TRACE: Real-Time Multimodal Common Ground Tracking in Situated Collaborative Dialogues

Hannah VanderHoeven[1], Brady Bhalla[2*], Ibrahim Khebour[1], Austin Youngren[1], Videep Venkatesha[1], Mariah Bradford[1], Jack Fitzgerald[1], Carlos Mabrey[1], Jingxuan Tu[3], Yifan Zhu[3], Kenneth Lai[3], Changsoo Jung[1], James Pustejovsky[3] and **Nikhil Krishnaswamy[1]**

[1]Colorado State University, Fort Collins, CO USA;
[2]California Inst. of Technology, Pasadena, CA USA; [3]Brandeis University, Waltham, MA USA

Correspondence: hannah.vanderhoeven@colostate.edu, nkrishna@colostate.edu

Abstract

We present TRACE, a novel system for live *common ground* tracking in situated collaborative tasks. With a focus on fast, real-time performance, TRACE tracks the speech, actions, gestures, and visual attention of participants, uses these multimodal inputs to determine the set of task-relevant propositions that have been raised as the dialogue progresses, and tracks the group's epistemic position and beliefs toward them as the task unfolds. Amid increased interest in AI systems that can mediate collaborations, TRACE represents an important step forward for agents that can engage with multiparty, multimodal discourse.

1 Introduction

When engaging in a shared task, collaborators continually exchange information about goals, obstacles, and next steps, thereby building a shared understanding of the problem, or a "common ground" (Clark and Brennan, 1991). In situations involving hybrid human-AI teams, although there is an increasing desire for AIs that act as collaborators with humans, modern AI systems struggle to account for such mental states in their human interlocutors (Sap et al., 2022; Ullman, 2023) that might expose shared or conflicting beliefs, and thus predict and explain in-context behavior (Premack and Woodruff, 1978). Additionally, in realistic scenarios such as collaborative problem solving (Nelson, 2013), beliefs are communicated not just through language, but through multimodal signals including gestures, tone of voice, and interaction with the physical environment (VanderHoeven et al., 2024b). Since one of the critical capabilities that makes human-human collaboration so successful is the human ability to interpret multiple coordinated modalities in real-time, collaborative AIs would need to likewise replicate this ability in live real-time settings, but this remains extraordinarily difficult for machines.

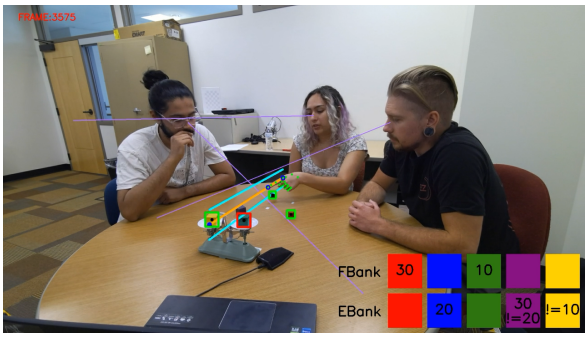

Figure 1: Three participants performing the Weights Task with overlay showing detected deixis, objects, and gaze directions, as well as banks of *evidence* (EBANK) and agreed-upon *facts* (FBANK) regarding the weights of each differently-colored block.

Our system, TRACE (Transparency in Collaborative Exchanges) addresses this problem with the following novel and unique contributions in a single system:

- Real-time tracking of participant speech, actions, gesture, and gaze when engaging in a shared task;
- On-the-fly interpretation and integration of multimodal signals to provide a complete scene representation for inference;
- Simultaneous detection of asserted propositional content and epistemic positioning to infer task-relevant information for which evidence has been raised, or which the group has agreed is factual;
- A modular, extensible architecture adaptable to new tasks and scenarios.

