS-generated dialogue (Liu et al., 2025), whereas HumDial assesses emotional understanding and empathetic generation using human-recorded conversations (Zhao et al., 2026). EMIS introduces conflicting semantic and vocal emotions in synthetic speech, revealing reliance on textual semantics (Corrêa et al., 2026). Beyond these benchmarks, UniSRM combines a benchmark spanning utterance quality and contextual coherence with a unified speech reward model (Wang et al., 2026).

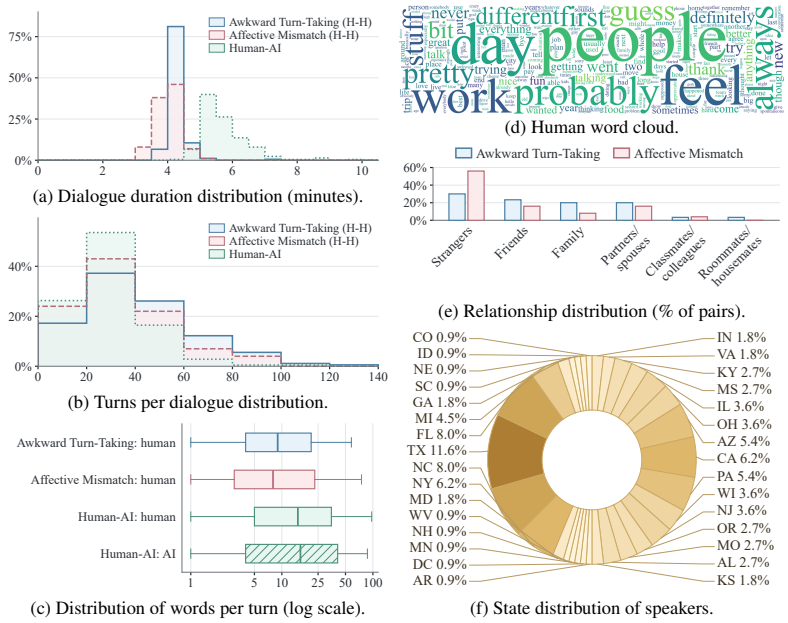
Compared to these studies, our focus is conversational affective appropriateness: whether a speaker's emotional response suits the preceding interpersonal context. We test whether surprisal-based affective readouts, combined with temporal coordination cues, correlate with human-rated affective appropriateness.

3 The CONTACT Human-Grounded Conversational Dataset

CONTACT supports the evaluation of perceived interactional naturalness through three subsets: Awkward Turn-Taking and Affective Mismatch, comprising H-H conversations and Open-Ended H–AI Conversations. Participants were recruited through the Prolific platform and were required to be native English speakers. Participants joined an online recording session, using their personal computer with a functioning microphone, webcam and a stable internet connection. They could see each other during the conversations in all H-H interactions. Participants were asked to choose a quiet environment without other speakers, were recommended to wear headphones, and were always recorded in separate channels. We also assessed recording-path delay using an OBS-based beep test (Appendix B.3.3). More details on recruitment and recording procedures are provided in Appendix B and on the Ethics statement.

3.1 The Awkward turn-taking subset

We collected dyadic H-H conversations by providing the speakers with a conversational context and condition-specific behavioral instructions, rather than a fixed script, with the goal of allowing participants to determine the development of the conversation. Each recording session began with a baseline recording in which both participants were instructed to converse normally, without deliberately modifying their turn-taking behavior. Then, we targeted five forms of turn-taking disruption: *Premature Entry*, where a speaker begins responding before their partner finishes; *Blocked Invited Response*, where a speaker appears to offer the floor but takes it back after the partner begins



(a) Dialogue duration distribution (minutes).

(b) Turns per dialogue distribution.

(c) Distribution of words per turn (log scale).

(d) Human word cloud.

(e) Relationship distribution (% of pairs).

(f) State distribution of speakers.

Figure 1: Visualization of static analysis of CONTACT.

responding; *Competitive Floor Conflict*, where both speakers continue substantive speech during overlap; *Fragmented Turn Holding*, where repeated turn-internal pauses make it difficult for the partner to determine when to respond; and *Unmarked Turn Abandonment*, where a speaker stops without clearly completing or releasing the turn.

In recordings with intentional turn-taking disruptions, the target behavior was privately explained and briefly practiced with one member of the dyad so that the conversational partner was unaware of the type of elicited behavior. The partner was instructed to continue the conversation naturally, and both participants were asked to exhibit behavior in a conversationally plausible manner. Topics were selected to encourage everyday dialogue-based conversations e.g., planning, advice, troubleshooting, or narrating an experience. Table 5 in Appendix C includes the topic families used.

3.2 The Affective Mismatch subset

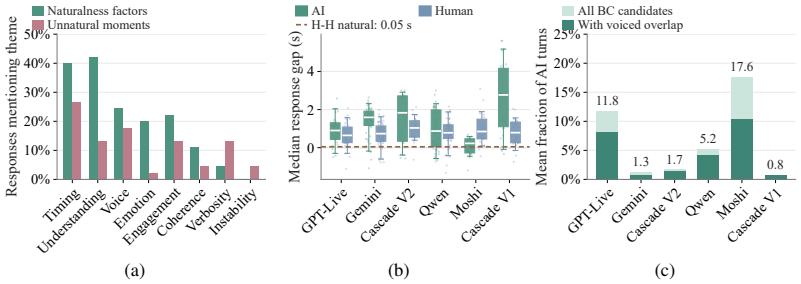
The affective subset uses the same general semi-structured H–H paradigm described in Subsection 3.1 but participants were asked to manipulate the relationship between the affective expression of speech and the context of the conversation. The sessions included a baseline recording, in which both participants responded without special affective instructions and, then, three manipulated conditions: *Under-Response*: one of the speakers preserved a broadly appropriate emotional interaction while reducing expected emotional engagement; *Disproportionate Overreaction*: one participant preserved an appropriate emotional direction while increasing its intensity beyond what the context would ordinarily warrant; and *Valence Mismatch*: one of the speakers preserved semantic relevance while making vocal affect incompatible with the conversational context, according to the speaker's own criterion (opposite emotional direction). Topics were selected to establish conversational contexts in which particular affective orientations were conversationally plausible (e.g., excitement, disappointment, frustration, anxiety, or surprise), as included in Table 6 in Appendix C. We did not attempt to infer or control participants' baseline latent affective state prior to each conversation. Participants were instructed to manipulate vocal affect while remaining semantically relevant and cooperative, avoiding deliberately contradictory or inappropriate lexical content.

3.3 The Open-Ended Human–AI Conversations subset

Each participant completed 5 spoken conversations with AI speech assistants. Topics were proposed by a research supervisor and spanned different emotional and conversational contexts, including everyday routines, enthusiasm, uncertainty, and frustration, to encourage variation in affect and timing. Appendix C includes a detailed list of topics and instructions. To avoid biases towards any model, the topics were always the same for each session but were randomly assigned to the models. The conversations were not held in the same order of models for different sessions, to avoid fatigue effect. Participants were asked to converse naturally, without preparing responses

Table 2: Participant demographics and conversation statistics by subset.

Subset	#Particip.	Sex (M/F)	#Convers.	#Turns	Mean age (std)	Hours
Awkward Turn-Taking (H–H)	60	21/39	180	7,378	42.2 (12.9)	12.8
Affective Mismatch (H–H)	50	24/26	100	3,428	41.3 (11.6)	6.7
Open-Ended H–AI Conversations	45	22/23	213	6,220	42.0 (13.0)	20.3



(a) (b) (c)

Figure 2: **Perceived interaction characteristics of human–AI conversations.** (a) Themes in participant feedback influencing naturalness. (b) Median response gaps of AI agents and humans; dashed line: H–H natural baseline (0.05 s). (c) Backchannel (BC) candidate rates (definitions in Appendix D.2). In (b) and (c), systems are ordered by combined turn-taking MOS (left: most natural).

or following a script. No fixed number of turns, explicit turn-taking behaviors or emotions were prescribed, allowing pauses, interruptions, overlaps, and backchannels to emerge spontaneously. The details of the recording protocol are included in Appendix B.

Spoken dialogue systems employed. We evaluated four end-to-end spoken dialogue systems and two cascaded Vapi pipelines that share GPT-4.1 for response generation but differ in speech components: Cascade V1 uses GPT-4o Transcribe and GPT-4o Mini TTS, whereas Cascade V2 uses xAI speech recognition and synthesis (Table 3*(c)*; details in Appendix C.2).

3.4 Human Naturalness Evaluation

After each conversation, participants reported either the perceived turn-taking naturalness, the emotions they experienced, their perceived emotional intensity, the emotions they perceived from their partner, or how well their own and their partner's or model emotions fit together in the conversation, depending on the subset. Appendix C includes all the questionnaires used in each subset and more details on the evaluation protocol.

In addition to the ratings provided by the participants, we added two new annotations to each conversation. First, the research supervisor managing the recording sessions answered the same questions after each conversation. Then, an independent evaluator blind to condition re-rated the conversations (Appendix D.1).

3.5 Statistics, Analysis and Insights of CONTACT

Tables 2 and 3 report corpus statistics and naturalness ratings. In these and the following tables, participant ratings are taken only from the naive speaker, who is not modifying their turn-taking or affective behavior; for natural baselines, where neither participant is manipulated, the two ratings are averaged. Agreement between the uninstructed participant and the supervisor (P–S) is given as single-rater absolute-agreement ICC(A,1) and Within-1, the proportion of conversations whose ratings differ by at most one point; H–AI MOS is averaged within each session–model pair and then across sessions. Figure 1 highlights some corpus characteristics. H–H dialogue durations cluster around four minutes, while H–AI dialogues are generally longer, (a). The most common dialogue length is 20–40 turns, with a wider spread in Awkward Turn-Taking, (b). Median turn lengths are around ten words, with substantial variability, (c). Frequent words such as "day," "people," "work," and "feel" reflect familiar everyday topics that encourage natural conversation, (d). H–H pairs span a range of relationships, from strangers to friends, family members, and partners (e), so elicited disruptions occur across varying levels of familiarity (see Appendix E.7 for a relationship-conditioned analysis). The sample covers multiple U.S. states, with Texas contributing the largest share (11.6%; f).

Which disruptions hurt most? All elicited disruptions were rated below their natural baselines, but unequally (Table 3): disruptions related to leaving the floor unresolved were rated as more unnatu-

Table 3: **Human naturalness ratings and recording statistics for the three CONTACT subsets. Combined** is the mean score of participant and supervisor MOS. Human–AI cells are ordered as turn-taking, affective appropriateness, and overall.

