GESBENCH: WHEN AND WHERE DO MULTIMODAL LANGUAGE MODELS PERCEIVE SOCIAL MEANING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Human social perception is incremental: we infer intent before an utterance is complete, revise our interpretation as new cues arrive, and attend to specific facial, gestural, or prosodic signals that shape social meaning. Current multimodal LLM benchmarks largely evaluate final-answer understanding over complete episodes, rewarding post-hoc recognition of social meaning rather than the unfolding perception required for real-time interaction. Drawing on interactionally embedded Gestalt principles of multimodal human communication, we introduce GESBENCH: a human-grounded benchmark for cue-level social perception in multimodal LLMs. GESBENCH uses cumulative sub-utterance clips to measure when models form stable interpretations of speaker intention and social affordance, and human gaze-guided corruption to evaluate whether regions attended by humans are functionally important for model judgments. Built from more than 6000 utterance-level clips from scripted and unscripted conversational datasets, GESBENCH includes human annotations of unfolding judgments and gaze on a representative subset. Across open and closed multimodal models, we find a clear human-model gap: humans form predictive understandings early and revise more often, while models often hedge or repeat across different clips. GESBENCH shifts social evaluation from whether models eventually recognize meaning to when and where that meaning is formed. Our codes are publicly available at `https://anonymous.4open.science/status/GesBench-share-212C`.

1 INTRODUCTION

Multimodal language models are increasingly evaluated on their ability to understand social meaning: recognizing intentions, emotions, conversational norms, or appropriate responses. Yet, most benchmarks test this ability *after* the model has observed a complete episode. This evaluation setting rewards post-hoc recognition of social meaning, but does not test a capability that is central to real-time interaction: perceiving meaning as the multimodal scene unfolds.

Human social perception is not delayed until the end of an utterance (Schlangen & Skantze, 2011). We often form provisional interpretations from partial words (Grosjean, 1980), facial expressions, gaze shifts, gestures, pauses, and prosody, and revise these interpretations as new cues arrive. For example, when a friend responds to a vacation suggestion with "Well, actually I...", we may already anticipate disagreement, especially if this fragment is accompanied by a furrowed eye brows. If the utterance continues as "...I was thinking of the exact same place!", and the expression changes into raised eye brows and a smile, the interpretation is rapidly revised. The social meaning of the utterance is thus not simply recovered from the complete signal but constructed, tested, and updated through *sub-utterance cues* (SUC).

This view is closely aligned with interactionally embedded Gestalt principles of multimodal human communication of Trujillo & Holler (2023): people bind partial multimodal signals into coherent utterance-level interpretations, while also predicting the speaker's intended social action and the response opportunities available to the listener. In other words, social perception is both predictive (Fan et al., 2019; Ishii et al., 2017) and affordance-oriented (Valenti & Gold, 1991): perceivers do not only infer what the speaker means, but also what actions the situation makes possible.

For embodied agents and multimodal LLMs that are expected to participate in face-to-face interactions, this distinction matters (Fung et al., 2025; Zhang et al., 2025b). An agent that can answer

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

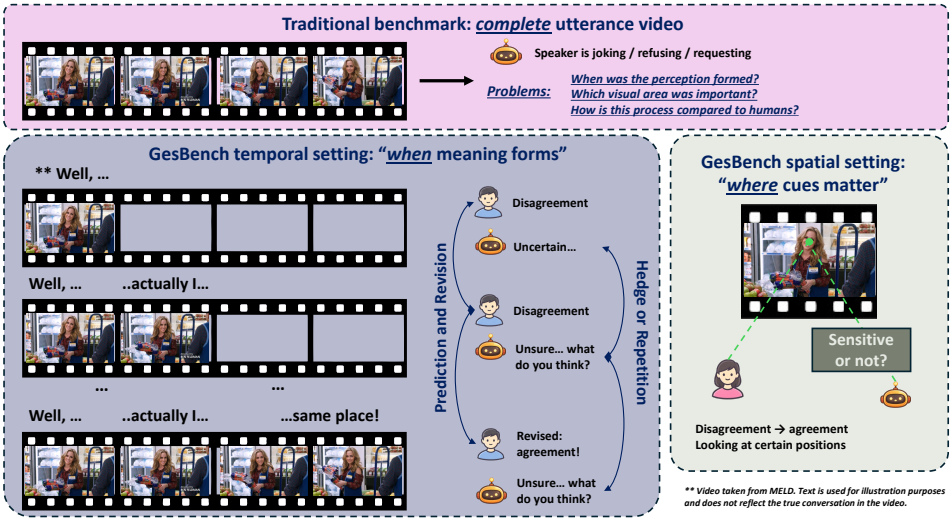


Figure 1: Instead of only asking model perception after the complete utterance, GESBENCH measures **when** and **how** interpretations update. It also collects human gaze during utterance interpretation to compare model judgments of sensitivity to human-attended visual cues. Together, these temporal and spatial probes evaluate when, where, and how social meaning is formed.

correctly after seeing a complete video may still fail to support fluent interaction if it cannot form useful provisional interpretations, attend to communicatively relevant cues (Qi et al., 2025; Upadhyay et al., 2026), or update its beliefs smoothly as the utterance progresses. Existing multimodal and social benchmarks evaluate many important abilities, including video understanding (Zhang et al., 2025a; Li et al., 2024), social reasoning (Mathur et al., 2025), emotion recognition (Lian et al., 2024), and social acting (Wu et al., 2025). However, they are less suited to measuring the unfolding process by which social meaning is formed from fine-grained multimodal cues.

We introduce GESBENCH, a human-grounded benchmark for evaluating cue-level social perception in MLLMs. GESBENCH operationalizes interactionally embedded Gestalt perception along two complementary axes. First, to measure temporal unfolding, we decompose each utterance into cumulative clips and ask both humans and models to infer the speaker's intended social action and the listener's possible response actions. This allows us to measure when are perceptions formulated and how they update through the process. Second, to measure spatial grounding, we collect human eye gaze while annotators interpret utterances, construct human-attended cue profiles, and corrupt relative regions to test whether those cues are functionally important for models.

Our contributions are three-fold: **(1)** We formulate cue-level social perception as an evaluation problem for MLLMs, shifting the focus from final-answer recognition over complete videos to the unfolding process by which social meaning is formed. **(2)** We propose a human-grounded spatio-temporal evaluation protocol: cumulative sub-utterance clips measure how interpretations update, while gaze-guided corruption tests whether perturbing human-attended regions changes model judgments. **(3)** We evaluate open and closed source MLLMs and identify a human model gap in unfolding social perception: humans make earlier commitment in predicting, revise more often while models tend to remain uncertain with little information and make fewer revisions.

Together, these results show that current multimodal LLMs can often recover social meaning after observing an utterance, but still differ from humans in how that meaning is formed. GESBENCH provides a benchmark for studying this gap and for developing models that are more predictive, interactive, and attentive to the fine-grained cues that structure real-time communication.

2 RELATED WORKS

MLLM benchmarks. An abundance of benchmarks have been proposed to measure multi-modal perception of MLLMs. Liu et al. (2024); Meng et al. (2024) proposes comprehensive image under-

standing evaluations for MLLMs, including general perception and reasoning tasks. There are also benchmarks for specializing domains, for example on object hallucination (Li et al., 2023), identifying truthful statements (Kang et al., 2025), and moral reasoning (Jiao et al., 2025). Specifically, for video understanding, Fu et al. (2025); Li et al. (2024) provide general evaluation frameworks regarding videos from film clips to everyday recordings. Other benchmarks focus on social abilities of MLLMs. Zhang et al. (2025a) benchmarks MLLMs in perceiving multi-modal communication signals with a distribution of social tasks. Strachan et al. (2024) has tested the theory of mind (ToM) in MLLMs by testing their abilities in interpreting the other party's minds through tasks such as identifying indirect requests. Mathur et al. (2025) benchmarks MLLMs in understanding long videos and through correct reasoning traces, but it does not explicitly address the temporal updating characteristic. For multi-agent tasks, Zhou et al. (2023); Wu et al. (2025) evaluates the performances of MLLM-powered agents under certain social interaction settings and against various social targets.

Gestalt evaluations in AI models. Attempts have been made to study gestalt principles on both traditional machine learning tasks like computer vision as well as modern MLLMs. Sha et al. (2016) has studied classification problems with regards to the gestalt principles in human vision to evaluate whether the models are recognizing holistic objects. Kim et al. (2021) also points out that some computer vision models are exhibiting gestalt principles in terms of closure recognition. Regarding MLLMs, Dayanandan et al. (2024) has leveraged the dual thinking framework (Kahneman, 2011) to help improve MLLMs' performances when facing adversarial data. However, these benchmarks either focus on low-level Gestalt principles, or do not specifically address Gestalt perception in social interaction. GESBENCH, to our knowledge, is the first benchmark built with real social data and designed to incorporate gestalt features in model and human perceptions.

Spatial grounding and visual cue evaluation in models. Prior work studies whether vision-language models ground answers in appropriate image regions using heatmaps (Selvaraju et al., 2016; Christian et al., 2026), attention analyses (Abnar & Zuidema, 2020), or region-level grounding objectives (Zhang et al., 2025c; Chen et al., 2023; Yuan et al., 2024; Qi et al., 2025). These methods provide useful diagnostics, but they are typically post-hoc or centered on static image understanding rather than the unfolding perception of social meaning in video. Recently more works have been proposed to study whether aligning with human eye gaze help model perception (Peng et al., 2025; Pani & Yang, 2025). GESBENCH instead uses human gaze collected during utterance interpretation to construct human-attended profiles, and evaluates whether corrupting these regions changes model judgments more than corrupting matched comparison regions.