We demonstrate TRACE on the task of tracking the *common ground* that emerges within triads performing a situated collaborative task called

*This work performed under a CalTech Summer Undergraduate Research Fellowship (SURF) program at Colorado State University.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 40–50
April 30, 2025 ©2025 Association for Computational Linguistics

the Weights Task (Khebour et al., 2024a) (Fig. 1). Importantly, our system jointly operationalizes methods previously evaluated in isolation (Khebour et al., 2024b; VanderHoeven et al., 2024a; Venkatesha et al., 2024), and we do this in *real-time* while balancing speed and performance. To our knowledge, no previous system has attempted this. TRACE can be adapted to similar situated collaborative tasks with sufficient data, making it useful for real-time analysis of collaborative problem solving and multimodal communication. We also assess the level of error introduced into multiple features by different levels of live automated processing when compared to manually-annotated ground truth. TRACE represents an important advance for for AI systems that can model group collaboration in real-time situated contexts. A video demonstration showcasing multiple aspects of a collaborative interaction is available here. Installable code and setup instructions may be found at `https://github.com/csu-signal/TRACE/releases/tag/naacl-demo`, available at present under the MIT license.

2 Related Work

Dialogue state tracking (DST) aims to update the representations of a speaker's (user's) needs at each turn in the dialogue, taking into account past dialogue moves and history (Budzianowski et al., 2018; Liao et al., 2021; Jacqmin et al., 2022). Dialogue studies provide a technical definition of a "common ground" as a set of shared beliefs among participants in an interaction (Grice, 1975; Clark and Brennan, 1991; Traum, 1994; Stalnaker, 2002; Asher and Gillies, 2003; Traum and Larsson, 2003; Hadley et al., 2022). This attribution of mental states to one's interlocutors is central to *Theory of Mind* (Premack and Woodruff, 1978). Such internal states may be communicated not just through language, but nonverbal behavior as well (Hall et al., 2019). Understanding nonverbal behavior in multimodal communication has been of longstanding interest in psychology and HCI (Kendon, 1997, 2004; McNeill, 2005; Beilock and Goldin-Meadow, 2010), and has recently found increasing relevance to AI systems (Sigurdsson et al., 2016; Gu et al., 2018; Li et al., 2020).

Our work is similar in spirit to the Dialogue State Tracking Challenge (DSTC; Williams et al. (2016)). While both are consistent with Clark (1996)'s notion of common ground and may involve a live evaluation, our work is novel in that we address

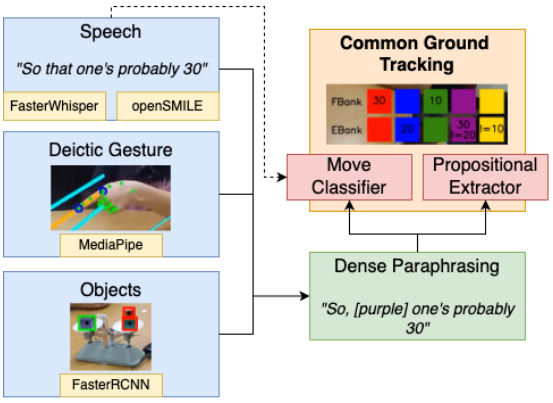

Figure 2: High-level schematic of information flow in real-time multimodal common ground tracking. We combine signals from speech, gesture, and objects in the environment to determine the task-relevant content being discussed, and the epistemic positioning expressed in each utterance. Logical closure rules unify these outputs into the set of common QUDs (QBANK—not displayed for space reasons), pieces of evidence (EBANK), and facts (FBANK).

the content of the common ground directly rather than proxies such as goal, and interpret multimodal signals in a situated collaborative task context. Similar work that involves situated interaction includes grounding of action descriptions (Beinborn et al., 2018), and previous work using interactive virtual avatars (Krishnaswamy et al., 2017; Pustejovsky et al., 2017; Krishnaswamy et al., 2020) where a common ground can be constructed *post hoc* (Krishnaswamy and Pustejovsky, 2020).