Rank	Condition / Model	Participant MOS	Supervisor MOS	Combined MOS	P–S ICC(A,1)	P–S Within-1	Recordings	Hrs.
(a) Awkward Turn-Taking (H–H)								
1	Natural Baseline	4.62	4.93	**4.78**	–	–	30	2.2
2	Premature Entry	3.32	3.61	**3.46**	–	–	30	2.1
3	Blocked Invited Response	3.23	2.80	**3.02**	–	–	30	2.1
4	Competitive Floor Conflict	3.02	2.73	**2.88**	–	–	30	2.1
5	Fragmented Turn Holding	2.78	2.22	**2.50**	–	–	30	2.2
6	Unmarked Turn Abandonment	2.50	1.96	**2.23**	–	–	30	2.1
	Total	–	–	–	**0.70**	**0.89**	**180**	**12.8**
(b) Affective Mismatch (H–H)								
1	Natural Coordination	4.26	4.80	**4.53**	–	–	25	1.7
2	Disproportionate Emotional Overreaction	3.32	3.16	**3.24**	–	–	25	1.7
3	Under-Response	2.24	2.36	**2.30**	–	–	25	1.7
4	Valence Mismatch	2.16	1.64	**1.90**	–	–	25	1.6
	Total	–	–	–	**0.59**	**0.80**	**100**	**6.7**
(c) Open-Ended Human–AI Conversations	*(cells: Turn-taking | Affective Appropriateness | Overall)*							
1	GPT-Live-1 mini	3.69|3.81|3.46	3.67|3.87|3.71	**3.68**|**3.84**|**3.59**	–	–	31	2.9
2	Gemini 3.1 Flash Live Preview	3.97|4.17|3.88	3.19|2.83|2.89	**3.58**|**3.50**|**3.38**	–	–	55	5.1
3	Cascade V2	3.39|4.11|3.36	2.83|3.72|3.22	**3.11**|**3.92**|**3.29**	–	–	22	2.4
4	Qwen3.5 Omni+	2.93|3.39|3.09	2.46|3.08|2.61	**2.69**|**3.23**|**2.85**	–	–	53	5.1
5	Moshi	1.84|2.34|1.79	1.74|1.92|1.89	**1.79**|**2.13**|**1.84**	–	–	22	2.0
6	Cascade V1	1.68|1.88|1.56	1.18|1.18|1.14	**1.43**|**1.53**|**1.35**	–	–	30	2.9
	Total	–	–	–	**0.53**|**0.39**|**0.51**	**0.73**|**0.68**|**0.73**	**213**	**20.3**

ral than those involving overlap, and responding with low affective engagement, below responding too strongly. Both patterns involve withheld contributions, which motivates the joint-silence target modeling proposed in Section 4 for automatic evaluation of naturalness. In H–AI, Cascade V1 and V2 share the same language model yet differ sharply on every dimension, indicating that naturalness depends on how responses are delivered, not only on what is said.

What determines the naturalness from a human perspective? From the answers of the participants to open-ended questions in the H-AI subset, timing and understanding are most frequently cited naturalness factors, and timing is also the most cited source of unnatural moments (Fig. 2 (a)). Emotion was often mentioned as a factor but rarely pinpointed as the cause of any specific unnatural moment, whereas verbosity showed the opposite pattern.

Shorter is not better. Systems in Fig. 2 (b) are sorted by human-rated turn-taking naturalness and the median response latency does not follow this order. The fastest system (Moshi, about 0.2 s) and the slowest (Cascade V1, 2.9 s) received the two lowest ratings, and Qwen usually responded faster than Gemini and Cascade V2 yet was rated below both. Human response latency to the agents was stable across systems (0.65–1.0 s).

Backchannel frequency does not imply naturalness. The two systems producing the most backchannel candidates, Moshi and GPT-Live-1 mini, rank fifth and first in naturalness, while Gemini and Cascade V2 rank second and third with under 2% (Fig.2 (c)), suggesting that placement and context matter more than rate.

4 Surprisal-Based Naturalness Scoring from Natural Dialogue

Perceived turn-taking naturalness is not a monotone function of latency, backchannel rate, or turn length: the same behavior can be preferred in one context and disruptive in another, and judgments are shaped by local breakdowns rather than averages (Roddy & Harte, 2020). We therefore treat naturalness as conformity with context-conditional expectations: in a conversation, the next turn or a continuation is natural to the extent that a model trained only on natural human dialogue would have anticipated it given the preceding interaction. This view dictates what the model predicts and over which horizons. A short 2 s horizon of joint voice activity captures who speaks when, including response onset and backchannel placement; a longer 4 s horizon of joint silence captures dead air and disengagement that can follow a disruption. The responder's arousal–valence state captures whether the affective continuation fits. Scoring a conversation through the *surprisal* of its realized continuation requires neither a matched reference conversation nor naturalness-label supervision of the predictor. We use this scoring to quantify turn-taking naturalness, affective appropriateness, and a combination of both that we call overall naturalness. In all experiments, the two speakers are recorded in two separate channels.

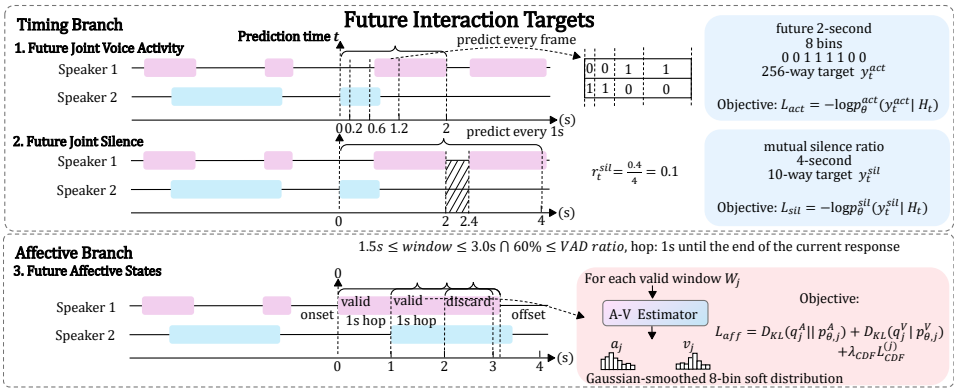


Figure 3: Future interaction targets used to model future turn-taking and affective activity.

4.1 Future Interaction Targets

In our approach, we consider the following three predictive targets, as illustrated in Fig. 3 .

Future joint voice activity. Following VAP (Ekstedt & Skantze, 2022), for a time t in a conversation, the next 2 s are divided into four non-uniform intervals in which voice activity is predicted. Ground truth is obtained with Silero VAD (Silero Team, 2024), which determines whether each speaker is active in each interval, giving eight binary indicators and a 256-way future-activity state.

Future joint silence. Every 1 s, we measure the fraction of the following 4 s during which neither speaker is active and quantize this joint-silence ratio into 10 categories. This complements the short-horizon activity target with a longer-horizon measure of conversational inactivity.

Future affective state. To capture affective dynamics, we extract arousal–valence (A–V) targets from future speech using a frozen A–V estimator released with Vox-Profile (Feng et al., 2025). Each valid future window yields continuous arousal and valence estimates in $[0, 1]$, which are converted into Gaussian-smoothed distributions over eight ordered bins for each dimension.

The precise window selection, target construction, and implementation details are in Appendix E.1.

4.2 Predictive Learning from Natural Human Dialogue

We train exclusively on a large corpus of two-speaker natural conversations. No human naturalness ratings, elicited interaction failures, or natural-versus-unnatural labels are used for training. Future joint activity and silence prediction use categorical negative log-likelihood loss functions ($\mathcal{L}_{\text{act}}$ and $\mathcal{L}_{\text{sil}}$, respectively), while the soft affective target uses KL divergence together with an ordinal regularizer over arousal and valence ($\mathcal{L}_{\text{aff}}$). The full objective is

$$\mathcal{L} = \lambda_{\text{act}}\mathcal{L}_{\text{act}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \lambda_{\text{aff}}\mathcal{L}_{\text{aff}} + \mathcal{L}_{\text{aux}}, \tag{1}$$

where $\mathcal{L}_{\text{aux}}$ contains the auxiliary DualTurn timing objectives for current voice activity, end-of-turn, hold, beginning-of-turn, and backchannel prediction (Rajaa, 2026). λ weights compensate for differences in the scale and density of the three predictive losses. Detailed target, loss, weights, and optimization definitions are provided in Appendix E.

4.3 Naturalness score from Predictive Surprisal

At evaluation time, we compare observed voice activity and affective states (extracted with Silero and Vox-Profile, respectively) with distributions predicted with our model from the preceding dialogue history through a sliding window. Voice-activity and joint-silence scores are mean log-likelihoods over valid prediction positions. For affective state, predictive gain is the target-to-prior KL divergence minus the target-to-prediction KL divergence, using the teacher-derived soft A–V target and the affect prior estimated from natural training data. Gains are averaged within responses and then equally across responses for each speaker; the negative absolute difference between speaker-level means forms the affective component. After standardization with development-set statistics, the components are fused with development-selected weights into timing (future joint voice activity and future joint silence), affective (future affective state and future joint silence), and overall (their mean) scores (Appendix E.3).

5 Experiments and Results

5.1 Experimental Setup

Data. We train on 28,038 paired English conversational recordings (1,636 h) from the naturalistic Interpersonal Circumplex (IPC) portion of Seamless Interaction (Agrawal et al., 2025). All scorers use the same CONTACT evaluation recordings: H–H Awkward Turn-Taking (180), H–H Affective Mismatch (100), and Open-Ended H–AI (213), with development/test splits of 60/120, 32/68, and 70/143, respectively. Session-level splits enforce speaker disjointness across all three subsets while balancing session metadata and participant–supervisor agreement.

Metrics. For H–H discrimination, paired accuracy measures how often natural baseline recordings score higher than manipulated ones within the same session. C-index considers all natural baseline–manipulated pairs within each subset and split, including cross-session pairs. Both assign half credit to ties. Spearman's ρ measures recording-level correlation with P, S, and C ratings. Higher is better for all metrics. Significance uses two-sided paired session-cluster bootstrap tests (20,000 resamples); 95% intervals for targeted ablations are in Appendix E.6.

Training and Architecture Details. We use a frozen Mimi encoder and a shared causal Qwen2.5-0.5B backbone initialized from DualTurn, with separate timing and affective Transformer branches. Training proceeds from future joint voice-activity prediction to affective-state prediction and finally joint-silence prediction, using 4 A100 GPUs, AdamW, and a global batch size of 4. Further details are provided in Appendix E.4.

Baselines. VAP (Ekstedt & Skantze, 2022) tests future-activity likelihood alone, while DualTurn (Rajaa, 2026) provides a reference for improvements over our pretrained initialization. Talking Turns (Arora et al., 2025) tests whether event-level turn-taking decision accuracy reflects perceived naturalness. UTMOSv2 (Baba et al., 2024) tests whether general speech quality alone captures conversational naturalness. TRACE (Ugandhar et al., 2026), an emotional entrainment detector for dyadic speech, tests whether affective coordination reflects perceived conversational naturalness. UniSRM (Wang et al., 2026) provides a contextual speech-reward baseline: we adapt its *Emotion & Prosody Match* and *Overall Naturalness* rubrics to assess affective appropriateness and overall naturalness, respectively. Our TimingStats control tests whether basic timing statistics suffice without context-conditioned prediction. Scoring details are provided in Appendix E.5.