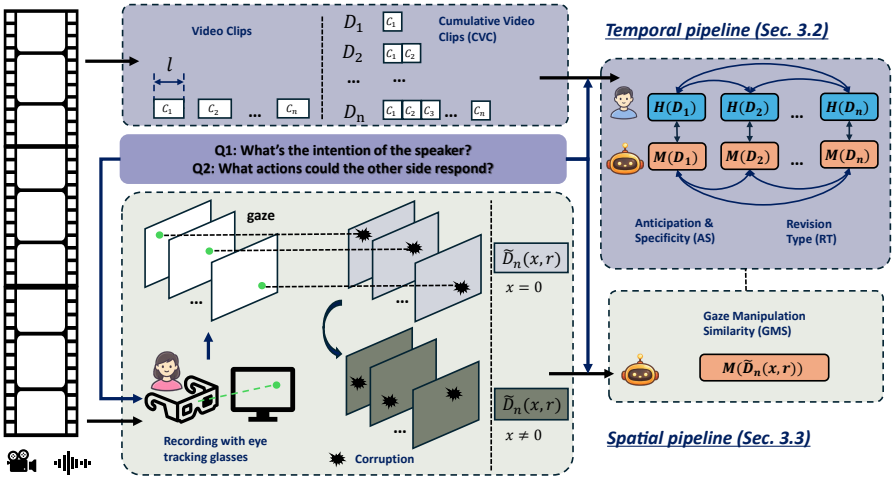


Figure 2: Illustration of the evaluation pipeline. The temporal pipeline takes video inputs, cut into small clips, and build CVC which is then given to both human and models. The spatial pipeline measures human eye gaze data, builds a gaze profile, and does corruption to make gaze based corruption videos with varying radius and offset to prompt the models. Finally, we compute specificity, anticipation, revision type, and gaze manipulation similarity for metrics.

3 THE GESBENCH BENCHMARK

GESBENCH combines scripted and unscripted conversational data to cover both clearly bounded utterances and natural dyadic interaction. Scripted clips are from MELD (Poria et al., 2019) and MIntRec2 (Zhang et al., 2024) while Seamless Interaction (Agrawal et al., 2025) provides unscripted face-to-face conversations. Our benchmark contains more than 6000 utterance-level clips, with human text and gaze annotations collected on representative subsets with around 600 videos. Additional dataset information and selection rationale are provided in Appendix A. The overarching structure of GESBENCH is shown in figure 2. In 3.1 we illustrate our design of tasks, prompts, and questions. Sections 3.2 and 3.3 introduces temporal/spatial pipelines. In 3.4 we demonstrate how metrics are defined and calculated.

3.1 PROMPTS AND QUESTIONS

We leverage single utterances as the basic unit of perception. Referring to the Gestalt Framework framework (Trujillo & Holler, 2023; Holler & Levinson, 2019) and the social affordance theory (Valenti & Gold, 1991), we ask two questions to both human and models, with the full prompts in Appendix B.1:

GESBENCH PROMPT (simplified)

```
Q1: What is the intention of the speaker?
Q2: What actions could the other side take in response?
```

3.2 TEMPORAL PIPELINE

Inspired from approaches which provide models with incremental language (Madureira & Schlangen, 2020; Kahardipraja et al., 2021) we create cumulative versions of each utterance video. Each cumulative clip contains the signal observed from the start of the utterance up to a given point and are given to the models incrementally context. Concretely, we start with a raw video V of single utterance. To create cumulative short clips, we first cut it into short clips of equal length l: $C_1, C_2 \ldots, C_n$ and let Concat$[C_1, C_2, \ldots, C_n]$ to be their sequential concatenation. We define the set of **Cumulative Video Clips (CVC)** with respect to V, by concatenating the clips cumulatively:

$$D_k = \text{Concat}[C_1, \ldots, C_k], \quad k = 1, 2, \ldots, n \tag{1}$$

By design, each D_k has length kl (we take $l = 0.5$ s in practice). The CVC set is then given to both human annotators and the models. We obtain an answer for each video and for both Q_1 and Q_2. We denote human response for a video D_k under question Q to be $H_Q(D_k)$, and model response to be $M_Q(D_k)$. In the temporal benchmarking pipeline, this would yield the human response set $\{H_Q(D_k)\}_{k=1}^n$ and model response set $\{M_Q(D_k)\}_{k=1}^n$, where $Q \in \{Q_1, Q_2\}$.

To obtain the human answers for CVC, we employ a crowdsourcing approach. Paid annotators are recruited on Prolific [1] platform, and the annotation is performed with Covfee (Quiros et al., 2022). We ask them to speak out the answers, which are then transcribed using WhisperX (Bain et al., 2023). The answers are then extracted per question. A model reply answers both questions in one go and we split it into Q_1 and Q_2 with an LLM instruction to extract the answers verbatim; Appendix C.1 to C.2 shows that our conclusions do not depend on this step. Annotators already answer each question in its own field. Sometimes an annotator says "nothing has changed" or similar even when we prompted them to provide reasons. To mitigate this, we carry their previous answer forward when this happens. Appendix A.2 and B details the annotation setup, model prompts, and postprocessing.

3.3 SPATIAL PIPELINE

In the spatial domain, interactionally embedded Gestalt perception is reflected in how perceivers bind communicatively relevant visual signals while filtering out incidental ones. Human viewers do not attend uniformly across a scene; they often orient toward regions that afford social interpretation,

[1] https://app.prolific.com/

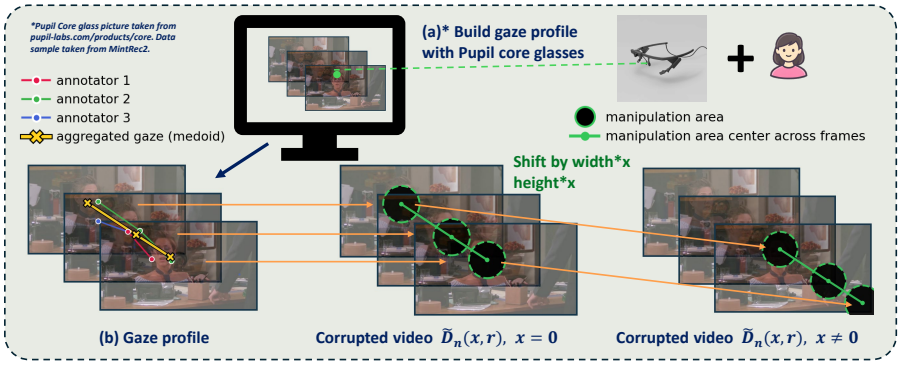


Figure 3: Process of building human attention profile. (a) We record participants' eye gaze data with Pupil Core glasses while watching videos; (b) eye gaze profile by aggregation. Each annotator's gaze for a frame is shown. (c) Build per-frame corrupted video $\widetilde{D}_n(x, r)$ with $x = 0$, by blocking the region centered at the gaze point. (d) Build $\widetilde{D}_n(x, r)$ with $x \neq 0$, shifting the manipulation area. A complete set of video manipulation sample is provided in the supplementary material.

such as faces, gaze direction, hands, body orientation, and visible articulators. This motivates us to measure whether certain regions attended by humans are important to model perception or not.

To obtain a human-grounded cue profile for a given clip, we collect eye gaze while annotators watch complete utterance videos and answer the two benchmark questions. For each video V, annotators wear eye-tracking glasses while viewing. The recorded gaze points are mapped back to the video, producing a per-frame profile of human attention(Figure 3). On top of this, we construct a series of corrupted videos. For a displacement x and a radius r, let $\widetilde{D}_n(x, r)$ be the video obtained from V by drawing on every frame a circular mask of radius r times the shorter edge of the video. Its center is determined by x: $x = 0$ means centered at human gaze point, and non-zero means shifting the disc by an offset of xw horizontally and xh from the gaze center: w, h means width and height of the video respectively. This way, we obtain a grid of varying (x, r)s and a control of untouched video with $r = 0$. In practice we choose $x \in \{0, 0.1, 0.3, 0.5\}$ and $r \in \{0.1, 0.2, 0.3\}$. The corrupted videos are then given to the models with the same prompts, and the answers are extracted as described in B.2.

3.4 Metrics

Most of our analysis are based on the semantic similarity of two texts T_1 and T_2. A pre-trained language encoder $\mathcal{E}$ is used to map responses into an embedding space. We employ the cosine similarity: $S(T_1, T_2) = \langle \mathcal{E}(T_1), \mathcal{E}(T_2) \rangle / (\|\mathcal{E}(T_1)\| \cdot \|\mathcal{E}(T_2)\|)$,where $\langle \cdot, \cdot \rangle$ denotes the inner product. We use all-MiniLM-L6-v2 [2] as the text encoder. Following Trujillo & Holler (2023); Holler & Levinson (2019), we measure 1) whether the annotator makes early and plausible answers to the entire utterance; 2) whether the annotator revises their answers in the process; and 3) whether human attended cues are important to models with following metrics:

Specificity. We first measure if an author (model/human) is actually providing perceptions specific to the current video V, a failure known as generic responses in dialogue (Li et al., 2016). We employ a similar measure as Wang et al. (2020) to distinguish model answers from generic ones: how far an annotation is distinctive compared to other answers from the same annotator. Specifically, for video V and V' with different clip numbers n and n', we define $\kappa(k, V, V') = \mathrm{round}(kn'/n)$ to be the clip of V' at the progress in the utterance as clip k of V. Following the context in 3.2, we use $D_k(V)$ below to denote the k-th cumulative clip of video V. We measure this on the subset with at least 2 annotators $\mathcal{T}_2$ as mentioned in Appendix A.3:

$$\mathrm{Spec}_Q(A, V, k) = 1 - \frac{1}{|\mathcal{T}_2 \setminus \{V\}|} \sum_{V' \in \mathcal{T}_2 \setminus \{V\}} S\Big(A_Q\big(D_k(V)\big),\, A_Q\big(D^{(V')}_{\kappa(k,V,V')}\big)\Big). \tag{2}$$

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

where $A = H$ or M means either model or human. This comparison is done per perceiver, by averaging over videos V and clip positions k into the final $\mathrm{Spec}_Q(A)$. This score is low for perceivers whose answers are very similar across all videos at similar progress positions, and is high for perceivers who write specific perceptions for each video. Note that natural similarities among videos may influence its value, so we compare on a relative instead of absolute scale for Spec.

Anticipation. We also measure if the model's early answers are similar to human final perception. We employ a ranking approach, aiming to measure how whether an perceiver's partial perception $A_Q(D_k(V))$ is more similar to *this* video's final human perception $H(V)$, than *other* videos' final human perceptions $\{H^a_Q(V')\}_{V' \neq V}$. Specifically,

$$\mathrm{Ant}_Q(A, V, k) = \frac{1}{|\mathcal{A}_V|} \sum_{a \in \mathcal{A}_V} \mathrm{R}\Big(A_Q\big(D_k(V)\big);\ H^a_Q(V),\ \big\{H^a_Q(V')\big\}_{V' \neq V}\Big). \tag{3}$$

where $\mathcal{A}_V$ is the annotator set (for humans, this means other annotators), and R is the ranking measure similarly employed in Hanley & McNeil (1982):

$$\mathrm{R}\big(q;\, t,\, \mathcal{D}\big) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \left(\mathbb{1}\big[S(q,t) > S(q,d)\big] + \frac{1}{2}\, \mathbb{1}\big[S(q,t) = S(q,d)\big] \right) \tag{4}$$

By this design, Ant measures whether models are aligning with annotator answers, and for humans this provides a comparable measure on agreement over annotators. Both Spec and Ant are averaged over the first third of each utterance: $k/n \leq 1/3$.