Khebour et al. (2024b), introduced a novel task of *common ground tracking* (CGT) that automatically identifies the set of shared beliefs and "questions under discussion" (QUDs) of a group with a shared situated task and goal, using multimodal signals to both extract the propositional content being expressed by task participants (Venkatesha et al., 2024), and their epistemic positionings toward them, to mark which are accepted as facts by the group vs. merely evidenced. VanderHoeven et al. (2024b) laid out the different modalities that may be used to give an AI system enough information to adequately interpret a collaborative dialogue. TRACE operationalizes and integrates the aforementioned works in real time.

3 System Description

TRACE is a modular system that combines features from speech, acoustic, RGB, and depth channels to interpret task participants' linguistic and nonverbal behavior to model their common task-relevant

beliefs. Descriptions of the individual modules and their relations to previous research are given in Sec. 3.1, and Fig. 2 shows how they interact. All feature modules specify an output *interface* or a class representing the data type a module outputs. Modules also specify zero or more input interfaces, which they require in order to calculate the output. For example, the Propositional Extraction module requires only text input while the Dense Paraphrasing module requires text, gesture, and object inputs (Fig. 2). TRACE enables modules to set their input interfaces as dependencies, and the contents of the required output interface will be automatically passed into the dependent input interface. Thus, the entire system, or any such system built with TRACE can be structured as a directed graph with features as vertices and edges connecting a module and all of its dependencies. This framework allows for swapping in and out different multimodal processing modules to create variants of the system.

TRACE is demonstrated on the Weights Task (Khebour et al., 2024a), a situated collaborative task where triads work together to determine the weights of five differently-colored blocks using a balance scale. The block weights follow the pattern of the Fibonacci sequence in increments of 10 grams. The correct weight assignments by color are: 10g (red), 10g (blue), 20g (green), 30g (purple), and 50g (yellow). Beliefs in the Weights Task constitute evidence for or against weight assignments for blocks, or agreement upon the weight of a given block, as discussed in Khebour et al. (2024b). Fig. 1 shows the physical task space, with blocks and the balance scale on a table with 3 participants seated around it. The task is recorded using an Azure Kinect RGBD camera, and either a single MXL AC-404 ProCon microphone or 3 individual lavalier or headset mics—one for each participant (Bradford et al., 2022). The system as presented in the demonstration video runs on an Alienware Aurora R12 tower with an NVIDIA RTX 3090 with 24GB of VRAM but can run on systems as small as a laptop with an RTX 3070 Ti (8GB VRAM). See Appendix C for further details.

3.1 Modules

Here we describe the individual modules used by TRACE for multimodal processing. Our choices of processing techniques were motivated by the need to simultaneously optimize for both performance and the speed necessary to run in real time while remaining within the aforementioned hardware limits when running all modules simultaneously. Thus, we combine older and newer techniques to provide sufficient performance while running quickly enough for real-time processing. The technical details of each are in the referenced papers. Details such as hyperparameters or minor modifications we made to the original models are deferred to Appendix A.

Automatic Speech Recognition For automatic speech recognition, we use the FasterWhisper variant of Whisper (Radford et al., 2023). Acoustic and prosodic features of utterances are extracted using openSMILE (Eyben et al., 2010).

Object Detection Detection of the blocks in the scene uses a FasterRCNN ResNet-50-FPN model (Lin et al., 2017) trained over block bounding box annotations from the original Weights Task Dataset (WTD; Khebour et al. (2024a)).

Deictic Gesture and Gaze Detection We use the 3-stage gesture recognition method from VanderHoeven et al. (2023) that operationalizes the gesture semantics of Kendon (1997) to detect the "stroke" or semantically-important phase of a gesture; e.g., for deictic gesture, this is the extension of a digit. We then use VanderHoeven et al. (2024a)'s method to calculate a "pointing frustum" (Kranstedt et al., 2006) from the extended digit into 3D space and intersect it with detected objects to determine what the targets of deixis are. A similar method is used to infer gaze direction from the direction of participants' heads (see Appendix A).