5.2 Results and Discussion

Controlled H–H Evaluation. Our scorer gives the strongest held-out MOS alignment on both CONTACT H–H subsets and the best affective discrimination (Table 4). Gains over the best displayed baseline are significant for affective-mismatch accuracy and combined turn-taking MOS correlation ($p < .05$). These results support the proposed scoring design for detecting affectively inappropriate interactions while remaining informative about their perceived naturalness. Among baselines, DualTurn provides the strongest timing results, consistent with the value of context-conditioned future-activity likelihood. However, its strong discrimination does not translate into equally strong MOS alignment; UniSRM and Talking Turns also distinguish perceived timing disruptions while tracking graded ratings poorly. TRACE shows the clearest baseline alignment with affective MOS, but its discrimination remains modest and its timing performance is weak. UTMOSv2 and TimingStats provide little MOS alignment, indicating that the evaluated speech-quality predictor and global timing summaries are insufficient proxies. These contrasts motivate CONTACT's human-grounded evaluation: detecting a disruption and capturing its perceived impact are distinct requirements.

Naturalistic H–AI Evaluation. Our scorer also achieves the highest combined-MOS correlations across all three H–AI dimensions, with significant gains in overall naturalness for participant and combined ratings. Its predictor thus supports both controlled and spontaneous interaction assessment despite being trained on natural human dialogue. DualTurn remains the strongest baseline by combined MOS, whereas UniSRM and Talking Turns show better timing and affective alignment than on the corresponding H–H Eval. This may reflect closer task alignment: UniSRM learns contextual response comparisons involving synthesized speech (Wang et al., 2026), while Talking Turns here evaluates agent-side decisions. Conversely, TRACE's H–H affective advantage does not persist in H–AI. These changes suggest that success in one interaction setting does not guarantee perceptual validity in another, reinforcing the need for both CONTACT settings.

Ablations and Analysis. Score-removal ablations reveal complementary contributions. Removing future joint silence from modeling lowers H–H affective accuracy more than removing A–V (21.6 vs. 7.8 points), although the difference is not significant, it is consistent with affective judgments reflecting temporal engagement as well as vocal emotion: awkward silence may weaken perceived

Table 4: **Naturalness evaluation on the CONTACT development and test sets.** Acc.: paired accuracy (%); C-I: C-index; P, S, C: Spearman ρ with Participant, Supervisor, and Combined ratings; inter-rater rows give P–S ρ. Bold marks column maxima, excluding reference rows. Ablations remove score components from the same checkpoint. †: gain over the best baseline; ‡: decrease relative to Ours in targeted component tests (both $p < .05$; see Metrics).

(a) H–H Eval	Discrimination								MOS correlation											
Scorer	Turn-Taking				Affective				Turn-Taking						Affective					
	Dev		*Test*		*Dev*		*Test*		*Dev*			*Test*			*Dev*			*Test*		
	Acc.	C-I	Acc.	C-I	Acc.	C-I	Acc.	C-I	P	S	C	P	S	C	P	S	C	P	S	C
Baselines																				
UTMOSv2	44.0	.47	50.0	.50	54.2	.51	51.0	.46	.03	.04	.04	-.01	-.02	-.01	.06	-.05	.00	-.21	-.07	-.16
VAP	70.0	.70	80.0	.76	45.8	.48	41.2	.47	.19	.12	.17	.12	.16	.15	.08	-.22	-.08	-.27	-.19	-.27
DualTurn	**84.0**	**.80**	85.0	.77	58.3	.59	60.8	.55	.27	.24	.28	.18	.17	.19	.15	.02	.06	.03	-.03	.00
Talking Turns	68.0	.65	72.0	.67	45.8	.49	54.9	.58	-.05	-.02	-.04	.06	.03	.06	.13	.17	.19	.04	.00	.03
UniSRM	68.0	.65	63.0	.61	62.5	.56	42.2	.43	.27	.19	.26	.02	.03	.03	-.06	.10	-.01	-.22	-.13	-.21
TRACE	64.0	.56	53.0	.51	50.0	.49	62.7	.61	.17	.24	.21	.00	.10	.06	-.06	.12	.06	.17	.15	.17
TimingStats	62.0	.64	61.0	.61	50.0	.50	56.9	.52	.08	.22	.16	.00	.18	.09	-.26	.00	-.10	-.07	.10	.01
Ours and score-component ablations																				
Ours	78.0	.72	**86.0**	**.80**	**83.3**	**.77**	**82.4**†	**.72**	**.28**	**.37**	**.35**	**.35**	**.33**	**.36**†	**.39**	**.33**	**.40**	**.26**	**.29**	**.28**
w/o A–V	78.0	.72	**86.0**	**.80**	**83.3**	.75	74.5	.71	**.28**	**.37**	**.35**	**.35**	**.33**	**.36**	.38	.31	.39	.24	**.29**	.27
w/o future joint silence	80.0	.78	84.0	.77	62.5	.62	60.8	.64	**.28**	.28	.31	.20	.19‡	.22	.13	.12	.15	.07	.17	.12
w/o future voice activity	60.0	.58	71.0‡	.69‡	**83.3**	**.77**	**82.4**	**.72**	.26	.32	.31	.33	.31	.34	**.39**	**.33**	**.40**	**.26**	**.29**	**.28**
Reference (not an audio scorer)																				
Inter-rater ρ (P–S)	–	–	–	–	–	–	–	–	–	–	.75	–	–	.76	–	–	.66	–	–	.69

(b) H–AI Eval	Turn-Taking						Affective						Overall					
Scorer	*Dev*			*Test*			*Dev*			*Test*			*Dev*			*Test*		
	P	S	C	P	S	C	P	S	C	P	S	C	P	S	C	P	S	C
Baselines																		
UTMOSv2	.12	.22	.20	.07	-.02	.02	.15	.10	.15	.08	-.09	-.01	.21	.12	.19	.06	-.06	-.01
VAP	.20	.20	.24	.23	.19	.23	.09	.09	.09	.21	.19	.22	.21	.12	.19	.20	.18	.21
DualTurn	.46	**.41**	.50	.36	.27	.35	.38	.22	.33	.29	.17	.26	.34	.30	.37	.30	.15	.26
Talking Turns	.05	.06	.08	.21	.09	.18	.06	-.13	-.06	.18	.04	.13	.05	-.02	.04	.18	.13	.19
UniSRM	.25	.29	.30	.18	.14	.18	.18	.16	.15	.19	.23	.24	.21	.16	.20	.18	.23	.22
TRACE	.10	.09	.09	.11	.09	.10	.08	.24	.17	.10	.04	.06	.16	.21	.19	.09	.14	.13
TimingStats	.02	-.07	-.05	-.08	.09	.01	.06	.13	.09	-.05	.18	.07	.09	.14	.09	.03	.18	.09
Ours and score-component ablations																		
Ours	.52	.39	.52	.40	**.31**	**.40**	**.49**	**.27**	**.41**	**.36**	.29	**.37**	.45	.38	.47	.41†	.23	.36†
w/o A–V	.52	.39	.52	.40	**.31**	**.40**	.47	.26	.39	.34	.23‡	.32‡	.44	.38	.47	.40	.21	.34
w/o future joint silence	.42	.37	.46	.33	.26	.33‡	.27	.07	.22	.22	**.32**	.33	.37	.26	.36	.35	.23	.33
w/o future voice activity	**.55**	.39	**.53**	**.41**	**.31**	**.40**	**.49**	**.27**	**.41**	**.36**	.29	**.37**	**.46**	**.40**	**.49**	**.43**	.22	**.37**
w/o timing outputs	–	–	–	–	–	–	–	–	–	–	–	–	.29	.10	.22	.25	**.25**	.27
Reference (not an audio scorer)																		
Inter-rater ρ (P–S)	–	–	.58	–	–	.55	–	–	.38	–	–	.39	–	–	.55	–	–	.54

emotional responsiveness. Nevertheless, A–V removal significantly reduces H–AI affective correlations with supervisor and combined ratings, demonstrating additional value beyond timing cues. Removing future joint silence from modelling also significantly reduces turn-taking MOS alignment for H–H supervisor and H–AI combined ratings. Removing the VAP-style future-activity component slightly improves H–AI participant timing correlation but significantly reduces both H–H timing discrimination metrics, exposing a trade-off between spontaneous-rating alignment and disruption sensitivity. The full scorer therefore offers balanced coverage rather than maximizing every individual metric. Overall, it combines complementary timing and affective signals to support both disruption detection and human-rating alignment across H–H and H–AI settings, establishing a strong reference for future research.

6 Conclusion And Future Work

In this paper, we present CONTACT, a publicly available human-grounded benchmark with perceived human naturalness ratings that comprises human conversations with a variety of turn-taking naturalness and affective perceived appropriateness, and a set of human-AI conversations involving several speech AI agents. To showcase the usefulness of the benchmark, we propose and evaluate a method to provide turn-taking naturalness and affective appropriateness metrics. Results suggest that CONTACT provides a useful benchmark for validating naturalness algorithms and metrics that can ultimately help improve speech AI agents. In future versions of CONTACT, we will provide perceptual annotations from more evaluators with and without using facial expression, will include turn-level emotion annotations as well as perceived dominance and stance. Lastly, we will reproduce these methods in other languages, such as Spanish, and other English dialects.

AI USE STATEMENT

We used generative AI tools to assist with participant-facing materials, code debugging and refinement, and manuscript preparation, as follows. AI was used to propose candidate conversation topics and speaking prompts under author-defined constraints, including that topics be low-stakes, broadly accessible, and conducive to natural unscripted conversation, and to refine the wording of participant instructions. AI tools were also used to improve the clarity and fluency of the manuscript text and detection of typos. AI-assisted, prompts, instructions, and manuscript suggested corrections were reviewed and verified by the authors before use. AI-generated suggestions were not used as participant data, annotations, or ground-truth labels. The study design, experimental conditions, annotation protocol, modeling methodology, analyses, interpretation of results, and conclusions were determined by the authors. The authors take responsibility for all final content and research artifacts.

ETHICS STATEMENT

The study's data collection, analysis, and public-sharing procedures were approved by an Institutional Review Board (IRB).[2] All participants provided written informed consent before recording. Participation was voluntary, and participants could decline questions, stop recording, or withdraw without penalty. They were informed whether their conversational partner would be another person or an AI system. The consent described potential discomfort, including fatigue, self-consciousness, and awkwardness during recorded interactions. All participants were compensated for their time.

Consent explicitly covered public release and subsequent research of the recordings and associated transcripts and annotations. Participants were informed that, despite removing direct identifiers, their voices and, where collected and released, images could remain recognizable. Complete confidentiality therefore cannot be guaranteed. The consent also explained the limits of withdrawal: requests to destroy recordings might no longer be actionable after identifiers had been removed, and withdrawal could not prevent continued use of data already shared publicly.