Revision Types (RT). Gestalt perception also includes actively revising one's own answers. A direct semantic comparison, however, is insufficient to tell the differences, since modifications such as adding a clause, restating a reading, and completely throwing out the previous reading can change an answer similarly in embedding space. To this end, also inspired from the Similarity-Num Steps measure in Social Genome (Mathur et al., 2025), we consider the natural-logic relations (MacCartney & Manning, 2009) with the entailment between adjacent clips. Specifically, we classify each pair of consecutive answers $\big(A_Q(D_{k-1}), A_Q(D_k)\big)$ with a NLI cross-encoder (Liu et al., 2019) which discerns the entailment between two texts in both directions. We use ρ for the word-sequence ratio between the two answers, and consider the relationship classes between adjacent responses:

copy	$\rho \geq \tau_c$: the same content; not sent to the encoder.
restate	the same reading in other words.
elaborate	the reading is kept and something is added.
generalize	the reading is kept and something is dropped or hedged.
contradict	contradiction in either direction: the reading is reversed.
shift	neither entailment nor contradiction: a different reading.

In practice we choose $\tau_c = 0.9$. Details of this classification is provided in Appendix C.3. The first four classes keep the previous reading and the last two replace it, so we call **contradict** and **shift** together a *reading change*. A perceiver that actively reorganizes its interpretation produces many reading changes; one that repeats itself produces many copies.

Gaze Manipulation Similarity (GMS). To compare the masked video perception to original ones, we do the cosine similarity comparison between masked and unmasked videos:

$$\mathrm{GMS}_Q(\widetilde{D}_n(x,r)) = S\Big(M_Q\big(\widetilde{D}_n(x,r)\big),\ M_Q\big(\widetilde{D}_n(0,0)\big)\Big) \tag{5}$$

We run with varying (x, r), and also experimented with and without the audio contents. Model responses M_Q are obtained over every masked/unmasked videos and GMS is computed by averaging.

Terminological caution. We intentionally avoid referring to GMS as saliency measures. This distinction is important because the term "saliency" has been used inconsistently across the literature for heatmaps, salient maps, etc. (Arrieta et al., 2020, Sec. 5.3). Just as Raman et al. (2023) argued in their work about information saliency, perturbation-based methods can create semantically invalid inputs (Adebayo et al., 2018; Lapuschkin et al., 2019) (Molnar, 2020, Sec. 9.5.3.3), making it unclear whether changes in model output reflect the importance of the removed cue or artifacts

introduced by the intervention. Our analysis is designed with these limitations in mind. We build an empirical profile of where human observers look while interpreting multimodal utterances; it is not a ground-truth map of perceptual or causal significance. Likewise, GMS measure how much a model's open-ended judgment changes when human-attended or comparison regions are corrupted. They therefore quantify intervention-based corruption sensitivity, not preattentive or causal necessity. A low GMS indicates that the model is sensitive to corrupting human-attended regions; it should not be interpreted as showing that those regions are intrinsically salient for the model.

4 EXPERIMENTS

We only consider MLLMs which handle both video and audio inputs for Gestalt principles. We choose **Qwen3-Omni-30B (Xu et al., 2025b), Qwen2.5-Omni-7B/3B** (Xu et al., 2025a), **Ming-Lite-Omni** (AI et al., 2025), and **Gemma 4-E4B** (Team et al., 2026) as open-source models, and **Gemini2.5/3.6-flash** (Comanici et al., 2025) as closed-source models. We run open-source models locally and the closed-source models through API.

4.1 TEMPORAL RESULTS: SPECIFICITY, ANTICIPATION, AND REVISION

Temporal results are done with text annotated videos. Spec and Ant (Appendix A.3) with 95% bootstrap intervals over videos. The Spec, Ant, and RT of Section 3.4 separate humans from models both in how specific they perceive and how they revise.

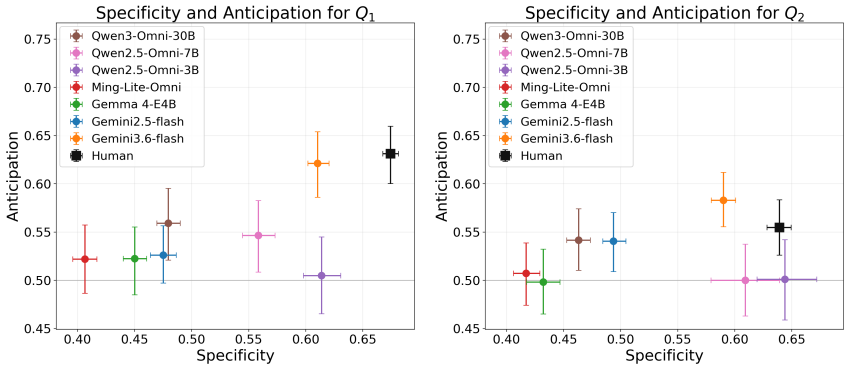


Figure 4: Specificity vs. anticipation over the first third of sub-utterance answers for each video, for Q_1 (left) and Q_2 (right). Anticipation is 0.5 at chance. Bars are 95% bootstrap intervals over videos.

Humans commit early and align more with others's final perception. Figure 4 places each perceiver's early answers (first one third of answers per video) in the (Spec, Ant) plane. Humans sit in the top-right corner on both questions, especially for Q_1. In the first third of an utterance, their readings are already specific to it and agree with where other annotators end up. This aligns wtih Gestalt account: a single coherent reading on incomplete signal. Models fail to reach human performance in general for both questions.

For Q_1, the distinction is clear: all models are not as committed as human, and many models' Ant are including chance level such as Gemma and Ming-Omni-Lite, showing that they are neither specific enough to the current video nor reaching to another human perception. Gemini3.6-flash is closest to human but still has a gap of around 0.07 in Spec. For Q_2, the gap between human and models are smaller. Gemini3.6-flash surpasses human at around 0.03 in Ant but at a cost of 0.05 inferiority in Spec. Qwen2.5-Omni-7B/3B, on the other hand, are reaching human level Spec, but with Ant at chance level. This shows that in general human annotators make predictions early and what they perceive is already in more agreement with other humans, compared to models.

Humans revise perceptions more often. Classification of RT also separates models and humans as shown in figure 5. Annotators replace their reading (shift/contradict) around 55% of the time

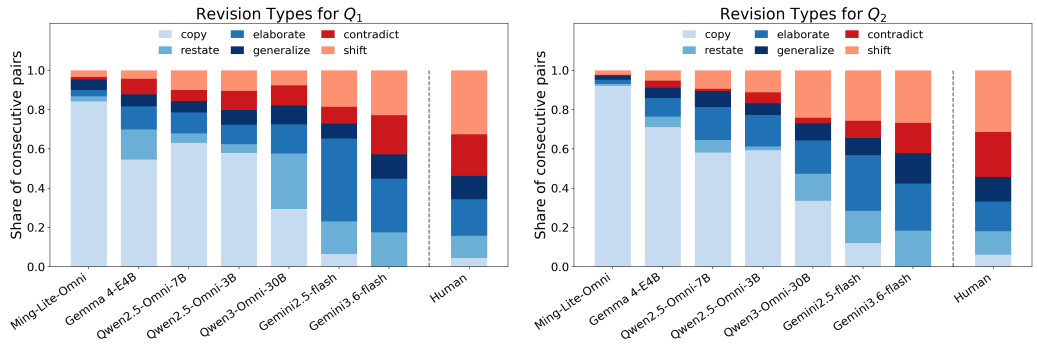


Figure 5: RT classification with NLI-roberta-base model. Humans show notably higher rate of shifting and contradicting than all models, both signs for significant revisions. Many models are occupied with simple copying, and do restating or elaboration more when they do revise.

on both questions, and copy their previous answer very sparingly. The closest model to human is Gemini3.6-flash at around 45% for both questions, still lower than humans by a gap at 10% level. On the other hand, models other than the two Gemini ones all show a much more repetition manner over adjacent responses: their pairs are often dominated instead by **copy**, especially for Ming-Lite-Omni which repeats itself at around 85% and 90% for Q_1 and Q_2, and also includes Qwen2.5-7B/3B and Gemma. In cases where they do move, they **elaborate** or **restate** more often: the previous reading is retained and new words are appended or rephrased around it, so the answer moves while the hypothesis is never at risk. In the Appendix C.3 to C.5 we provide more results in support for our conclusion.

We note that frequent revision and early anticipation are not in tension. Anticipation is a rank that identifies which utterance an early reading is about, whereas a contradiction concerns what is claimed about it: a reading can keep its scene and still reverse its claim. Only a reading specific enough to be wrong can be retracted, so frequent revision is the cost of committing early, which humans do but most models don't.

4.2 Spatial results: gaze manipulation similarity

We compare GMS by $x \in \{0, 0.1, 0.3, 0.5\}$ and $r \in \{0.1, 0.2, 0.3\}$ against the untouched control. Figure 6 shows GMS for Q_1 and Q_2, and figure 7 provides a qualitative illustration.

Perception gaps exists where human-attended regions are corrupted. We observe that corrupting human-attended regions have an impact on model perception, and it is not only due to corruption itself. In almost all cases, GMS increases as offset x increases. This shows that corrupting human attended regions pushes away the final perception more than corrupting other regions with equally. For some models corrupting more area impacts the perception more. Qwen2.5-Omni-7B is representative: the three lines show a clear gap for different rs. However, for some models this effect is not obvious, for example Gemini2.5/3.6-flash, where GMS is nearly flat as (x, r) changes. In C.6 we also report the results without audio input.

GMS Sensitivity is strongly model-dependent. GMS plot pattern over (x, r) is highly influenced by models. The Qwen2.5-Omni models are strongly gaze-sensitive, each spanning roughly 0.2 of the similarity range on the Q_2 grid. Ming-Lite-Omni and the two Gemini models are nearly flat, varying by well under 0.2 across all results. Note that here flatness on the grid can mean either model robustness or barely reading the visual channel. However, combined with the fact that Ming-Lite-Omni copies its previous answer in $80\% - 90\%$ of the time, a flat response is consistent with output that is largely insensitive to input. The Gemini models, on the other hand, copy rarely, have the revision profile closest to humans, so their flat grids are likely showing robustness.