Multimodal Dense Paraphrasing (MMDP) In situated dialogue, objects are often referenced with demonstratives ("this," "that one," etc.). Fully interpreting these demonstratives requires recourse to one or more non-linguistic modality. We follow a *multimodal dense paraphrasing* (MMDP) procedure (Tu et al., 2023, 2024), which uses additional context to merge multimodal channels into enriched LLM prompts that query the state of the common ground. TRACE uses MMDP to build a list of potential referents from the objects selected by deixis, and then takes demonstratives in utterances that overlap with the deictic gesture and replaces them with the names of the objects, depending on the objects in the list (ordered by distance from the pointing digit) and the grammatical number of the demonstrative pronoun. Table 1 provides examples.

Common Ground Tracking (CGT) CGT follows Khebour et al. (2024b)'s method, combin-

Blocks	Utterance	Dense paraphrase
purple	So, *that*'s more than 20	So, *[purple block]*'s more than 20.
red, green	So *that*'s a 10 and *that*'s a 20 right there?	So *[red block]*'s a 10 and *[green block]*'s a 20 right there?
green, purple	So, *these* are 50 on here?	So, *[green block, purple block]* are 50 on here?

Table 1: Utterances, retrieved blocks, and corresponding (ground truth) MMDPs.

ing an epistemic "move" classifier, a propositional extractor, and a set of logical closure rules to enforce consistency over the facts, evidence, and questions under discussion within the group's common ground. Utterances are classified as expressing an epistemic $STATEMENT$ of evidence toward the currently or most-recently expressed proposition, $ACCEPT$ance of previously-surfaced evidence as fact, $DOUBT$ of evidence or a fact, or none of the above. These classifications are performed on the basis of the MMDPed text of the transcribed utterance encoded through BERT (Devlin et al., 2019), and the acoustic/prosodic features extracted with openSMILE, and does not include other features like Gesture-AMR (GAMR; Brutti et al. (2022)) or collaborative problem solving facets (Sun et al., 2020), which require manual annotations or an auxiliary model (Bradford et al., 2023).

Propositions are extracted from the text of the dense-paraphrased utterance, and take the form of relations between blocks or between blocks and weight values (e.g., $red = 10$ or $red = blue$). Here we use the cross-encoder method from Venkatesha et al. (2024), who report improved performance over the cosine similarity method used in Khebour et al. (2024b). Further technical specifications are given in Appendix A.

Logical closure rules consistent with those in Khebour et al. (2024b) unify the extracted propositions and epistemic moves into the contents of the common ground. $STATEMENT(p)$ raises evidence consistent with p to EBANK. $ACCEPT(p)$ raises p (if in EBANK) to FBANK. $DOUBT(p)$ lowers p (if in FBANK) to EBANK. For $ACCEPT$s and $DOUBT$s, we assume p to be the most-recently stated proposition if no proposition is extractable from the utterance (e.g., utterances like "yeah" or "wait, I don't think so").

4 Evaluation

We evaluate over Groups 1, 2, 4, and 5 of the Weights Task Dataset (WTD). These groups have been fully manually annotated with ground truth labels for all speech transcriptions, gestures, block locations, epistemic "move" labels, and expressed propositions. We use move and proposition models that exclude the relevant test group from the training data. We also present an evaluation of the demonstration video (linked in Sec. 1) where the models used were trained over the entire WTD.

Following Khebour et al. (2024b), our primary metric is Sørensen-Dice Coefficient (DSC; Dice (1945); Sørensen (1948)). This can be computed against each utterance (as in Fig. 3), or averaged over a dialogue (as in Table 2), and indicates the match between the set of propositions extracted by TRACE using all the component models, and the set of propositions in the ground truth, while also normalizing for the size of the two sets.