Participants were instructed not to disclose personal or sensitive information during recording and were trained to use made-up names when needed. If the research supervisor detected any potential personal identifiers being shared during the recording, the recording was stopped and deleted after checking with the speaker.

CONTACT is intended to support the evaluation of conversational systems. Its perceptual ratings should be interpreted within the linguistic, cultural, and interactional contexts represented in the dataset, rather than as universal standards of conversational behavior. The study received industry support; the sponsor will be identified in the non-anonymous version.

REPRODUCIBILITY STATEMENT

The redistributable CONTACT recordings, including audio, video, transcripts, participant-facing prompts and instructions, experimental conditions, and perceptual annotations, are publicly released[3]. GPT-Live-1 mini recordings are excluded from public release due to OpenAI terms and conditions[4]; their annotations and aggregate evaluation results remain included in the benchmark. We also provide a public repository containing the code, data splits, configuration files, and implementation details for preprocessing, predictive-target construction, model training, scoring, baseline evaluation, and metric computation. The main paper and appendices describe the recording protocol, questionnaire items, model architecture, training objectives, score aggregation, and evaluation procedures, and identify the third-party models and datasets used in the pipeline.

ACKNOWLEDGMENTS

To be included after anonymization period.

[2]The approving institution and protocol identifier are withheld for anonymous review and will be disclosed after acceptance.

[3]Link available after acceptance of the paper or end of the anonymization period. A link with illustrative samples is shared within the article's main text

[4]https://openai.com/policies/service-terms/

REFERENCES

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 56898–56918. Curran Associates, Inc., 2024. doi: 10.52202/079017-1813. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/681fe4ec554beabdc9c84a1780cd5a8a-Paper-Datasets_and_Benchmarks_Track.pdf.

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. Talking Turns: Benchmarking Audio Foundation Models on Turn-Taking Dynamics. In *International Conference on Learning Representations*, 2025. URL https://mlanthology.org/iclr/2025/arora2025iclr-talking/.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

Boaz M Ben-David, Namita Multani, Vered Shakuf, Frank Rudzicz, and Pascal H H M van Lieshout. Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *J. Speech Lang. Hear. Res.*, 59(1):72–89, February 2016.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42: 335–359, 2008. URL https://api.semanticscholar.org/CorpusID:11820063.

Pedro Corrêa, João Lima, Victor Moreno, Lucas Ueda, and Paula Costa. Evaluating emotion recognition in spoken language models on emotionally incongruent speech. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17952–17956, 2026. doi: 10.1109/ICASSP55912.2026.11462441.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. Technical report, Kyutai, September 2024. URL http://kyutai.org/Moshi.pdf.

Yayue Deng, Guoqiang Hu, Haiyang Sun, Xiangyu Zhang, Haoyang Zhang, Fei Tian, Xuerui Yang, Gang Yu, and Eng Siong Chng. Multi-bench: A multi-turn interactive benchmark for assessing emotional intelligence ability of spoken dialogue models, 2025. URL https://arxiv.org/abs/2511.00850.

Yuhao Du, Qianwei Huang, Guo Zhu, Zhanchen Dai, Shunian Chen, Qiming Zhu, Yuhao Zhang, Li Zhou, Benyou Wang, and Haizhou Li. Mtalk-bench: Evaluating speech-to-speech models in multi-turn dialogues via arena-style and rubrics protocols, 2025. URL https://arxiv.org/abs/2508.18240.

Erik Ekstedt and Gabriel Skantze. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*, pp. 5190–5194, 2022. doi: 10.21437/Interspeech.2022-10955.

Tiantian Feng, Jihwan Lee, Anfeng Xu, Yoonjeong Lee, Thanathai Lertpetchpun, Xuan Shi, Helin Wang, Thomas Thebaud, Laureano Moro-Velazquez, Dani Byrd, et al. Vox-profile: A speech foundation model benchmark for characterizing diverse speaker and speech traits. *arXiv preprint arXiv:2505.14648*, 2025.

Freeman Jiang, Ramon Sanabria, Soham Deshmukh, Bandhav Veluri, Simon Michael Vuch Williams, Elliott K. Suen, Garreth Lee, Kevin Yoonho Choi, Takuya Umeki, Riku Kubo, Sathvik Udupa, Chien-yu Huang, Shih-Yun Shan Kuan, Zhuoyan Tao, Satyapriya Krishna, Sefik Emre Eskimez, Yu Tsao, Hung-yi Lee, and Shinji Watanabe. Turnbench: A multi-domain benchmark for turn-taking dynamics in spoken dialogue, 2026. URL https://arxiv.org/abs/2608.25218.

Namkje Koudenburg, Tom Postmes, and Ernestine H Gordijn. Disrupting the flow: How brief silences in group conversations affect social needs. *Journal of Experimental Social Psychology*, 47(2):512–515, 2011.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-Yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2025, Honolulu, HI, USA, December 6-10, 2025*, pp. 1–8. IEEE, 2025a. doi: 10.1109/ASRU65441.2025.11433838. URL https://doi.org/10.1109/ASRU65441.2025.11433838.

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. Full-duplex-bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 27–36, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-391-3. doi: 10.18653/v1/2026.acl-short.4. URL https://aclanthology.org/2026.acl-short.4/.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, Shinji Watanabe, and Hung-yi Lee. Full-duplex-bench v1. 5: Evaluating overlap handling for full-duplex speech models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19447–19451. IEEE, 2026b.

Yuxin Lin, Yinglin Zheng, Ming Zeng, and Wangzheng Shi. Predicting turn-taking and backchannel in human-machine conversations using linguistic, acoustic, and visual signals. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15310–15322, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.743. URL https://aclanthology.org/2025.acl-long.743/.

Jingwen Liu, Kan Jen Cheng, Jiachen Lian, Akshay Anand, Rishi Jain, Faith Qiao, Robin Netzorg, Huang-Cheng Chou, Tingle Li, Gopala Anumanchipalli, et al. Emo-reasoning: Benchmarking emotional reasoning capabilities in spoken dialogue systems. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025.

OpenAI. Introducing gpt-4.1 in the api, April 2025. URL https://openai.com/index/gpt-4-1/.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Shangeth Rajaa. Dualturn: Learning turn-taking from dual-channel generative speech pretraining. In *Interspeech 2026*, pp. 7228–7232, 2026. doi: 10.21437/Interspeech.2026-2424.

Matthew Roddy and Naomi Harte. Neural generation of dialogue response timings. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2442–2452, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.221. URL https://aclanthology.org/2020.acl-main.221/.

Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178, 2021. doi: 10.1016/j.csl.2020.101178. URL https://doi.org/10.1016/j.csl.2020.101178.

Sathvik Manikantan Napa Ugandhar, Hao Zhang, Alison Gunzler, Yuzhe Wang, Thomas Thebaud, Georgi Tinchev, Venkatesh Ravichandran, and Laureano Moro-Velázquez. Trace: Temporal relationship-aware conversational entrainment detection in dyadic speech, 2026. URL https://arxiv.org/abs/2606.30543.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. Turn-taking and backchannel prediction with acoustic and large language model fusion. 2024. URL https://www.amazon.science/publications/turn-taking-and-backchannel-prediction-with-acoustic-and-large-language-model-fusion.

Yuanyuan Wang, Dongchao Yang, Yayue Deng, Zhiyong Wu, Steven Y. Guo, Helen M. Meng, and Xixin Wu. UniSRM: A unified speech reward model for reasoning-based fine-grained assessment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46346–46366, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2150. URL https://aclanthology.org/2026.acl-long.2150/.

Anfeng Xu, Yashesh Gaur, Naoyuki Kanda, Zhicheng Ouyang, Katerina Zmolikova, Desh Raj, Simone Merello, Anna Sun, and Ozlem Kalinli. Conversational Speech Naturalness Predictor, 2026. arXiv:2603.01467v1.

Shu-wen Yang, Ming Tu, Ting-Wei Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *International Conference on Learning Representations*, volume 2026, pp. 70704–70729, 2026.

Zhixian Zhao, Shuiyuan Wang, Guojian Li, Hongfei Xue, Chengyou Wang, Shuai Wang, Longshuai Xiao, Zihan Zhang, Hui Bu, Xin Xu, Xinsheng Wang, Hexin Liu, Eng Siong Chng, Hung yi Lee, and Lei Xie. The icassp 2026 humdial challenge: Benchmarking human-like spoken dialogue systems in the llm era, 2026. URL https://arxiv.org/abs/2601.05564.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Wang Benyou, and Haizhou Li. Echomind: An interrelated multi-level benchmark for evaluating empathetic speech language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 55554–55588, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5b0f168359c170e6c8d2d7be9a713be4-Paper-Conference.pdf.

A GRAPHICAL ABSTRACT

Fig. 4 includes a graphical abstract of the study.

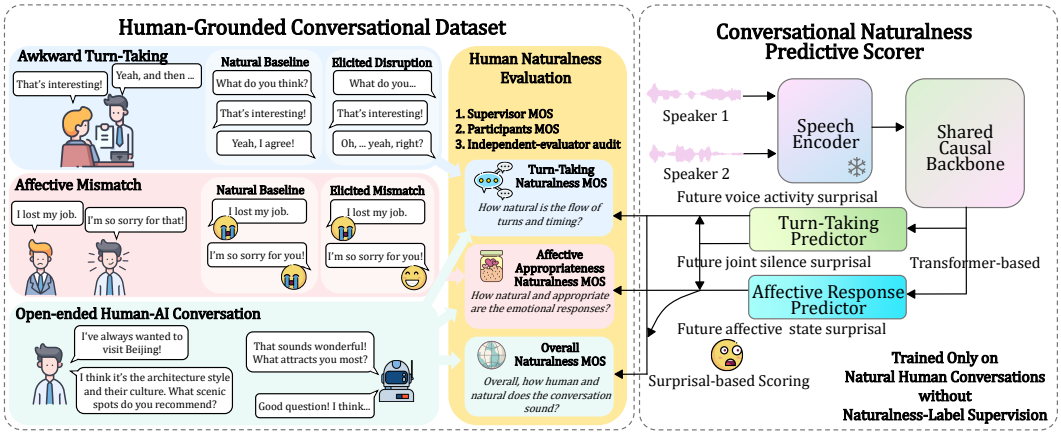


Figure 4: Overview of **CONTACT**. **Left:** H–H and H–AI conversations with human naturalness evaluation. **Right:** Surprisal-based predictive naturalness scorer.

B Detailed Data Recording Protocol

This section documents participant recruitment, recording arrangements, and the operator workflow for the CONTACT recordings.

B.1 Participant Recruitment and Requirements

Participants were recruited through Prolific and scheduled for remotely supervised recording sessions after providing informed consent.

General Participant Requirements and Recording Preparation

Native English Speaker. Be a fluent or native-level speaker of English.

Age. Be age 18 or older.

Conversation Ability. No impediments in hearing, talking or any conversational related ability.

Environment and equipment. A quiet location, a stable Internet connection, and a computer with a functioning microphone are required. Headphones or earbuds are recommended. The human–AI invitation additionally specifies speakers and a working camera for the audio and video session.