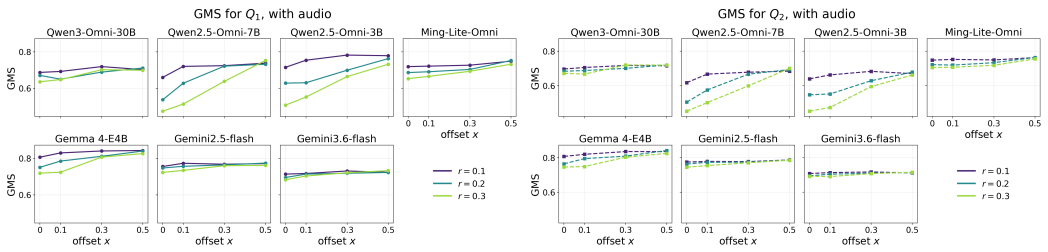


Figure 6: GMS plot for Q_1 and Q_2. Qwen2.5 models has lower GMS with larger r and smaller x, showing they are sensitive to corruption. Gemma shows same pattern for Q_1 but not obvious for Q_2. Other models are not sensitive to gaze based manipulation.

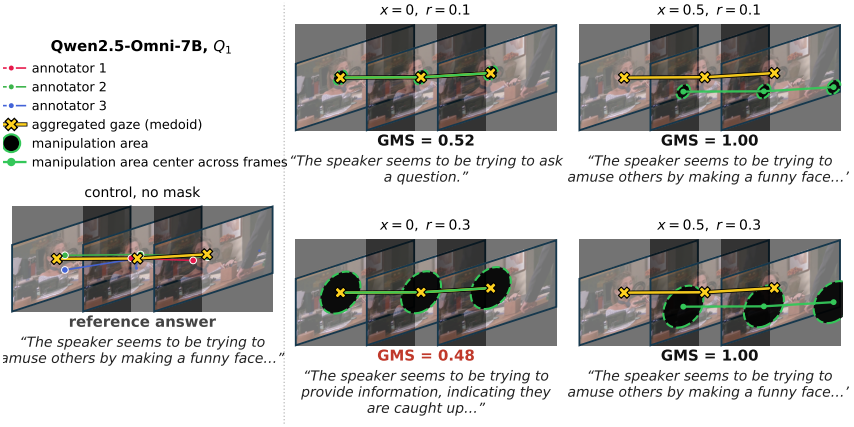


Figure 7: A qualitative illustration of GMS. For Qwen2.5-Omni-7B model on Q_1, $(x, r) = (0, 0.3)$ gives the lowest GMS, while $(0.5, 0.1)$ gives 1 on GMS, showing its sensitivity over (x, r). Video source from Mintrec2. A complete sample of video manipulation and the answers is provided in the supplementary material.

5 Limitations and Conclusion

Limitations. GesBench is limited to short, face-to-face utterances with clearly bounded speakers, drawn from a mixture of scripted and unscripted datasets. Broader settings such as longer conversations, multi-party interactions, or noisier environments, may require additional annotation protocols. Human annotations are collected on a representative subset rather than the full benchmark. Our metrics use sentence-embedding similarity as a scalable proxy for semantic agreement, which may miss subtle pragmatic distinctions between open-ended responses. Finally, human gaze based corruption should be interpreted as a human-grounded corruption-sensitivity test, not as a causal account of human perception or as a saliency estimator.

Conclusion. We introduced GesBench, a human-grounded benchmark for evaluating how multimodal LLMs form social interpretations as utterances unfold. Its temporal pipeline uses cumulative sub-utterance clips to measure whether models could make early predictive perceptions in alignment with human understanding, while its spatial pipeline uses human gaze to test model sensitivity to human-attended visual cues. Across several datasets, we find a consistent human-model gap: humans make more specific, and more aligning perceptions with other humans at early stages when they see the video. They also revise more often without being sparing with shifting or contradicting their previous answers. Finally, models also vary significantly on whether their perception is influenced by human gaze patterns. These results suggest that final-answer accuracy is insufficient for evaluating socially interactive agents; models must also form useful provisional interpretations, revise them incrementally, and ground the judgments in communicatively meaningful cues.

AI DISCLOSURE

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, and interpret results. We did not use generative AI tools to clean or reformat datasets. The generation of synthetic datasets, development of theoretical models or conceptual frameworks, formulation of mathematical claims, provision of critical ingredients for proving mathematical claims, assistance in writing proofs, and translation were not applicable to this work. We did not use generative AI tools to propose or refine hypotheses.

Additionally, we used generative AI tools to draft parts of the research paper and to summarize or analyze existing literature. We reviewed all AI-assisted work. All literature suggested by generative AI tools was manually verified to ensure that the cited sources exist and that they were cited and summarized accurately in support of our narrative and claims. AI-assisted methodological suggestions and method implementations were reviewed by the authors for appropriateness and correctness. AI-assisted qualitative and thematic analyses were checked against the underlying research materials, and AI-assisted interpretations of results were manually verified against the reported results, figures, and tables. We also manually checked all reported numerical values and verified that figures and tables were interpreted correctly. LLM-generated research ideas were checked for potential overlap with existing work through a manual literature survey. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICAL STATEMENT

In the data collection procedure of GESBENCH, participants were informed about the study content, and required to sign a consent form, as mentioned in Appendix A.2. There were informed of our measures to store, process, and use their recorded data. No dangerous content or measures appear or is exposed in any of the annotation tasks.

REPRODUCIBILITY STATEMENT

To support reproducibility, we open-source our codes, including data curation procedure, model prompting strategies, and metric analysis, as mentioned in the abstract. We also include a sample in the supplementary material to illustrate the gaze manipulation mechanism. Due to the volume of the data and the double-blind reviewing policy, we will include links to human annotation data upon publication.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

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.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al.

Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

Monika Bednarek. *Salient Features of TV Dialogue: A Corpus Linguistic Approach*, pp. 121–154. Cambridge Applied Linguistics. Cambridge University Press, 2018.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Isaac R Christian, Udith Haputhanthrige, Hanna Hornfeld, Declan Campbell, Samuel Nastase, Taylor Webb, and Michael Graziano. Attention alignment between humans and vision-language models. *arXiv preprint arXiv:2606.17410*, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Kailas Dayanandan, Nikhil Kumar, Anand Sinha, and Brejesh Lall. Dual thinking and logical processing–are multi-modal large language models closing the gap with human vision? *arXiv preprint arXiv:2406.06967*, 2024.

Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5724–5733, 2019.

Siqi Fan, Olga Dal Monte, and Steve WC Chang. Levels of naturalism in social neuroscience research. *IScience*, 24(7), 2021.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the fourteenth conference on computational natural language learning–Shared task*, pp. 1–12, 2010.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

François Grosjean. Spoken word recognition processes and the gating paradigm. *Perception & psychophysics*, 28(4):267–283, 1980.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Judith Holler and Stephen C Levinson. Multimodal language processing in human communication. *Trends in cognitive sciences*, 23(8):639–652, 2019.

Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. Prediction of next-utterance timing using head movement in multi-party meetings. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pp. 181–187, 2017.

Iiro P Jääskeläinen, Mikko Sams, Enrico Glerean, and Jyrki Ahveninen. Movies and narratives as naturalistic stimuli in neuroimaging. *NeuroImage*, 224:117445, 2021.

Junfeng Jiao, Saleh Afroogh, Abhejay Murali, Kevin Chen, David Atkinson, and Amit Dhurandhar. Llm ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. *Scientific Reports*, 15(1):34642, 2025.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. Towards incremental transformers: An empirical analysis of transformer models for incremental nlu. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1178–1189, 2021.

Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

Caixin Kang, Yifei Huang, Liangyang Ouyang, Mingfang Zhang, and Yoichi Sato. Can mllms read the room? a multimodal benchmark for verifying truthfulness in multi-party social interactions. *arXiv preprint arXiv:2510.27195*, 2025.

Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C Mozer. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 4(3):251–263, 2021.

Kurt Koffka. *Principles of Gestalt psychology*. routledge, 2013.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 110–119, 2016.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. Affectgpt: Dataset and framework for explainable multimodal emotion recognition. *arXiv preprint arXiv:2407.07653*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL `https://arxiv.org/abs/2307.06281`.

Bill MacCartney and Christopher D Manning. An extended model of natural logic. In *Proceedings of the eight international conference on computational semantics*, pp. 140–156, 2009.

Brielen Madureira and David Schlangen. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental nlu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 357–374, 2020.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. *arXiv preprint arXiv:2502.15109*, 2025.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Anupam Pani and Yanchao Yang. Gaze-vlm: Bridging gaze and vlms through attention regularization for egocentric understanding. *Advances in Neural Information Processing Systems*, 38: 163544–163577, 2025.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38, 2025.

Martin J Pickering and Chiara Gambi. Predicting while comprehending language: A theory and review. *Psychological bulletin*, 144(10):1002, 2018.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Zheng Qi, Chao Shang, Evangelia Spiliopoulou, and Nikolaos Pappas. Capturing gaze shifts for guidance: Cross-modal fusion enhancement for vlm hallucination mitigation. *arXiv preprint arXiv:2510.22067*, 2025.

Paulo Quaglio. Television dialogue and natural conversation: Linguistic similarities and functional differences. In *Corpora and discourse: The challenges of different settings*, pp. 189–210. John Benjamins Publishing Company, 2008.

Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Laura Cabrera-Quiros, and Hayley Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. In *Understanding social behavior in dyadic and small group interactions*, pp. 265–293. PMLR, 2022.

Chirag Raman, Alec Nonnemaker, Amelia Villegas-Morcillo, Hayley Hung, and Marco Loog. Why did this model forecast this future? information-theoretic saliency for counterfactual explanations of probabilistic regression models. *Advances in Neural Information Processing Systems*, 36: 33222–33240, 2023.

Elizabeth Redcay and Dustin Moraczewski. Social cognition in context: A naturalistic imaging approach. *NeuroImage*, 216:116392, 2020.

Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Björn Þór Jónsson, Onanong Kongmeesub, Hoang-Bao Le, Stevan Rudinac, Klaus Schöffmann, Florian Spiess, et al. The castle 2024 dataset: Advancing the art of multimodal understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12629–12635, 2025.

David Schlangen and Gabriel Skantze. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2:83–111, 2011.