We compare TRACE's live performance to *post hoc* results from Khebour et al. (2024b), who considered only utterances that were annotated as expressing some epistemic position, and used human-annotated dense paraphrases and gesture annotations using GAMR (Brutti et al., 2022). We present DSC over all three common ground banks, as well as over the union of FBANK and EBANK, which approximates the quality of propositional extraction independent of epistemic move classification (because misclassified moves may raise a proposition p to the wrong bank). Due to the challenge of real-time processing, our reported numbers are often lower, though we do find a few cases where we match or slightly exceed previous results, such as extracting QUDs in Group 5. Generally, live processing does fairly well at tracking the set of QUDs over time but struggles to assign facts and evidence to the right level. This was also a challenge noted in the original Khebour et al. (2024b) results.

As such, we also compare to results reported in Tu et al. (2024), who focus on using multimodal dense paraphrasing to identify common ground in the aftermath of human-labeled $ACCEPT$ moves, and hence only report results on FBANK. Thus, their results can be directly compared to the union of facts and evidence (F $\cup$ E) in the live condition, as they implicitly assume the contents of other preceding utterances accumulate evidence which is then moved to fact status upon the occurrence of a human-labeled $ACCEPT$.

This represents comparisons to all previously-reported SOTA on this task and data, however our numbers represent real-time automated processing of *all features* considering all utterances (unlike Khebour et al. (2024b)), meaning that we are simultaneously *detecting and classifying* epistemic

positioning, *and* considering all banks of the common ground (unlike Tu et al. (2024)). Table 2 shows the results and comparisons.

	Group 1	Group 2	Group 4	Group 5
		TRACE		
QBank	0.349	0.656	0.741	0.546
EBank	0.063	0.135	0.231	0.214
FBank	0.000	0.205	0.000	0.000
F ∪ E	0.246	0.377	0.231	0.464
		Khebour et al. (2024b)		
QBank	0.767	0.911	0.817	0.514
EBank	0.344	0.713	0.812	0.335
FBank	0.000	0.528	0.045	0.165
F ∪ E	1.000	0.922	0.832	0.959
		Tu et al. (2024)		
F (TA)	0.883	0.580	0.450	0.652
		GPT-4O (from Tu et al. (2024))		
F (TA)	0.841	0.331	0.321	0.478

Table 2: Comparison of TRACE live tracking performance to *post hoc* results from other methods. **F (TA)** represents the contents of FBANK when only "True Accepts" are considered, as in Tu et al. (2024), which is equivalent to **F** ∪ **E** in the real-time condition.

Like previous results, we see significant variance across groups, indicating the intrinsic challenge of the common ground tracking task. TRACE overpredicts $STATEMENT$s and underpredicts $ACCEPT$s in Groups 4 and 5, just like Khebour et al. (2024b). This leads to propositions correctly being surfaced as evidence but never raised to facts according to the model. We also find that TRACE approaches or outperforms GPT-4o (as reported in Tu et al. (2024)) on fact retrieval in Groups 2 and 5 given the assumption of true $ACCEPT$ classification.

Fig. 3 shows an evaluation over the demonstration video, showing detected common ground vs. the annotated ground truth over time per utterance. This shows a typical pattern in the evolution of common ground as the task unfolds, where the group begins with a full set of QUDs, over time evidence is surfaced (shown as peaks in EBANK), and certain correct facts are agreed upon over time.

Sequences in the demonstration video can also be explained by individual module performance. The utterance "okay, I guess this one's 10" is correctly paraphrased as "okay, I guess [red] one's

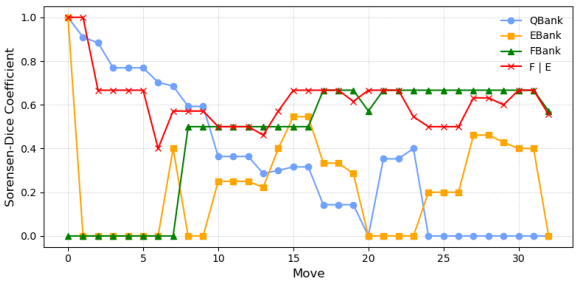

Figure 3: DSC of each common ground bank vs. moves in the demo video dialogue. A value may be zero if there is no intersection between the predicted and ground truth sets, but also if the union of the ground truth and predicted sets for that bank is empty, resulting in zero denominator. We treat this case as no similarity.