Participation. Be comfortable participating in a live audio- and video-recorded conversation. Be willing to interact with another participant during the recording. Be willing to follow simple conversational instructions provided during the session.

Consent and privacy. Consent is completed before recording. Participants are instructed to avoid personal identifiers and highly sensitive information and may end the session at any time.

B.2 Participant-Facing Recruitment Instructions

B.2.1 Human–Human Turn-Taking

Human–Human Turn-Taking Recruitment Instructions

In this study, you will take part in a live conversation session with a matched participant (**stranger/chosen partner**) recruited through Prolific who is also a native English speaker. You will complete a series of short spoken conversations. The goal of the study is to examine conversational turn-taking, including both natural interaction and slightly awkward turn-taking patterns.

1. You will first book a valid time slot and complete a **consent form**.
2. Each session will include **6 short conversations**, each lasting approximately **4–5 minutes**.
3. Conversation topics will involve everyday activities and experiences, such as planning, opinions, storytelling, advice, or mild disagreement.
4. In some conversations, you will be asked to speak naturally. In others, you may receive a simple instruction about how or when to enter, continue, pause, or respond during the conversation.
5. Some conversations may therefore feel slightly less natural than an ordinary conversation. Please follow the instructions provided by the study monitor.

B.2.2 HUMAN–HUMAN AFFECTIVE MISMATCH

Human–Human Affective Recruitment Instructions

In this study, you will take part in a live conversation session with a matched participant (**stranger/chosen partner**) recruited through Prolific who is also a native English speaker. You will complete a series of short spoken conversations. The goal of the study is to examine emotional expression during two-person conversations, including both natural emotional responses and responses that may be less emotionally aligned with the conversational context.

1. You will first book a valid time slot and complete a **consent form**.
2. Each session will include **4 short conversations**, each lasting approximately **4–5 minutes**.
3. Conversation topics will involve everyday experiences or hypothetical situations intended to support natural discussion of different emotional contexts.
4. In some conversations, you will be asked to respond naturally. In others, you may receive a simple instruction about how to express your response during the interaction.
5. Some conversations may therefore feel slightly less natural than an ordinary conversation. Please follow the instructions provided by the study monitor while continuing to respond meaningfully and naturally to what your conversation partner says.

B.2.3 HUMAN–AI CONVERSATIONS

The following invitation specifies 5 conversations.

Human–AI Participant Recruitment Instructions

In this study, you will have **5 short spoken conversations** with AI voice agents. You will be given different topics to chat about, please **feel free to talk** and treat the AI agent as you would in a normal conversation with other people.

1. You will first review and complete a **consent form** before any recording begins.
2. After consent, we can arrange a time slot for the recording.
3. You will be invited to join an online **audio and video recording** session via a secure link. Please participate from a **quiet** location with a **stable internet** connection. You will need a computer with a functioning microphone, speakers, and a working camera. Headphones or earbuds are recommended to ensure audio quality.
4. You don't have to be prepared with the given topic, chat as you normally would, and treat the AI agent as a human.
5. You will receive a link to a questionnaire after you join the studio. After each interaction with the AI voice agent, you will complete a brief set of questions. And at the end of the session, you will also complete a short final survey about your overall experience.

B.3 RECORDING SETUP AND SESSION PROCEDURE

B.3.1 PLATFORM AND AUDIO ROUTING

A research supervisor managed the recording setup and transitions between conversations.

For the human–human subsets, participants joined a remotely supervised recording session from their own computers, while the research supervisor managed the session from a Windows laptop.

When a manipulated condition required private behavioral instructions, these were delivered to the instructed participant through a separate video-call channel so that the conversational partner remained unaware of the manipulation.

For the human–AI subset, recordings were operated from a macOS computer using BlackHole 2ch[5] as a virtual audio-routing device. The routing configuration supported bidirectional audio between the remote participant, the spoken dialogue agent, and the recording system while allowing the supervisor to monitor the interaction locally.

B.3.2 Session Procedure

Each session began after informed consent and scheduling. Participants joined a remotely supervised recording session, and the research supervisor introduced the study procedure and provided the relevant conversational topic before each interaction.

For the human–human subsets, baseline conversations were conducted without special behavioral instructions. For manipulated conditions, one participant received the corresponding instruction privately and briefly practiced the requested behavior before the conversation began. The other participant was not informed of the manipulation and was instructed to continue the conversation naturally. Preparation and practice occurred outside the evaluated dialogue. After each conversation, participants completed the corresponding questionnaire before proceeding to the next condition.

For the human–AI subset, each participant completed five short conversations with spoken dialogue agents. Before each conversation, the research supervisor provided the assigned topic and prepared the corresponding agent. Participants were asked to converse naturally without preparing responses or following a fixed script. After each interaction, both the participant and research supervisor completed the corresponding evaluation questionnaire, and participants completed a final survey after all five conversations.

B.3.3 Exploratory Recording-Path Latency Check

We conducted a separate two-laptop probe of the remote audio path. Both laptops joined the same online recording session. Laptop A routed browser-generated beeps through VB-CABLE[6] to both the platform microphone input and a reference track in Open Broadcaster Software (OBS Studio)[7]. Laptop B played the received audio through its built-in speakers, and A's built-in microphone captured this playback as the return signal. OBS recorded both signals on separate tracks within the same recording, providing a common timeline while retaining any source-specific capture offsets.

The test used 20 sinusoidal beeps at 1 kHz, each lasting 30 ms, at randomized 3–7 s intervals. Both tracks were recorded at 48 kHz. All 20 beep pairs were matched using band-limited RMS envelopes, with onsets defined by a 20% local-peak threshold. Return-minus-reference onset differences had a median of 516 ms (IQR: 503–545 ms), a mean of 537 ms (SD: 63 ms), and a range of 459–667 ms, without capture-offset correction. These measurements include audio routing, live transmission, buffering, playback, acoustic recapture, and possible relative OBS source offsets; they do not isolate network propagation, network round-trip time, or model inference latency.

Response latencies in Fig. 2(b) are computed from the participant and agent WAV tracks exported by the recording platform, not from the OBS probe, as is the H–H natural baseline latency. The probe does not establish per-session transport delays or cross-device alignment offsets, so no post-hoc timing correction is applied; we therefore report recorded response latencies, measured from speech offset to response onset, rather than isolated human reaction times or agent inference times. Because all systems were recorded through the same platform path in varied order, this delay should not systematically favor any system; consistent with this, human response latencies are similar across systems (Fig. 2(b)).

[5] https://existential.audio/blackhole/
[6] https://vb-audio.com/Cable/
[7] https://obsproject.com/

C Dataset Details: Prompts, Questionnaires, and Models

We present topic prompts, condition guidance, and questionnaire items for the human–human and human–AI recordings. Questionnaires appear in blue boxes, topic prompts in tables, and turn-taking condition guidance in cream boxes. Questionnaire headings and layout are adapted for presentation; asterisks indicate required fields. The human–human affective questionnaire is an excerpt.

C.1 Human–Human Conversations

C.1.1 The Awkward Turn-Taking subset

After each conversation, participants rated how natural the interaction felt in terms of conversational flow, including the timing of responses, pauses, interruptions, and how smoothly the two speakers took turns. This subsection includes the questionnaire's details.

Human–Human Turn-Taking Questionnaire

Your relationship to each other:*

Options (single choice): Friend; Romantic partner or spouse; Family member; Roommate or housemate; Classmate or colleague; Strangers; Other: _____

Turn-Taking Naturalness

Overall, how natural did the conversation feel in terms of ***turn-taking***? Think about how smoothly you and your partner took turns speaking (the flow), including the **timing of responses, pauses, interruptions**, and short listener responses such as "yeah," "right," "uh-huh," or "mm-hmm". Did it feel like a normal conversation with your partner during your **daily communication** in terms of ***Turn-Taking***?*

Scale: 1, 2, 3, 4, 5. 1: Very Unnatural; 5: Very Natural.

Topic prompts. Table 5 presents representative prompts organized into six topic families. The prompts provide conversational starting points rather than fixed utterance scripts.

Table 5: Representative human–human turn-taking topic prompts, organized by topic family.

Topic family	Representative topic prompt
Explanation / Instruction	One speaker explains how they would cook a dish they know well or prepare a meal they often make. Walk through the steps, timing, ingredients, possible mistakes, and tips. The other speaker should listen naturally, ask clarifying questions, or respond at natural points.
Advice / Problem Discussion	Ask your partner for advice about how to handle a minor awkward situation with a friend, classmate, coworker, or family member. Explain the situation, what makes it awkward, what options you are considering, and what kind of advice you want.
Story / Experience Telling	One speaker tells a story about a time when a plan did not go as expected. Explain what the original plan was, what went wrong, how people reacted, and how the situation ended.
Hypothetical Reasoning	If you had to move to a completely new city tomorrow, how would you decide where to go? Discuss what factors would matter most, such as work, cost, weather, friends, family, lifestyle, and risk.
Opinion / Preference Discussion	Some people think trips should be carefully planned in advance, while others think trips are better when they are spontaneous. Discuss which style you prefer and why. Try to compare concrete examples such as transportation, budget, activities, and stress.

Continued on next page

Table 5 (continued)

Topic family	Representative topic prompt
Collaborative Decision-Making	Plan a weekend trip together. Decide where you would go, how long you would stay, what you would do, how much planning is needed, and what tradeoffs you would make if time or money were limited.

C.1.2 The Affective Mismatch subset

After each conversation, participants reported the emotions they experienced, their perceived emotional intensity, the emotions they perceived from their partner, and how well their own and their partner's emotions fit together in the conversation.

Human–Human Affective Questionnaire

Your relationship to each other:*

Options (single choice): Friend; Romantic partner or spouse; Family member; Roommate or housemate; Classmate or colleague; Strangers; Other: _____

1. What emotions did you feel during the **early part of the recording** (approximately the first half of the conversation)?*

Options (multiple selections): Neutral; Happy/Joy; Sad; Annoyed; Fear/Anxious; Surprised; Disgust/Contempt; Excited; Calm; Other: _____

2. How intense did your emotion seem relative to the situation?*

Scale: 1, 2, 3, 4, 5. 1: Not intense at all; 5: Very intense.

3. What emotions did you feel **during the later part of the recording** (approximately the second half of the conversation)?*

Options (multiple selections): Neutral; Happy/Joy; Sad; Annoyed; Fear/Anxious; Surprised; Disgust/Contempt; Excited; Calm; Other: _____

4. How intense did your emotion seem relative to the situation?*

Scale: 1, 2, 3, 4, 5. 1: Not intense at all; 5: Very intense.

5. What emotions do you think your partner expressed during the conversation?*

Options (multiple selections): Neutral; Happy/Joy; Sad; Annoyed; Fear/Anxious; Surprised; Disgust/Contempt; Excited; Calm; Other: _____

6. ***Emotion Appropriateness:*** How well do your emotions and your partner's emotions fit together in this conversation? For example, if you are talking about your challenges, does your partner try to support you, or if you are happy, does your partner sound happy as well?*

Scale: 1, 2, 3, 4, 5. 1: Not well at all; 5: Very well.