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

Jingyuan Sha, Hikaru Shindo, Kristian Kersting, and Devendra Singh Dhami. Gestalt vision: A dataset for evaluating gestalt principles in visual perception. In *19th International Conference on Neurosymbolic Learning and Reasoning*, 2016.

Shihab A Shamma, Mounya Elhilali, and Christophe Micheyl. Temporal coherence and attention in auditory scene analysis. *Trends in neurosciences*, 34(3):114–123, 2011.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature human behaviour*, 8(7):1285–1295, 2024.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang,

Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Pia Tikka, Mauri Kaipainen, and Juha Salmi. Narrative simulation of social experiences in naturalistic context–a neurocinematic approach. *Neuropsychologia*, 188:108654, 2023.

James P Trujillo and Judith Holler. Interactionally embedded gestalt principles of multimodal human communication. *Perspectives on Psychological Science*, 18(5):1136–1159, 2023.

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. Time blindness: Why video-language models can't see what humans can? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30906–30918, 2026.

S Stavros Valenti and James MM Gold. Social affordances and interaction i: Introduction. *Ecological Psychology*, 3(2):77–98, 1991.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Compare and reweight: Distinctive image captioning using similar images sets. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2020.

Yusen Wu, Junwu Xiong, and Xiaotie Deng. How social is it? a benchmark for llms' capabilities in multi-user multi-turn social agent tasks. *arXiv preprint arXiv:2505.04628*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

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

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel Understanding with Visual Instruction Tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28202–28211, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02664. URL https://ieeexplore.ieee.org/document/10656218/.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Wenrui Li, Yanting Chen, et al. Mintrec2. 0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. *arXiv preprint arXiv:2403.10943*, 2024.

Hanlei Zhang, Zhuohang Li, Yeshuang Zhu, Hua Xu, Peiwu Wang, Haige Zhu, Jie Zhou, and Jinchao Zhang. Can large language models help multimodal language analysis? mmla: A comprehensive benchmark. *arXiv preprint arXiv:2504.16427*, 2025a.

Ruoxuan Zhang, Qiyun Zheng, Zhiyu Zhou, Ziqi Liao, Siyu Wu, Jian-Yu Jiang-Lin, Bin Wen, Hongxia Xie, Jianlong Fu, and Wen-Huang Cheng. Mindpower: Enabling theory-of-mind reasoning in vlm-based embodied agents. *arXiv preprint arXiv:2511.23055*, 2025b.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest, 2025c. URL https://arxiv.org/abs/2307.03601.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

Appendices

A Data Curation and details

A.1 Dataset selection rationale

We selected source datasets according to four criteria. First, the data should contain face-to-face conversational utterances in which social meaning is expressed through multimodal cues such as speech, facial expression, gaze, etc. Second, utterance boundaries should be sufficiently clear to support cumulative sub-utterance clipping. Third, the data should include short-to-medium utterances, since very long videos make dense human annotation and eye-tracking collection less practical. Fourth, the benchmark should include both controlled and naturalistic interaction settings.

MELD (Poria et al., 2019) (GPL3.0 license) and MIntRec2 (Zhang et al., 2024) (CC-BY-NC-SA-4.0 license) provide scripted utterances from television dialogue. Although scripted, these clips retain many perceptual demands of everyday social understanding: viewers must integrate lexical, prosodic, facial, gestural, interpersonal, and turn-taking cues to infer speaker intention and possible responses. Prior work also suggests that television dialogue shares important linguistic properties with spontaneous conversation (Quaglio, 2008; Bednarek, 2018), while offering cleaner utterance boundaries and greater accessibility for annotation. Also, simulated stimuli has been used for studying human responses in social interactions in neuro-science (Redcay & Moraczewski, 2020). This aligns with our goal, as gestalt principles are also highly based on findings from neuro-science side (Koffka, 2013; Shamma et al., 2011). We deem that the task of social signal perception is of the similar end for what simulated signals could afford. In fact, these controlled stimuli are frequently used for simulated narratives (Tikka et al., 2023; Jääskeläinen et al., 2021) and there are also arguments that they offer good balance by having controlled scenarios as well as decent level of ecological validity (Fan et al., 2021). Seamless Interaction (Agrawal et al., 2025)(CC-BY-NC-SA-4.0 license) also complements these sources with unscripted dyadic conversations, providing more natural face-to-face interaction. We therefore combine scripted and unscripted sources to balance cue diversity, annotation feasibility, and ecological validity.

We do not include in-the-wild activity-video datasets as primary sources because many clips contain weakly bounded utterances, unclear speaker roles, fewer face-to-face exchanges, or audio/visual conditions that make sub-utterance annotation difficult (Grauman et al., 2022; Rossetto et al., 2025). Future versions of GesBench could extend the protocol to longer, noisier, multi-party, and more culturally diverse interactions.

A.2 Annotation setup

GesBench records two types of human data: (1) human answers for CVCs: $H_Q(D_1), \ldots, H_Q(D_n)$, and (2) eye-gaze recordings. Both types of responses are recorded by the Covfee tool (Quiros et al., 2022). It provides clear and smooth setup process of annotation tasks. Figure 8 shows a sample screenshot for monitoring the progress of the annotation tasks.

A.2.1 Human text annotations

We employ a crowd sourcing approach for collecting human text annotations on CVC. We utilize the Prolific [3], and provided them with a link to the Covfee server which is set up locally. Every annotator sign a consent form before taking part in. During the task, we asked the annotators to provide the following for both Q_1 and Q_2:

- Their perception and reasons over the current clip D_k, by speaking it out and audio recorded;
- Their current confidence, with a scale from 1-5;
- Whether their perception has changed compared to the previous clip (D_{k-1}, only applies to clips with $k \geq 2$)

[3] https://app.prolific.com/

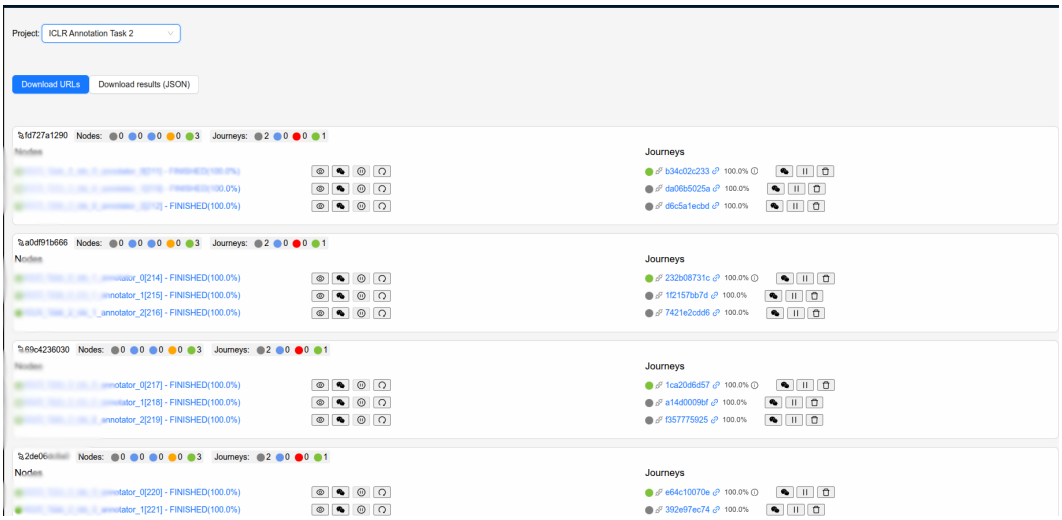


Figure 8: Illustration of the Covfee main page for monitoring the annotation tasks.

and a screenshot of the Covfee page for the annotators is shown in figure 9. After collection, their audio are transcribed into text using WhisperX (Bain et al., 2023) model. In practice the implementation is done with fast WhisterX [4].

A.2.2 Human Eye Gaze Data

To collect annotators' eye focus when they watch a video, we utilize the Pupil Core Eye Tracking Glasses [5]. It takes videos for people's eyeballs and then infers their eye gazes in the world camera. If a surface is provided and detected in the scene, it can automatically calibrate and calculate relative positions of eye gaze on the specific surface. Since we also have fixed region of the video display area on the screen, we can then infer the eye gaze position within the video region. Annotators sign a consent form informing them about the fact that their eye gaze data will be processed and published before they annotate.

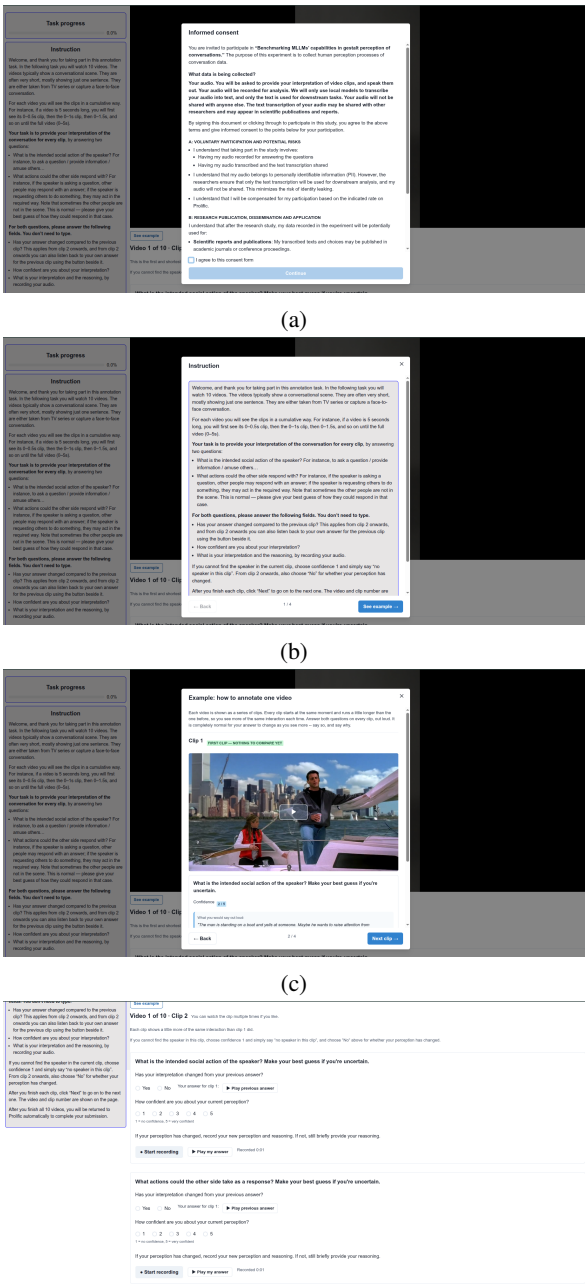
We use AprilTags provided on the official site of Pupil [6] to mark the computer screen in the scene of the world camera in Pupil Core. All annotators did a calibration session [7] before they started the experiment. During the experiment, they were asked to think of two questions, Q_1 and Q_2, during watching the videos, and then they move on to the next video. There is no video cutting in this experiment, as only continuous gazing during watching an entire video would enable us to construct a per-frame gaze-position on the videos. An example layout of one annotation page is provided in figure 10.