10" using gesture and object signals, and the correct proposition $red = 10$ is extracted. However, the move classifier predicts that the utterance is an $ACCEPT$ (correct label is $STATEMENT$). Since $red = 10$ is not already in EBANK here, $red = 10$ is not raised to FBANK. Later, the utterance sequence "so purple is 30 and blue is 10?" (with pointing) and "yeah, that should be 40 right there" raise both $purple = 30$ and $blue = 10$ from EBANK to FBANK simultaneously.

4.1 Substitution Study

Because we have the ground truth annotations for speech transcriptions, gestures, and block locations, we perform a substitution study following Cohen and Howe (1988) to quantify of the level of error introduced into the final output by automated processing of these features. This study is conducted by evaluating over a video as if live, except instead of passing the model outputs a given feature into the CGT pipeline, we pass the ground truth values. This allows us to evaluate the impact of each module's actual performance on the whole pipeline when compared to a hypothetical scenario where that module performs perfectly. Because of the nature of dependencies between features (see Sec. 3), fully removing these features would prevent the system for operating entirely, and so a standard ablation study is not realistic, hence our framing of a substitution study (see Appendix B for more). However, given TRACE's many interlinked components, such evaluation is critical to understand where specific components can be improved.

Table 3 shows substitution study results over the 4 WTD test groups. When using "ground truth utterances," these are passed into the automated move classifier and propositional extraction models, and MMDP is performed using the automated pointing outputs. "Ground truth gestures" indicates

	Ground truth utterances				Ground truth gestures				Ground truth objects			
	Group 1	**Group 2**	**Group 4**	**Group 5**	**Group 1**	**Group 2**	**Group 4**	**Group 5**	**Group 1**	**Group 2**	**Group 4**	**Group 5**
QBank	0.423	0.498	0.714	0.549	0.343	0.634	0.783	0.570	0.351	0.657	0.762	0.554
EBank	0.031	0.042	0.248	0.263	0.050	0.147	0.280	0.290	0.067	0.135	0.231	0.247
FBank	0.054	0.183	0.247	0.000	0.053	0.202	0.000	0.000	0.204	0.228	0.000	0.000
F ∪ E	0.383	0.324	0.419	0.555	0.384	0.377	0.368	0.608	0.220	0.405	0.255	0.508

Table 3: Substitution study results over the 4 WTD test groups, where instead of automatically processing the indicated feature, the ground truth value from the annotated data is passed into the rest of the pipeline.

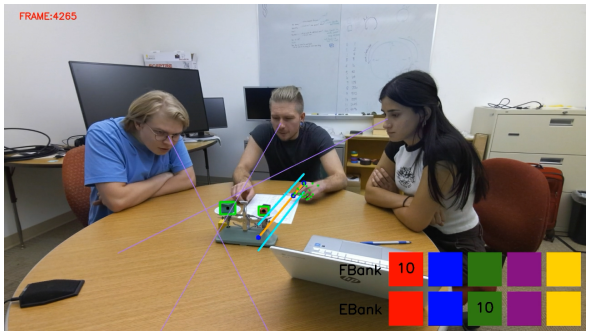

Figure 4: Still from Group 2 showing both a false positive and false negative pointing detection.

that MMDP uses ground truth pointing annotations, automatically transcribed utterances, and automatically detected blocks. Likewise, "ground truth objects" indicates that MMDP uses automatically transcribed utterances and automatically detected points, but ground truth object bounding boxes to ensure no missed or misclassified blocks (e.g., the object detector model often confuses the blue and purple blocks due to their similar colors).