Topic prompts. Table 6 presents the seven affective topic prompts. The listed emotions describe the intended topic framing, not observed emotion annotations or manipulation-condition labels.

Table 6: Human–human affective topic prompts and their intended affective framing.

Topic emotion / expected affect	Topic prompt
Happy / Joy *Expected affect:* happiness, enjoyment, pleasant surprise	Talk about an ordinary experience that unexpectedly turned out much better than you expected. For example, you might discuss trying a restaurant or food that was surprisingly good, visiting a place that was more enjoyable than expected, joining an activity you did not expect to like, attending an event that turned out to be fun, or receiving a small unexpected benefit or piece of good news. Discuss what happened, what surprised you, and why it made you feel happy.
Sad / Disappointed *Expected affect:* disappointment, mild sadness	Talk about a situation where something you had been looking forward to was unexpectedly cancelled or could no longer happen. For example, a trip being cancelled, an event no longer happening, friends changing plans at the last minute, bad weather ruining an outdoor activity, or discovering that something you had anticipated was unavailable. Discuss why you were looking forward to it and how you would feel when the plan fell through. A hypothetical example is fine.
Angry / Annoyed / Frustrated *Expected affect:* annoyance, irritation, frustration	Talk about an everyday behavior that you find particularly annoying or inconsiderate. For example, someone repeatedly interrupting, cutting in line, playing loud audio in a quiet shared space, arriving late without warning, leaving a shared area messy, or blocking a walkway while not paying attention. Discuss why the behavior bothers you and how people should behave instead. Do not identify a specific person.
Fear / Anxious *Expected affect:* anxiety, nervousness, apprehension	Talk about a situation where you have to perform well in front of other people and do not know how it will go. For example, a job interview, an academic or professional interview, giving a presentation, speaking publicly, answering unexpected questions, or waiting before an evaluation. Discuss which part would make you most nervous. The situation can be hypothetical, and you should not name a real employer, school, interviewer, or identifying detail.
Surprised *Expected affect:* surprise, amazement, curiosity	Talk about an event, fact, coincidence, or story that would make you genuinely think, 'Wait, really?' or 'I did not expect that.' For example, unexpectedly meeting someone in an unlikely place, discovering that two unrelated people know each other, finding something unusual inside an ordinary-looking place, hearing a strange but true fact, or seeing an unexpected result in a game or event. Avoid private gossip or identifiable information.
Disgust / Contempt *Expected affect:* disgust, revulsion, irritation	Talk about a hygiene or cleanliness situation that you would find particularly unpleasant or disgusting. For example, discovering bugs in a kitchen or bedroom, seeing residue on supposedly clean dishes, finding a shared kitchen covered in old food or oil, entering a very dirty bathroom, finding overflowing garbage, or discovering that a place that should be clean has not been cleaned for a long time. Discuss what would bother you most and how you would react.
Excited *Expected affect:* excitement, enthusiasm, anticipation	Talk about a trip, event, or activity that you would be especially excited to experience. For example, visiting a city or country you have wanted to see, going to a concert, sports event, festival, or theme park, trying an activity such as skiing, diving, or hiking, or planning a particularly fun weekend or vacation. Discuss what you would look forward to most, what you would do first, and why it feels exciting. The example can be hypothetical.

C.2 The Open-Ended Human–AI Conversations subset

After conversing with each model, participants assessed turn-taking naturalness by considering how closely the conversational flow resembled everyday human interaction. The question explicitly directed attention to response timing, pauses, turn shifts, turn holds, interruptions, overlaps, and backchannels. Participants were instructed to evaluate these cues for an overall judgment of turn-taking naturalness. Participants reported their own emotions and those perceived in the agent's speech, and assessed whether the agent's emotional responses suited their state and the conversational context. They also responded to open-ended questions about what made the conversation sound more or less natural. They also rated whether the interaction resembled an everyday human conversation on the given topic. Together, these responses captured turn-taking quality, experienced and perceived emotions, emotional appropriateness, and overall naturalness.

The following questionnaire was completed after each human–AI recording.

Human–AI Per-Recording Evaluation Questionnaire

1. *Turn-Taking Naturalness*

Overall, how natural did the conversation feel in terms of ***Turn-Taking***? Think about how smoothly you and the voice agent took turns speaking (the flow), including the **Timing of responses**, **pauses**, **interruptions**, and short listener responses such as "yeah," "right," "uh-huh," or "mm-hmm". Did it feel like a normal conversation talking to a real person during your **daily communication** in terms of ***Turn Taking***?*

Scale: 1, 2, 3, 4, 5. 1: Very Unnatural; 5: Very Natural.

Follow-up: A brief annotation for low score (1,2,3) or high score (4,5) you gave for the last question:_____

2. Participant Emotions

What emotions did you feel during the conversation?*

Options: Neutral; Happy/Joy; Sad; annoyed; Fear/Anxious; Surprised; Disgust/Contempt; Excited; Calm; Unsure; Other: _____

3. Perceived Agent Emotions

What emotions do you think the AI agent exhibited during the conversation?*

Options: Neutral; Happy/Joy; Sad; Angry; Fear/Anxious; Surprised; Disgust/Contempt; Excited; Calm; Unsure; Other: _____

4. *Emotion Appropriateness Naturalness*

How well do your emotions and AI agent's emotions fit together in this conversation? For example, if you are talking about your challenges, does the AI agent try to support, or if the you are happy, does the AI agent sound happy as well?*

Scale: 1, 2, 3, 4, 5. 1: Not aligned at all; 5: Very aligned / fits well.

Follow-up: A brief annotation for low score (1,2,3) or high score (4,5) you gave for the last question:*_____

5. Overall Conversational Naturalness

In general, how natural did this conversation feel? Did it feel like talking to a real person in your daily life, **given this current topic**?*

Scale: 1, 2, 3, 4, 5. 1: Very unhuman; 5: Very human-like.

Follow-up: A brief annotation for low score (1,2,3) or high score (4,5) you gave for the last question:*

Topic prompts. Table 7 lists the six human–AI topic prompts.

Table 7: Human–AI conversation topics and prompts.

Topic	Topic prompt
Sports fandom (team / player / season)	Share your experience being a fan of a sports team. What first drew you to them, what keeps you supporting them, and how you feel about them these days.
Daily routine / normal day	Share what a normal day looks like for you. What parts feel natural, and which parts feel a bit forced.
Life stress (last year)	Share something that's been stressing you out in the last year. It could be your work, study, or an intimate relationship. What makes it hard, and how you deal with it.
Embarrassing moment	Share an awkward or embarrassing moment. What happened, and how you look back on it now. Do you regret it?
Dating 'rules' / social norms	Share a dating 'rule' you've noticed people assume (who texts first, who pays, how fast things move). Do you agree with it?
Technology imagination / future impact	Think about a technology you believe will significantly change everyday life in the future. What excites you about it, what worries you, and how confident you feel about those expectations.

C.2.1 EMPLOYED MODELS

We report the deployed component names exposed by the interfaces rather than inferring undocumented backend snapshots. The following models were used in our dataset:

- **GPT-Live-1 mini**[8] was accessed through ChatGPT Voice; OpenAI describes GPT-Live-1 as a full-duplex voice model capable of continuously listening and responding during pauses and interruptions.
- **Gemini 3.1 Flash Live**[9] was accessed through Google AI Studio using the real-time Gemini Live interface.
- **Qwen3.5-Omni+** was accessed through Qwen Chat and represents the speech-interaction variant of the Qwen3.5-Omni family (Qwen Team, 2026).
- **Moshi** was accessed through Kyutai's hosted real-time dialogue interface (Défossez et al., 2024).
- We additionally evaluated two cascaded systems implemented with Vapi[10]. The **Cascade V1** used GPT-4o Transcribe for ASR, GPT-4.1(OpenAI, 2025) for response generation, and GPT-4o Mini TTS with the Alloy voice for speech synthesis[11]. The **Cascade V2** used xAI speech recognition[12], GPT-4.1 (OpenAI, 2025) for response generation, and xAI speech synthesis with the Eve voice[13].

[8] https://deploymentsafety.openai.com/gpt-live
[9] https://deepmind.google/models/model-cards/gemini-3-1-flash-audio/
[10] https://docs.vapi.ai/quickstart/introduction
[11] https://openai.com/index/introducing-our-next-generation-audio-models/
[12] https://docs.x.ai/developers/model-capabilities/audio/voice
[13] https://x.ai/news/new-flagship-voices

Table 8: **Agreement involving the independent evaluator.** P, S, and I denote the participant, supervisor, and independent evaluator. Human–AI values are ordered as turn-taking | affective response | overall.

CONTACT subset	P–S–I ICC(A,1)	P–S–I Within-1
Awkward Turn-Taking (H–H)	0.70	0.79
Affective Mismatch (H–H)	0.52	0.50
Open-Ended Human–AI Conversations	0.61|0.42|0.57	0.58|0.54|0.58

D Extended Dataset Analysis

D.1 Independent Evaluator Agreement

An independent evaluator, who was not involved in data collection and was informed only of the evaluation criteria, but not the recording conditions or elicited manipulations, rated 24 conversations per subset to provide an additional check on annotation consistency[14]. As shown in Table 8, the preliminary agreement is encouraging for turn-taking: the three-rater ICC(A,1) is 0.70 for H–H turn-taking and 0.61 for H–AI turn-taking, with a Within-1 score of 0.79 for the former. Agreement is more modest for affective mismatch (ICC 0.52) and H–AI affective response (ICC 0.42), while H–AI overall naturalness reaches 0.57. These results suggest that turn-taking judgments are reasonably consistent with those of an additional evaluator, although affective judgments show greater variability. Because independent ratings currently cover only 72 (24×3) conversations, we treat these estimates as preliminary rather than corpus-wide reliability estimates. We plan to extend independent evaluation to the full corpus in a future release.

D.2 Interaction Measures for Human–AI Conversations

Audio timing is measured using WebRTC VAD (30 ms frames, mode 2) on the two speaker-specific recordings. Segments separated by pauses of at most 450 ms for humans or 800 ms for AI are merged; merged segments shorter than 300 ms are discarded. Backchannel detection additionally uses the transcripts.

Response latency. The response latency (response gap in Fig. 2(b)) is the responder's onset minus the paired source speaker's offset. Both segments must last at least 0.40 s; latencies outside $[-2, 8]$ s are excluded. Negative gaps indicate overlap between merged segments. Backchannel candidates are not explicitly excluded. We compute one median per conversation for each response direction. The dashed line shows the median of conversation-level medians for the H–H natural baseline, pooling both directions within each conversation and using 0.45 s merging for both speakers.