A.3 Dataset Details

We show the full details of our dataset in table 1. $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ represent the subset with *at least* $1/2/3$ text annotators respectively, so they are subsuming instead of exclusive. $\mathcal{S}$ denotes the set with gaze annotations, all clips have 3 annotators.

In the experiments, Spec and Ant are measured on $\mathcal{T}_2$, since the Ant needs at least two annotators to compute human value; RT is done on $\mathcal{T}_1$. GMS is measured on $\mathcal{S}$. Due to API costs, both Gemini models are prompted on only $\mathcal{T}_1$ and $\mathcal{S}$.

[4] https://github.com/SYSTRAN/faster-whisper
[5] https://pupil-labs.com/products/core
[6] https://docs.pupil-labs.com/core/software/pupil-capture/#surface-tracking
[7] https://docs.pupil-labs.com/core/software/pupil-capture/#calibration-process



(a)

(b)

(c)

(d)

Figure 9: Sample of paid annotators user interface on Covfee when doing the text annotation. They are first asked to sign the consent form to procees(9a), then go through the instruction(9b) and example(9c). 9d shows the user actual panel they see during annotation: specifying whether perception has changed, choosing confidence, and recording audios about their perception and reasoning.

B Experimental Setup

B.1 Model prompting

In this subsection we provide details for configuring and prompting the models. For all models, we employ a zero-shot strategy, and each video clip in CVC and corrupted video sets are prompted

	Full		$\mathcal{T}_1$		$\mathcal{T}_2$		$\mathcal{T}_3$		$\mathcal{S}$	
Dataset	# utt.	AL (s)	# utt.	AL	# utt.	AL	# utt.	AL	# utt.	AL
MELD	1,870	3.273	216	3.229	141	3.284	8	2.938	219	3.242
MIntRec2	2,974	2.968	215	2.928	143	2.871	8	2.500	219	2.922
Seamless	1,364	4.117	159	4.157	105	4.171	4	5.000	162	4.154
Overall	6,208	3.312	590	3.369	389	3.371	20	3.175	600	3.372

Table 1: GESBENCH data details. Seamless means Seamless Interaction. AL means average utterance length.



Figure 10: Example of the annotator task page for providing gaze annotations. The left is an instruction panel, and the right shows the video. The annotators are asked to think of the two questions (Q_1, Q_2) during watching, and they click "Submit Annotation" button to proceed to the next video. The scrolling bar is intentionally disabled for this task to ensure that the video displaying region stays fixed within the screen.

separately by launching a new conversation. all models, the user inputs within the chat template are as follows:

GESBENCH PROMPT (full, for CVC clips $\{D_k\}, k = 1$)

Watch the following video. You will see it in a cumulative way: all clips are cut from the same single source video. For instance, if a video is 5 seconds long, you will first see its 0-0.5s clip, then the 0-1s clip, then 0-1.5s, and so on until the full video (0-5s).

The current and first video clip is given in the media. Answer the following two questions:

Q1: What is the intended social action of the speaker? For instance, to ask a question / provide information / amuse others, etc.
- State your interpretation.
- How confident are you about your interpretation? Answer with a number in the scale of 1-5. 1 is no confidence and 5 is very confident.

Q2: What actions could the other side respond with? For instance, if the speaker is asking a question, other people may respond with an answer; if the speaker is requesting others to do something, they may act in the required way.

```
- State your interpretation.
- How confident are you about your interpretation? Answer with a
↪  number in the scale of 1-5. 1 is no confidence and 5 is very
↪  confident.
```

GESBENCH PROMPT (full, for CVC clips $\{D_k\}, k \geq 2$)

```
This is the <clip_ordinal> clip, cut from the same source video as
↪  the previous clips. Your answer for previous clip was:
↪  <previous_answer>
The current video clip is given in the media. Regarding current
↪  clip, still think of the two questions:

Q1: What is the intended social action of the speaker? For instance,
↪  to ask a question / provide information / amuse others, etc.
Q2: What actions could the other side respond with? For instance, if
↪  the speaker is asking a question, other people may respond with
↪  an answer; if the speaker is requesting others to do something,
↪  they may act in the required way.

For each of Q1 and Q2, answer the following questions separately:
- Has your perception changed compared to previous clip? Say
↪  explicitly "yes" or "no".
- If yes, state your new perception, and provide reasons. If no,
↪  also provide your reasons.
- How confident are you about your interpretation now? Answer with a
↪  number in the scale of 1-5. 1 is no confidence and 5 is very
↪  confident.
```

where <previous_answer> tag is for the model's answer at the previous round. The prompt is done in a multi-turn way, ensuring that models have the full context of their answer previously and in the same way as humans, since human would naturally remember what they have just seen.

Below we list some necessary parameters used when prompting:

Qwen2.5-Omni-7B/3B. We follow the official chat template provided on HuggingFace [8] [9], where it includes fields for video and audio inputs. The model weights and configurations are downloaded from the above HuggingFace repositories respectively. All of the model inference is done on Intel Xeon Gold 6448Y 32C 2.1GHz CPU and NVIDIA A100 GPU(80G memory), with CUDA version 12.6.

Qwen3Omni30B. We follow the official chat template provided on HuggingFace [10]. The model weights and configurations are downloaded from the above HuggingFace repositories. All of the model inference is done on AMD EPYC 9334 2.7 GHz 210W and NVIDIA H100 GPU(80G memory), with CUDA version 12.8 and PyTorch version 2.7.1.

Ming-Lite-Omni. We follow the official chat template provided on HuggingFace [11]. The model weights and configurations are downloaded from HuggingFace repository as above. The inference for CVC is done on Intel Xeon Gold 6448Y 32C 2.1GHz CPU and NVIDIA A100 GPU, with CUDA version 12.8, and the inference for GMS is done on Intel Xeon Gold 5128 CPU and NVIDIA RTX 6000 Pro GPU(96G memory), with CUDA version 12.8.

Gemma 4. We follow the chat guidlines provided by Google [12]. We download the models weights and configurations from HuggingFace [13]. Max-frame is set to 32 and thinking is disabled. All of

[8] https://huggingface.co/Qwen/Qwen2.5-Omni-7B
[9] https://huggingface.co/Qwen/Qwen2.5-Omni-3B
[10] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct
[11] https://huggingface.co/inclusionAI/Ming-Lite-Omni
[12] https://ai.google.dev/gemma/docs/core
[13] https://huggingface.co/google/gemma-4-E4B

the model inference is done on AMD EPYC 7413 CPU and NVIDIA A40 GPU(48G memory), with CUDA version 12.6.

Gemini2.5/3.6-Flash. We implement according the official documentation provided [14]. The evaluations are done through API calls where we request batch processing.

B.2 Answer Postprocessing

We postprocess the answers from both the models and humans.

Model answer extraction. For models, we employ a multi-turn conversation manner: prompting them with a media and their previous answers. This is intended to match the human experiment environment, where humans can naturally remember what they have actually seen. Since humans are answering both Q_1 and Q_2 under the same context, their answers maybe naturally associated. To ensure maximum comparability, we also prompt the models with the two question combined (see B.1) instead of separately, and then use another LLM to extract both fields. We asked the model to extract both fields verbatim as much as possible and not adding, deleting, summarizing, or modifying the content. Specifically, we leverage OpenAI's GPT-4o-Mini [15] by API call, with the following prompt for clips of $k = 1$ and $k \geq 2$ respectively:

GESBENCH ESTRACTION PROMPT for CVC clips $\{D_k\}, k = 1$

```
You are extracting structured fields from an answer another model
↪  gave about the FIRST clip of a video. That model was shown the
↪  video cumulatively -- clip 1 is the opening fragment -- and was
↪  asked two questions about it:

Q1 (intention): What is the intended social action of the speaker?
Q2 (response): What actions could the other side respond with?

The user message is that model's answer, verbatim. Report what it
↪  said about each question, as JSON.

- "changed" is always null here. This is the first clip, so there is
↪  nothing for a perception to have changed from.
- "confidence" is the 1-5 number the model gave for that question,
↪  as a number. Use null if it gave none for that question.
- "perception" is everything the model wrote in answer to that
↪  question, copied verbatim: the same words in the same order,
↪  including its perception, reasons, hedges and alternative
↪  readings. Do not paraphrase, summarise, shorten, correct or
↪  reorder it, and do not add anything of your own. Leave out only
↪  the question label or a restatement of the question ("Q1:",
↪  "Intended social action:") and the confidence rating itself. If
↪  one passage answers both questions together and cannot be split,
↪  copy it into both. Use "" if it said nothing about that
↪  question.
- "insufficient_information" is true only when the model declines to
↪  answer that question outright -- "there is not enough
↪  information", "I don't know", "it is impossible to say from this
↪  clip" -- and offers no interpretation alongside saying so.
↪  Record that statement as the "perception" when it happens, so
↪  that declining to answer is never confused with an answer nobody
↪  extracted. A tentative, hedged or low-confidence guess is still
↪  a guess: false, however unsure the model sounds. An answer that
↪  simply never mentions the question is false as well -- saying
↪  nothing is not the same as saying it cannot be known.
```

[14] https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash

[15] https://developers.openai.com/api/docs/models/gpt-4o-mini

```
- Report only what is in the answer. If the answer is empty, is an
↪  error message, or does not address one of the questions, use
↪  null and "" for that question rather than guessing.
- The two questions are usually labelled Q1 and Q2, but not always.
↪  Match on what is being answered -- the speaker's own intent goes
↪  in "intention", what someone else might do in reply goes in
↪  "response" -- rather than on the labels.
```