Using veridical values for different features often significantly boosts live performance of the other modules across the board. This is most evident when using ground truth utterance transcriptions, indicating that small improvements in live ASR (e.g., correctly transcribing "that" instead of "the") would have a pronounced positive effect. Using ground truth pointing annotations is most helpful in situations like the one shown in Fig. 4. Here, gesture recognition falsely detects deixis on the middle participant's left hand but misses it on the right hand. The accompanying utterance is "now the first go through *it* bounced twice and actually..." When using the ground truth pointing (annotated on the right hand, which is pointing to the green block), the MMDPed utterance is "now the first go through *[green]* bounced twice and actually...", which later helps in the correct classification of $STATEMENT(green = 20)$. This shows how small improvements in an individual feature can result in substantial overall performance increase.

Where using ground truth values adversely impacts performance, this indicates that the ground truth annotations themselves may be noisy. For instance, overlapping speech in the original Group 2 video led some utterances to be omitted from the manual transcription, meaning that ASR picked up some contentful speech that was absent in the ostensible "ground truth". Annotations of pointing frames are likewise also somewhat conservative.

5 Conclusion

TRACE addresses the already-challenging problem of common ground tracking in a situated collaborative task, and undertakes the added novel challenge of doing so in real time with live processing of multimodal signals. We integrate epistemic state classification (Khebour et al., 2024b), propositional extraction (Venkatesha et al., 2024), dense paraphrasing (Tu et al., 2023, 2024), and gesture detection (VanderHoeven et al., 2024a), using techniques that appropriately balance speed and performance. TRACE's dependency graph-based architecture facilitates study of multimodal fusion (Khebour et al., 2025), and adaptation to other situations and tasks by easily substituting or adding models and features. For example, modules for posture classification can be introduced to model social dynamics and level of individual task engagement (Moulder et al., 2022; Adams-Wiggins and Dancis, 2022). Additionally, TRACE's codebase has already been leveraged in ongoing work as a flexible platform that can support multiple different research and demonstration efforts. One such example is Palmer et al. (2025), which uses the underlying TRACE platform to track nonverbal indicators of group engagement, such as joint visual attention and posture. TRACE will be of use to researchers in dialogue studies and collaborative problem solving, and can be used in building AI systems that mediate collaboration, such as by inserting probing questions (Karadzhov et al., 2023; Nath et al., 2024) at key moments.

Adaptation of real-time common ground tracking with TRACE to other collaborative task scenarios is straightforward. Many modules use off-the-

shelf processors like Whisper ASR, openSMILE, and MediaPipe (Lugaresi et al., 2019). Models for epistemic classification and propositional extraction can be trained on annotated data. Propositions for a new task can be deterministically enumerated following Venkatesha et al. (2024), and the faster but less accurate cosine similarity method requires no new model. Our gesture recognition models can be reused as long as participants' are positioned similarly to the Weights Task. There is a practical limit of ∼5 bodies within the camera FOV.

Future improvements to TRACE as used in the Weights Task include also tracking the individual beliefs about the task not shared by the group, moving away from specialized depth cameras through RGB versions of modules like gesture recognition, and improving epistemic move classification through richer representation of modalities like gesture and facial expression. Further improvements to and use cases for TRACE include deploying it in less-constrained, more flexible tasks where conversations may be more ambiguous or more diverse, range in different directions, and cover a wider potential space of propositions. We are currently working on expanding TRACE's usage into such tasks, such as collaborative construction and annotation of non-verbal indicators in general collaborative settings. Additionally, we continue to improve TRACE's flexibility and codebase organization to allow it to accommodate new models and custom technologies, further permitting researchers to deploy individualized solutions for each modality and scenario of interest.

Acknowledgments

This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, and the National Science Foundation (NSF) under subcontracts to Colorado State University and Brandeis University on award DRL 2019805 (Institute for Student-AI Teaming). Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government. We would also like to thank the anonymous reviewers whose feedback helped improve the final copy of this manuscript. All errors are the responsibility of the authors.