Backchannel candidates. Consecutive transcript blocks from the same speaker are merged into runs. An AI run is a candidate if it contains only acknowledgment phrases (e.g., "yeah," "okay," or "I see"), occurs between human runs, the preceding human run is not question-like, and the surrounding context contains no closing expression. No duration or word-count cutoff is applied. For the *voiced overlap* subset, the transcript onset is aligned to the nearest AI VAD onset within 1.0 s, and human VAD must be active at that onset; no minimum overlap duration is required. Candidate counts and overlap-supported candidate counts are each divided by all AI transcript runs in a conversation, then averaged across conversations, including those with zero candidates. These are heuristic candidates, not systematically validated backchannels.

[14]The corpus is released with all the conversations annotated by the independent evaluator

E METHOD: EXTENDED DETAILS

This appendix gives the target construction, predictive objectives, recording-level readout, and implementation details omitted from Section 4.

E.1 FUTURE INTERACTION TARGETS

Future joint voice activity. Let $V_n^{(s)} \in \{0,1\}$ denote the 50-Hz voice-activity detection (VAD) value for speaker $s \in \{1,2\}$ at frame n. Following Voice Activity Projection (VAP) (Ekstedt & Skantze, 2022), the next 2 s are partitioned into four non-uniform intervals

$$\mathcal{I} = \{[0, 0.2), [0.2, 0.6), [0.6, 1.2), [1.2, 2.0]\} \text{ s}.$$

For prediction time t, let $\mathcal{F}_{t,k}$ denote the VAD frames in future interval k. Speaker s is active in interval k when at least half of those frames contain speech:

$$\begin{aligned} b_{t,k}^{(s)} &= \mathbf{1}\left[\frac{1}{|\mathcal{F}_{t,k}|}\sum_{n\in\mathcal{F}_{t,k}} V_n^{(s)} \geq 0.5\right], \\ y_t^{\text{act}} &= \text{index}\Big(\{b_{t,k}^{(s)}\}_{s=1,2;\, k=1,\ldots,4}\Big) \in \{0,\ldots,255\}. \end{aligned} \tag{2}$$

The four activity bits per speaker form an eight-bit joint state with $2^8 = 256$ categories. Speaker activity is obtained using Silero VAD (Silero Team, 2024).

Future joint silence. A silence target is constructed every 1 s. Let $\mathcal{G}_t$ denote the 50-Hz VAD frames in the following 4 s. The fraction of the window during which neither speaker is active is uniformly quantized into ten categories:

$$\begin{aligned} r_t^{\text{sil}} &= \frac{1}{|\mathcal{G}_t|}\sum_{n\in\mathcal{G}_t} \mathbf{1}\Big[V_n^{(1)} = 0 \wedge V_n^{(2)} = 0\Big], \\ y_t^{\text{sil}} &= \min\big(9, \lfloor 10 r_t^{\text{sil}} \rfloor\big) \in \{0,\ldots,9\}. \end{aligned} \tag{3}$$

Future affective state. Affect targets are constructed from speaker-specific, response-aligned future windows. Within each response region, candidate windows begin at the response onset and advance with a 1 s hop. Each window extends for at most 3 s and is truncated at the corresponding response offset. Windows shorter than 1.5 s or containing less than 60% VAD-active speech from that speaker are discarded.

For each valid window j, a frozen arousal–valence (A–V) estimator released with Vox-Profile (Feng et al., 2025) provides continuous $(a_j, v_j) \in [0,1]^2$. Rather than hard-quantizing these values, arousal and valence are each mapped to a Gaussian-smoothed distribution over eight ordered equal-width bins. Let $c_k = (k - \frac{1}{2})/8$ denote the center of bin $k \in \{1,\ldots,8\}$. For $d \in \{a_j, v_j\}$,

$$q_d(k) = \frac{\exp\Big[-\frac{(d-c_k)^2}{2(\sigma_{\text{bin}}/8)^2}\Big]}{\sum_{r=1}^{8}\exp\Big[-\frac{(d-c_r)^2}{2(\sigma_{\text{bin}}/8)^2}\Big]}, \tag{4}$$

where σ_{bin} controls the Gaussian smoothing width in units of one affect bin. We use $\sigma_{\text{bin}} = 0.75$, corresponding to a standard deviation of 0.75 bin widths.

We denote the resulting arousal and valence targets by q_j^A and q_j^V, and the corresponding predictions by $p_{\theta,j}^A$ and $p_{\theta,j}^V$. At each valid window start, the prediction for the corresponding speaker is supervised by these soft A–V targets.

E.2 PREDICTIVE OBJECTIVE

For categorical activity and silence, the predictive losses are negative log-likelihoods:

$$\begin{aligned} \mathcal{L}_{\text{act}} &= -\frac{1}{|\mathcal{T}_{\text{act}}|}\sum_{t\in\mathcal{T}_{\text{act}}} \log p_\theta^{\text{act}}\big(y_t^{\text{act}} \mid H_t\big), \\ \mathcal{L}_{\text{sil}} &= -\frac{1}{|\mathcal{T}_{\text{sil}}|}\sum_{t\in\mathcal{T}_{\text{sil}}} \log p_\theta^{\text{sil}}\big(y_t^{\text{sil}} \mid H_t\big). \end{aligned} \tag{5}$$

Here, $\mathcal{T}_{\text{act}}$ contains valid activity-prediction frames, $\mathcal{T}_{\text{sil}}$ contains valid 1-s-spaced silence prediction times, and H_t denotes the causally available dialogue history.

For affect window j, the per-window objective is

$$\mathcal{L}_{\text{aff}}^{(j)} = D_{\text{KL}}\big(q_j^A \| p_{\theta,j}^A\big) + D_{\text{KL}}\big(q_j^V \| p_{\theta,j}^V\big) + \lambda_{\text{CDF}}\, \mathcal{L}_{\text{CDF}}^{(j)}, \tag{6}$$

where

$$\mathcal{L}_{\text{CDF}}^{(j)} = \frac{1}{14} \sum_{m=1}^{7} \left[\left(\sum_{k=1}^{m} p_{\theta,j}^A(k) - \sum_{k=1}^{m} q_j^A(k) \right)^2 + \left(\sum_{k=1}^{m} p_{\theta,j}^V(k) - \sum_{k=1}^{m} q_j^V(k) \right)^2 \right]. \tag{7}$$

The CDF term preserves the ordinal geometry of the eight A/V bins. We use $\lambda_{\text{CDF}} = 0.25$ heuristically so that the ordinal term acts as a secondary regularizer to the primary KL objectives. The affect objective averages over valid future windows:

$$\mathcal{L}_{\text{aff}} = \frac{1}{|\mathcal{T}_{\text{aff}}|} \sum_{j \in \mathcal{T}_{\text{aff}}} \mathcal{L}_{\text{aff}}^{(j)}. \tag{8}$$

The predictive loss terms are combined as follows:

$$\mathcal{L} = \lambda_{\text{act}} \mathcal{L}_{\text{act}} + \lambda_{\text{sil}} \mathcal{L}_{\text{sil}} + \lambda_{\text{aff}} \mathcal{L}_{\text{aff}} + \mathcal{L}_{\text{aux}}, \tag{9}$$

with $\lambda_{\text{act}} = 0.125$, $\lambda_{\text{sil}} = 0.02$, and $\lambda_{\text{aff}} = 1.0$. $\mathcal{L}_{\text{aux}}$ contains the auxiliary timing objectives for current VAD, end-of-turn, hold, beginning-of-turn, and backchannel prediction (Rajaa, 2026). These weights control the relative contributions of the predictive losses.

E.3 Recording-Level Predictive Readout

Sliding-window inference. The two time-aligned speaker tracks are jointly processed in 20 s windows with 8 s overlap (12 s stride), yielding up to 250 frames per window at 12.5 Hz. The first approximately 3 s provide context without being scored, and the final 2 s are excluded as scoring positions. A deterministic window-assignment rule ensures that each retained frame or prediction anchor contributes exactly once. Padded positions and positions lacking complete future targets are excluded. Predictions use only causal dialogue history; future audio is used solely to construct the corresponding evaluation targets.

For timing, define local predictive discrepancies

$$\ell_t^{\text{act}} = -\log p_\theta^{\text{act}}\big(y_t^{\text{act}} \mid H_t\big), \qquad \ell_t^{\text{sil}} = -\log p_\theta^{\text{sil}}\big(y_t^{\text{sil}} \mid H_t\big).$$

The raw timing components are

$$F = -\operatorname{Mean}\big(\ell^{\text{act}}\big), \qquad S = -\operatorname{Mean}\big(\ell^{\text{sil}}\big). \tag{10}$$

Activity is averaged over all valid, de-duplicated frames in the recording, while silence is averaged over all valid, de-duplicated 1-s-spaced prediction anchors. Each frame or anchor receives equal weight in its respective recording-level mean.

For each response-aligned affective target window, the corresponding speaker's predictive distribution is obtained from the last causal representation available before the window onset. Both interlocutors are included in H–H and H–AI scoring. For affect window j, let

$$q_j^{AV}(k,l) = q_j^A(k) q_j^V(l), \qquad p_{\theta,j}^{AV}(k,l) = p_{\theta,j}^A(k) p_{\theta,j}^V(l). \tag{11}$$

We define the context-conditioned predictive gain

$$g_j = D_{\text{KL}}\big(q_j^{AV} \| \pi^{AV}\big) - D_{\text{KL}}\big(q_j^{AV} \| p_{\theta,j}^{AV}\big), \tag{12}$$

Here, π^{AV} is the fixed affect-target prior estimated from natural training data. We first average g_j over valid target windows within each response, then equally average the valid response-level means for each speaker, yielding $\bar{g}_0$ and $\bar{g}_1$. The raw affective component is the negative absolute difference between these speaker-level means:

$$R_{\text{aff}} = -\,|\bar{g}_0 - \bar{g}_1|. \tag{13}$$

Table 9: Key implementation settings for the predictive naturalness model.

Component	Setting
Natural-dialogue training data	1,635.83 h
Speech encoder	Mimi, frozen
Causal backbone	Qwen2.5-0.5B / DualTurn initialization
Model frame rate	12.5 Hz
VAD frame rate	50 Hz
Training chunk / stride	30 s / 15 s
Timing branch	2 layers, 512 hidden, 8 heads
Affective branch	2 layers, 512 hidden, 8 heads
Dropout	0.1
Silence-head hidden size	128
Future activity	2 s; 4 intervals; 256 states
Future silence	4 s horizon; 1 s stride; 10 bins
Affect window / hop	up to 3 s / 1 s
Minimum affect-window duration	1.5 s
Minimum speaker VAD-active ratio	0.60
Affect targets	8-bin soft A and V distributions
Affect smoothing width	$\sigma_{\mathrm{bin}} = 0.75$ bin
Affect ordinal weight	0.25
λ_{act}	0.125
λ_{sil}	0.02
λ_{aff}	1.0
Backbone learning rate	1×10^{-6}
Activity-head learning rate	1×10^{-4}
Auxiliary-head learning rate	5×10^{-6}
Branch/head learning rate	1×10^{-4}
Weight decay	0.01
Gradient clipping	1.0
Precision	bfloat16
Global batch size	4
Optimizer (refinement stages)	AdamW
GPUs (refinement stages)	4× 80G A100
Random seed	42

For each CONTACT subset, each raw component is standardized using the mean and standard deviation of its recording-level scores on that subset's development split. The corresponding test split uses these fixed statistics. Fusion weights are selected on development data and fixed for held-out evaluation. Let $\widetilde{F}$, $\widetilde{S}$, and $\widetilde{R}_{\mathrm{aff}}$ denote the standardized values. The final scores are

$$T = 0.45\,\widetilde{F} + 0.55\,\widetilde{S}, \quad E = 0.25\,\widetilde{R}_{\mathrm{aff}} + 0.75\,\widetilde{S}, \quad O = \frac{1}{2}(T + E), \tag{14}$$

where T, E, and O denote timing, affective, and overall conversational naturalness, respectively. The affective score combines A–V and joint-silence readouts to capture vocal affect and temporal coordination; the A–V removal ablation supports its complementary contribution.