GESBENCH ESTRACTION PROMPT for CVC clips $\{D_k\}, k \geq 2$

```
You are extracting structured fields from an answer another model
↪  gave about a later clip of a video. That model was shown the
↪  video cumulatively, had already answered about the previous
↪  clip, and was asked, for each of two questions, whether its
↪  perception had changed, what its perception is now, and how
↪  confident it is:

Q1 (intention): What is the intended social action of the speaker?
Q2 (response): What actions could the other side respond with?

The user message is that model's answer, verbatim. Report what it
↪  said about each question, as JSON.

- "changed" is the model's own explicit "yes" or "no" for that
↪  question, lowercased. If it does not say so in those words, take
↪  it from whether it describes its perception as revised or as
↪  unchanged. Use null only when the answer does not address that
↪  question at all.
- "confidence" is the 1-5 number the model gave for that question,
↪  as a number. Use null if it gave none for that question.
- "perception" is everything the model wrote about that question for
↪  this clip, copied verbatim: the same words in the same order,
↪  including its perception, reasons, hedges and alternative
↪  readings. Do not paraphrase, summarise, shorten, correct or
↪  reorder it, and do not add anything of your own. When it says
↪  its perception has not changed, copy what it wrote anyway -- the
↪  restated interpretation and the reasons alike. Leave out only
↪  the question label or a restatement of the question ("Q1:",
↪  "Intended social action:"), the confidence rating itself, and a
↪  bare "yes" or "no" that does nothing but answer whether the
↪  perception changed. If one passage answers both questions
↪  together and cannot be split, copy it into both. Use "" only if
↪  it wrote nothing about that question.
- "insufficient_information" is true only when the model declines to
↪  answer that question outright -- "there is still not enough
↪  information", "I don't know", "it is impossible to say from this
↪  clip" -- and offers no interpretation alongside saying so.
↪  Record that statement as the "perception" when it happens, so
↪  that declining to answer is never confused with an answer nobody
↪  extracted. A tentative, hedged or low-confidence guess is still
↪  a guess: false, however unsure the model sounds. An answer that
↪  simply never mentions the question is false as well -- saying
↪  nothing is not the same as saying it cannot be known.
- Report only what is in the answer. If the answer is empty, is an
↪  error message, or does not address one of the questions, use
↪  null and "" for that question rather than guessing.
- The two questions are usually labelled Q1 and Q2, but not always.
↪  Match on what is being answered -- the speaker's own intent goes
↪  in "intention", what someone else might do in reply goes in
↪  "response" -- rather than on the labels.
```

In the GMS tests, since we are comparing perceptions, we ask the extraction model to drop reasoning and just kept the perception itself. The difference between extraction prompt does not invalidate

GMS test as a model's answers are always compared to its own, and no human answers is involved in this process. Prompt is as below:

GESBENCH ESTRACTION PROMPT for GMS test $\widetilde{D}_n(x, r)$

```
You are extracting structured fields from an answer another model
↪  gave about a video. That model was shown the whole clip once and
↪  was asked two questions about it:

Q1 (intention): What is the intended social action of the speaker?
Q2 (response): What actions could the other side respond with?

The user message is that model's answer, verbatim. Report what it
↪  said about each question, as JSON.

- "confidence" is the 1-5 number the model gave for that question,
↪  as a number. Use null if it gave none for that question.
- "perception" is the model's own interpretation for that question,
↪  in its words or closely paraphrased. Do not add your own reading
↪  of the video, do not compress it to a one-word label, and leave
↪  out its supporting reasons unless the interpretation cannot be
↪  stated without them. Use "" if it said nothing about that
↪  question.
- "insufficient_information" is true only when the model declines to
↪  answer that question outright -- "there is not enough
↪  information", "I don't know", "it is impossible to say from this
↪  video" -- and offers no interpretation alongside saying so.
↪  Record that statement as the "perception" when it happens, so
↪  that declining to answer is never confused with an answer nobody
↪  extracted. A tentative, hedged or low-confidence guess is still
↪  a guess: false, however unsure the model sounds. An answer that
↪  simply never mentions the question is false as well -- saying
↪  nothing is not the same as saying it cannot be known.
- Report only what is in the answer. If the answer is empty, is an
↪  error message, or does not address one of the questions, use
↪  null and "" for that question rather than guessing.
- The two questions are usually labelled Q1 and Q2, but not always.
↪  Match on what is being answered -- the speaker's own intent goes
↪  in "intention", what someone else might do in reply goes in
↪  "response" -- rather than on the labels.
```

Human answer processing Human answers are naturally separated in the two fields so they don't experience this extraction process. With the transcribed human text answer for both questions, we employ a *carry-forward* operation. Sometimes humans say "no, nothing has changed" or similar words when their perception did not change, and in this case it can be misleading as the pure linguistic meaning is different from the true perception. We detect this pattern and carry their previous answers forward to replace the explicit indication that their perception did not change in such cases.

We argue, however, that this operation does not significantly impact our methodologies. For Spec and Ant, the answers are compared to either the same perceiver's own ones at different stages, or another human annotator's final answer. This also makes RT a conservative comparison instead of an inflated one: carry-forward is actually making *copy* patterns for humans, and humans are still having larger proportion of shift/contradict and very few copies even after it.

B.3 Gaze manipulation details

Every video in $\mathcal{S}$ has gaze from 3 annotators, recorded with Pupil Core. Samples with confidence below 0.6 or that fall outside the video area are dropped. For each frame of a clip, each annotator's gaze position is the per-axis median of their samples within $\pm 0.1\,\mathrm{s}$ of the frame time; if there are none, their nearest sample within $0.5\,\mathrm{s}$ is used. The frame's gaze point is the *medoid* of the annotators' positions, i.e. the one with the smallest summed distance to the others, so it is always a point an annotator actually looked at. Frames on which no annotator has a position are not masked.

The disc has radius $r \cdot \min(w, h)$. For $x > 0$, the offset direction is fixed per clip: we count which quadrant the gaze point falls in over all frames and move the disc towards the opposite corner (gaze mostly top-right is shifted to the bottom-left, and so on), by xw horizontally and xh vertically. If the shifted center stays inside the frame, a disc that crosses the frame edge is cut off there. If it would leave the frame along one axis, the shift along that axis is reversed, placing the disc on the other side of the gaze point.

C Additional results

C.1 Extraction fidelity

Model	Answers	Coverage	Exact	Cov. <0.9
Qwen3-Omni-30B	72735	0.999	0.85	0.00
Qwen2.5-Omni-7B	44656	1.000	0.99	0.00
Qwen2.5-Omni-3B	40659	0.955	0.89	0.05
Ming-Lite-Omni	80142	0.999	0.99	0.00
Gemma 4-E4B	80230	1.000	1.00	0.00
Gemini2.5-flash	11506	1.000	0.96	0.00
Gemini3.6-flash	11406	1.000	0.96	0.00

Table 2: Extraction fidelity, with Q_1 and Q_2 answers pooled. Coverage: mean share of an extracted answer's words found, in order, in the model's reply. Exact: share of answers that are one unbroken stretch of the reply. Cov. <0.9: share of answers with coverage below 0.9.

To verify that our extraction strategy keeps loyalty to the original content, we do a fidelity check on the raw answers and extracted ones, and show the statistics in table 2. We conclude that the extractor copies rather than rewrites. Mean coverage is at least 0.999 for six of the seven models, and at least 85% of answers are exact.

C.2 Early commitment profile

Human comprehension is incremental (Schlangen & Skantze, 2011) and predictive (Pickering & Gambi, 2018): listeners commit to an interpretation before the evidence licenses it (Holler & Levinson, 2019). To further understand the nature in early responses for models and human, we tag every answer with five non-exclusive features by regular-expression matching on the lowercased text, whole words only, using the keywords below:

undecided: *"can't / cannot / unable to / not able to / impossible to / not possible to / hard to / difficult to / too early to"* followed, up to two words later, by *"tell / say / determine / know / judge / interpret / infer / see"*; or *"not sure, unsure, no idea, don't know, do not know, unclear, not clear, insufficient, not enough, nothing yet, nothing to go on"*;

hedged: Referring to hedge detection (Lakoff, 1973; Farkas et al., 2010), we choose *"might, may, could, possibly, perhaps, maybe, probably, likely, seem(s), appear(s), suggest(s), potentially, guess"*;

alternatives: *"or, either, such as, whether"*;

speech: *"saying / says / said / tells / telling / told / replies / replying"*, *"is / was / are / were / 's / 're like,"*, or a quotation mark. It only counts in a clause (split on *". ! ? ;"* or a line break) that contains no prediction marker, so "he will say 'get out' " counts as prediction, not speech;

prediction: *"will / 'll / won't / shall / gonna"*; *"going to / about to / set to"* followed by a word other than an article or pronoun (so not "going to the door"); *"expect(s) / expecting / anticipate(s) / anticipating / predict(s) / predicting"*; *"soon / shortly / afterward(s)"*; and *"next"* followed by *"he / she / they / the"*.

We define $f\big(A_Q(D_k)\big)$ to represent whether $A_Q(D_k)$ belongs to category f, with $A = H$ or M to represent either human or model answer. Let $\mathcal{K}_\delta(V) = \{k : k/n(V) \leq \delta\}$ be the *early window* of

an utterance. The early rate is

$$\mathrm{Rate}_f(A, V) = \frac{1}{|\mathcal{K}_\delta(V)|} \sum_{k \in \mathcal{K}_\delta(V)} \mathbb{1}\Big[f\Big(A_Q\big(D_k(V)\big)\Big)\Big]. \tag{6}$$

We compute it on two texts: the *extracted* answers used everywhere else, and the *unextracted* text, which is the model's stored reply as generated and the annotator's two answers joined.

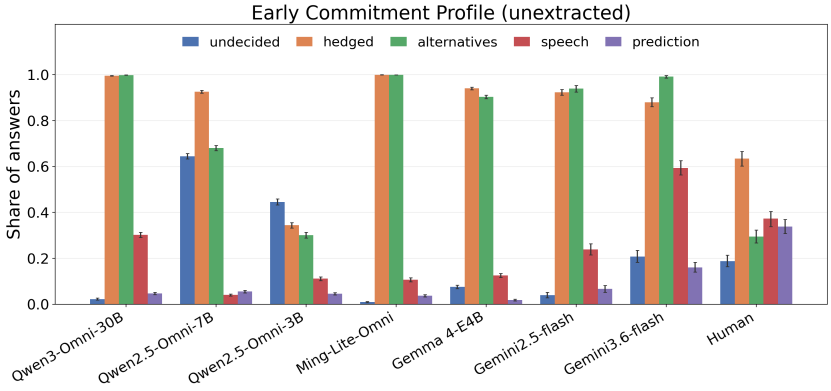


Figure 11: Comparison of early commintment type over categories, with raw texts from models without extraction. Humans are the highest in mentioning predictions, with most models providing hedges or saying there is insufficient information.