E.4 Architecture and Optimization

The model uses a frozen Mimi encoder and a shared causal Qwen2.5-0.5B backbone initialized from DualTurn, followed by separate timing and affective branches without learned cross-branch coupling. Architecture and hyperparameters are summarized in Table 9.

Staged optimization. After initial adaptation for future joint voice activity prediction, we train the affective branch with factorized arousal–valence heads, followed by a final refinement stage for future joint-silence prediction. The shared backbone and inactive pathway remain frozen during both refinement stages.

E.5 BASELINE SCORING

We use pretrained checkpoints with fixed model weights. Except for UniSRM, each baseline provides one recording-level score for all evaluation dimensions. Higher scores indicate greater naturalness. Any additional calibration uses development data only, with settings fixed for test evaluation.

VAP and DualTurn. Both models process separate speaker channels, with future-activity targets derived from channel-wise Silero VAD. VAP uses 16 kHz audio and 50 Hz predictions. Its next-2 s target has four bins of widths 0.2, 0.4, 0.6, and 0.8 s; VAD occupancy of at least 0.5 defines activity in each bin. The resulting eight bits encode 256 joint states, scored using categorical NLL. DualTurn uses 24 kHz audio and 12.5 Hz predictions, retaining its eight-output FVAD head with cumulative bin boundaries at 0.24, 0.48, 0.96, and 2.00 s. Its continuous VAD-occupancy targets are scored using binary cross-entropy summed over the eight outputs. Inference uses non-overlapping 60 s chunks for VAP and 30 s chunks for DualTurn. We exclude the first 3 s of each recording and positions lacking complete future targets: chunk-end positions for VAP and recording-end positions for DualTurn. Our recording-level readout averages losses over all valid frames $\mathcal{V}_r$, rather than averaging chunk scores:

$$s_r = -\frac{1}{|\mathcal{V}_r|} \sum_{t \in \mathcal{V}_r} \ell_{r,t}. \tag{15}$$

Talking Turns. The released Whisper-medium predictor receives 16 kHz mono audio averaged from both channels. Every 40 ms, it predicts the next 40 ms event using up to 30 s of preceding audio. We retain the author-provided thresholds without CONTACT retuning: $p_T - p_C > 0$ for turn change and yielding, $p_{\mathrm{BC}} > 0.1$ for backchanneling, $p_I - p_C > -0.45$ for interruption, and $p_T - p_C > -0.1$ for accepting interruption. Here, p_C, p_T, p_{BC}, and p_I denote continuation, turn-change, backchannel, and interruption probabilities.

Reference events are automatically derived from channel-wise pyannote VAD, faster-whisper-medium word timestamps, a backchannel lexicon, and speaker-floor tracking. These follow the authors' VAD–ASR approach but are not claimed to reproduce their labels exactly. Predictions are aligned to complete next-40 ms intervals. We compute ten directional accuracies covering turn taking, backchanneling, interruption, yielding, and interruption acceptance, together with their complementary decisions. Our recording score is their unweighted mean over metrics with evaluable events:

$$\begin{aligned} s_r^{\mathrm{TT}} &= \frac{1}{|K_r|} \sum_{k \in K_r} \frac{n_{rk}^{\mathrm{correct}}}{n_{rk}^{\mathrm{evaluable}}}, \\ K_r &= \left\{ k : n_{rk}^{\mathrm{evaluable}} > 0 \right\}. \end{aligned} \tag{16}$$

Missing event metrics are omitted, not assigned zero. H–H counts are pooled across both focal speakers before computing accuracies; H–AI uses only the AI as focal speaker. Development-calibrated thresholds and aggregation weights are evaluated separately as a sensitivity analysis.

UTMOSv2. We use five pretrained fold models, each evaluated with five stochastic repetitions through the official random-crop frontend. Input is 16 kHz mono audio with silence retained. Averaging the resulting 25 predictions gives the recording score. This is stochastic crop evaluation, not full-recording sliding-window coverage, and serves as a general speech-quality and naturalness control.

UniSRM. We adapt UniSRM's Task-4 pairwise rubric to evaluate one observed response given its preceding dialogue history. The model returns five scores on a 0–10 scale: Intent Matching & Dialogue Act, Speaker Consistency, Contextual Consistency, Emotion & Prosody Match, and Overall Naturalness. This is a single-response adaptation, rather than the original candidate-A-versus-B decision. Responses are segmented using VAD, merging same-speaker segments separated by at most 0.5 s unless the partner starts speaking in the gap. Eligible responses last at least 0.5 s and have at least 0.5 s of preceding speech from each speaker. History contains audio from the recording start to the response onset, mixed to mono; the response contains only the evaluated speaker's channel. We retain complete eligible responses and disable the processor's default audio truncation. Decoding is greedy, with a standard limit of 2,048 new tokens. H–H scores first average responses within

each speaker and then average the two speaker means equally. H–AI scores average all eligible AI responses. Emotion & Prosody Match supplies the affective score; Overall Naturalness supplies the overall score and a proxy for timing. We do not fit rubric weights to CONTACT MOS or use duration weighting or tail pooling.

TRACE. We use the speech-only checkpoint without relationship or context text. Each 16 kHz speaker channel is divided into 3 s windows with a 1 s hop. A pretrained Whisper-large-v3 emotion encoder produces temporally pooled, $1{,}280$-dimensional embeddings, which are L_2-normalized and interleaved by speaker before classification. The recording score is the raw naturalness logit. The checkpoint retains at most 256 windows per speaker, covering approximately the first 258 s.

TimingStats. This in-house statistical control extracts eight features from 50 Hz two-channel Silero VAD: the mean and standard deviation of signed floor-transfer offsets and speech-run durations, overlap and joint-silence ratios, and speaker-switch and short-segment rates per minute. Same-speaker segments are merged across gaps of at most 200 ms when the partner is inactive; short segments last at most 0.8 s. Floor-transfer offsets measure the gap (positive) or overlap (negative) between consecutive different-speaker runs. Features are standardized using $4{,}263$ natural training-reference recordings. A Ledoit–Wolf covariance estimator is fitted to these standardized features, and recordings are scored by their negative squared Mahalanobis distance. No CONTACT ratings or manipulation labels are used for fitting.

E.6 Paired Ablation Intervals

Table 10: Paired score-component ablations on the test split for the comparisons discussed in Section 5.2, including those marked with ‡ in Table 4. Δ is Full minus ablation, computed before rounding. (C) and (S) denote Combined and Supervisor ratings. Accuracy is in percent; its differences and intervals are in percentage points. Intervals are unadjusted, exploratory 95% paired session-cluster bootstrap intervals (20,000 resamples). The difference between the H–H affective accuracy decreases from removing silence and from removing A–V is 13.7 points (95% CI [−11.8, +37.3]; $p = .32$).

Removed	Metric	Full	Ablation	Δ	95% CI
w/o future voice activity	H–H timing Acc.	86.0	71.0	+15.0	[+9.0, +21.0]
w/o future voice activity	H–H timing C-index	0.798	0.687	+0.111	[+0.051, +0.183]
w/o silence	H–H timing ρ (C)	0.364	0.215	+0.149	[−0.010, +0.316]
w/o silence	H–H timing ρ (S)	0.33	0.19	+0.142	[+0.031, +0.264]
w/o silence	H–H affect Acc.	82.4	60.8	+21.6	[0.0, +43.1]
w/o silence	H–H affect ρ (C)	0.281	0.122	+0.159	[−0.036, +0.388]
w/o A–V	H–H affect Acc.	82.4	74.5	+7.8	[−2.0, +19.6]
w/o silence	H–AI timing ρ (C)	0.397	0.333	+0.065	[+0.006, +0.134]
w/o silence	H–AI affect ρ (C)	0.371	0.327	+0.044	[−0.163, +0.256]
w/o A–V	H–AI affect ρ (C)	0.371	0.319	+0.052	[+0.016, +0.097]
w/o A–V	H–AI affect ρ (S)	0.29	0.23	+0.060	[+0.026, +0.101]

E.7 Relationship-Conditioned Evaluation

CONTACT records the relationship between H–H partners (Fig. 1(e)), which can shape what affective behavior is appropriate: the same intensity may be expected between partners but excessive between strangers. We therefore compare the speech-only TRACE checkpoint reported in Table 4 with a relationship-conditioned TRACE checkpoint that additionally receives the partner relationship as input (Table 11). For H–AI conversations, the relationship input is set to strangers.

Relationship conditioning trades off rather than improves performance. The only change consistent across development and test is a higher H–H affective-MOS correlation ($\rho = .061 \rightarrow .166$ and $.169 \rightarrow .235$); H–H affective paired accuracy improves on development but is unchanged on test (62.7%). In contrast, H–H timing discrimination degrades on both splits (paired accuracy $64.0 \rightarrow 42.0\%$ and $53.0 \rightarrow 41.0\%$), and H–AI correlations change inconsistently (e.g., overall $\rho = .191 \rightarrow .187$ on development and $.126 \rightarrow .148$ on test). Relationship information thus appears relevant to affective appropriateness but not to timing. Because the two checkpoints were trained separately, the comparison should not be interpreted as an isolated causal effect of relationship information.

Table 11: Speech-only vs. relationship-conditioned TRACE on the same development and test splits. Acc./C-index: H–H paired accuracy (%) and C-index. MOS columns are Spearman ρ with Combined ratings.

		H–H timing		H–H affective mismatch		H–AI MOS ρ		
Scorer	Split	Acc./C-index	ρ	Acc./C-index	ρ	Timing	Affective	Overall
TRACE (speech-only)	Dev	64.0/.556	.213	50.0/.495	.061	.093	.170	.191
TRACE (+relationship)	Dev	42.0/.416	.096	66.7/.547	.166	.095	.158	.187
TRACE (speech-only)	Test	53.0/.507	.057	62.7/.612	.169	.102	.061	.126
TRACE (+relationship)	Test	41.0/.439	.051	62.7/.641	.235	.117	.081	.148