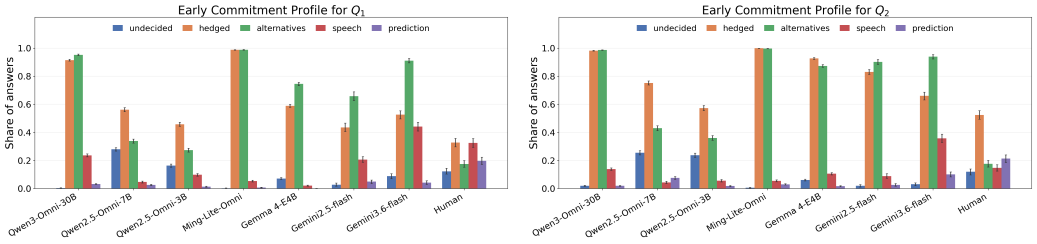


Figure 12: Comparison of early commintment type over categories, on extracted texts for models with both questions. Humans are the highest in mentioning predictions, with most models providing hedges or saying there is insufficient information.

Figure 11 shows the unextracted profile, computed on **full set** when applicable (local models), and on $\mathcal{T}_1$ for Gemini models and humans. Annotators predict what comes next in around 35% of early answers; no model exceeds 18% (Gemini3.6-flash being max), and the other six stay below 10%. Models instead hold readings open: Ming-Lite-Omni and Qwen3-Omni-30B offer alternatives and hedge in almost every early answer, and Gemini2.5/3.6-flash and Gemma offer alternatives in 90% to 99%, against around 30% for annotators. The Qwen2.5-Omni models fail differently: they are undecided in around 50% (3B) and 60% (7B) of early answers against 20% for annotators. The per-question profiles on the extracted text (Figure 12) show the same ordering: above 20% human predictions on Q_1 against at most 6% for any model. Both behaviors run against the Gestalt account: a perceiver who lists the possibilities has not closed on a reading, and one who declines has not begun.

C.3 Revision classification details

We use the pre-trained cross-encoder `nli-roberta-base` (Liu et al., 2019) without fine-tuning. For each pair it outputs one of entailment, neutral and contradiction by argmax over three classes. We run it twice, with the earlier answer as premise ($\ell_{\rightarrow}$) and with the later one ($\ell_{\leftarrow}$). Pairs with

$\rho \geq 0.9$ are **copy**. Otherwise, the pair is **contradict** if either direction is contradiction, **restate** if both are entailment, **elaborate** if only $\ell_{\leftarrow}$ is entailment, **generalize** if only $\ell_{\rightarrow}$ is, and **shift** otherwise.

C.4 Answer length

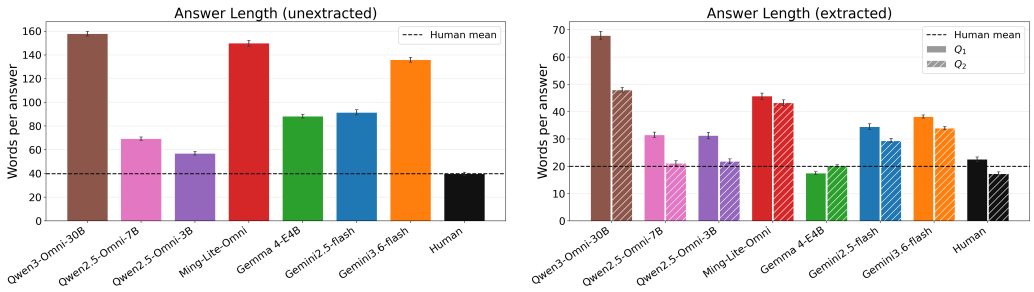


Figure 13: Comparison of answer lengths before and after extraction done on $\mathcal{T}_1$. The left part shows original answers while the right part shows after extraction.

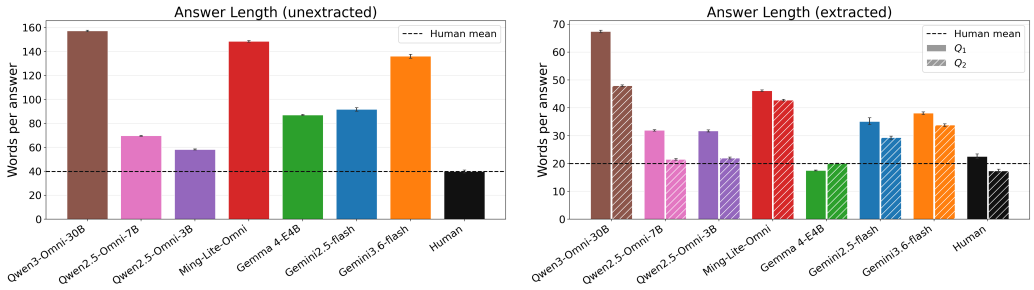


Figure 14: Comparison of answer lengths before and after extraction done on **full set** when applicable. The left part shows original answers while the right part shows after extraction.

To examing the loyalty of GPT extraction, also ablating on the possible influence of answer lengths on the RT metric, we report the distribution of answer length before and after extraction. Long, self-contained answers typically restate their own context and may entail their predecessors more often, which could lower a model's revision rate for reasons unrelated to perception. This check is done on $\mathcal{T}_1$. Figure 13 shows words per extracted answer. Annotators answered around 20 words on Q_1 and 18 on Q_2, in comparable scale with most models except Qwen3-Omni-30B and Ming-Lite-Omni. Length also does not order the other models: Gemma and the Qwen2.5-Omni models don't write significantly longer answers than other perceivers, yet still change their readings sparingly(Figure 5). This indicates that the difference observed in figure 5 is not a result of answer length or extraction but indeed related to the perception itself. In figure 14 we report the length check done on the **full set** with similar patterns.

C.5 Value of updating

Revising often is evidence of active perception only if the revisions help. We score each answer with the rank of Eq. 4 against what an annotator said *from the same evidence* with target $H_Q^a\big(D_k(V)\big)$ and distractors $H_Q^a\big(D_{\kappa(k,V,V')}(V')\big)$ on other videos. Referring to the SCoRe's $\Delta(t_1, t_2)$ (Kumar et al., 2025), we then subtract the same rank computed with the perceiver's *first* answer $A_Q\big(D_1(V)\big)$ in place of its current one:

$$\Delta_Q(A, V, k) = \mathrm{R}\Big(A_Q\big(D_k(V)\big); \cdot\Big) - \mathrm{R}\Big(A_Q\big(D_1(V)\big); \cdot\Big). \tag{7}$$

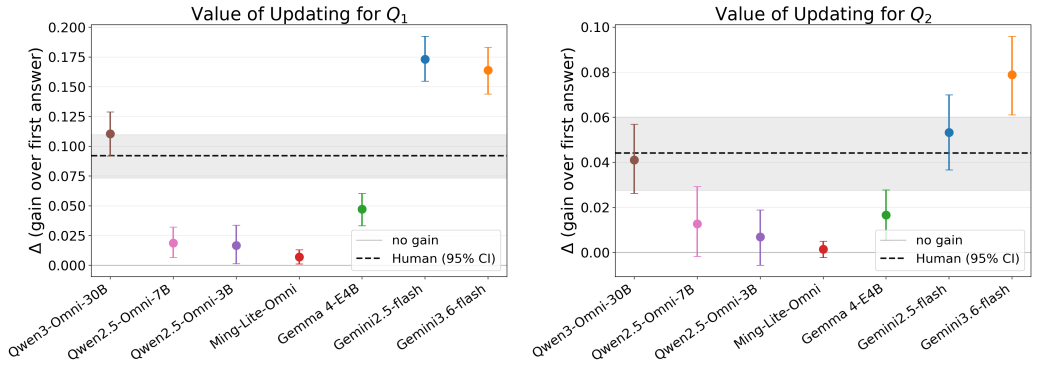
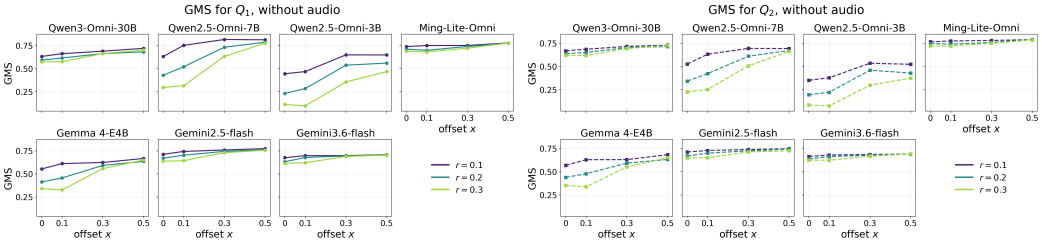


Figure 15: Rank gain Δ computed for all models with human interval shown in black dotted line. Qwen2.5-Omni-7B/3B and Ming-Lite-Omni are close to 0, while Gemini models are gaining much more positively. The comparison should be made with 0 instead of human standards.

Δ is what the perceiver's updates added over repeating its first answer. A perceiver that copies forward scores exactly 0, and $\Delta > 0$ means later readings moved towards what people perceived.

We do this experimnet on $\mathcal{T}_2$ in alignment with the Ant measure. Figure 15 shows that annotators gain around $+0.09$ on Q_1 and $+0.044$ on Q_2, so their revisions are productive rather than noise. Ming-Lite-Omni, Qwen2.5-Omni-7B/3B gain at most 0.021 on either question, with intervals at or near zero: their whole cumulative sequence is worth little more than its first answer. The Gemini models gain more than annotators (above $+0.15$ on Q_1, around $+0.05$ and $+0.08$ for Q_2). This does not mean they update better than people: Δ is measured from each perceiver's own starting point, and annotators often start closer to the target (Figure 4), leaving less headroom. We therefore read Δ against zero, not across perceivers.

C.6 Gaze masking without audio



(a) (b)

Figure 16: GMS plot for Q_1 and Q_2. Qwen2.5 models has lower GMS with larger r and smaller x, showing they are sensitive to corruption. Gemma shows same pattern for Q_1 but not obvious for Q_2. Other models are not sensitive to gaze based manipulation.

Figure 16a and 16b compares the GMS without audio. We observe that without audio the effect of masking is significantly higher. For sensitive models such as Qwen2.5-Omni-7B/3B, the GMS drops below 0.20 with $r = 0.3$ and $x \leq 0.1$, indicating that they are even more influenced by the masking. On the other hand, Ming-Lite-Omni and Gemini models are still not being influenced sharply with varying (x, r).