Ethical Statement

Multimodal processing entails modeling people's speech and gesture patterns, body language, facial expression, etc., and raises questions about such technologies being used for tracking and surveillance. For example, modeling how individuals collaborate also involves at least tacitly modeling their linguistic and reasoning patterns, which may be sensitive. The WTD used for training the core modules—common ground tracking, pointing, object detection, etc.—is publicly-available anonymized data that was collected under protocols reviewed by institutional review boards for ethical research, and were conducted with subjects who consented to the release of the data. However, collaboration modeling technology should be treated cautiously when it comes to ingesting multiple modal channels from specific people.

References

Karlyn R Adams-Wiggins and Julia S Dancis. 2022. Marginality in inquiry-based science learning contexts: the role of exclusion cascades. *Mind, culture, and activity*, 29(4):356–373.

Nicholas Asher and Anthony Gillies. 2003. Common ground, corrections, and coordination. *Argumentation*, 17:481–512.

Sian L Beilock and Susan Goldin-Meadow. 2010. Gesture changes thought by grounding it in action. *Psychological science*, 21(11):1605–1610.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal Grounding for Language Processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339.

Mariah Bradford, Paige Hansen, J Ross Beveridge, Nikhil Krishnaswamy, and Nathaniel Blanchard. 2022. A deep dive into microphone hardware for recording collaborative group work. In *Proceedings of the 15th International Conference on Educational Data Mining*, page 588.

Mariah Bradford, Ibrahim Khebour, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2023. Automatic detection of collaborative states in small groups using multimodal features. In *International Conference on Artificial Intelligence in Education*, pages 767–773. Springer.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract Meaning Representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Herbert H Clark. 1996. *Using language*. Cambridge University Press.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Paul R Cohen and Adele E Howe. 1988. How evaluation guides AI research: The message still counts more than the medium. *AI magazine*, 9(4):35–35.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. OpenSMILE: the Münich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056.

Lauren V Hadley, Graham Naylor, and Antonia F de C Hamilton. 2022. A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology*, 1(1):42–54.

Judith A Hall, Terrence G Horgan, and Nora A Murphy. 2019. Nonverbal communication. *Annual review of psychology*, 70(1):271–294.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–350, Edinburgh, UK. Association for Computational Linguistics.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. DeliData: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–25.

Adam Kendon. 1997. Gesture. *Annual review of anthropology*, 26(1):109–128.

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne M Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, Pustejovsky James, and Nikhil Krishnaswamy. 2024a. When Text and Speech are Not Enough: A Multimodal Dataset of Collaboration in a Situated Task. *Journal of Open Humanities Data*, 10(1).

Ibrahim Khebour, Changsoo Jung, Jack Fitzgerald, Huma Jamil, and Nikhil Krishnaswamy. 2025. Feature Contributions to Multimodal Interpretation of Meaning. In *International Conference on Human-Computer Interaction*. Springer.

Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024b. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia. ELRA and ICCL.

Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2006. Deixis: How to determine demonstrated objects using a pointing cone. In *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop, GW 2005, Berder Island, France, May 18-20, 2005, Revised Selected Papers 6*, pages 300–311. Springer.

Nikhil Krishnaswamy, Pradyumna Narayana, Rahul Bangar, Kyeongmin Rim, Dhruva Patil, David McNeely-White, Jaime Ruiz, Bruce Draper, Ross Beveridge, and James Pustejovsky. 2020. Diana's World: A Situated Multimodal Interactive Agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13618–13619.

Nikhil Krishnaswamy, Pradyumna Narayana, Isaac Wang, Kyeongmin Rim, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Ross Beveridge, Jaime Ruiz, Bruce

Draper, et al. 2017. Communicating and acting: Understanding gesture in simulation semantics. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)—Short papers*.

Nikhil Krishnaswamy and James Pustejovsky. 2020. A Formal Analysis of Multimodal Referring Strategies Under Common Ground. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5919–5927.

Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The Ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*.

Lizi Liao, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. 2021. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9:557–569.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019.

David McNeill. 2005. *Gesture and thought*. University of Chicago Press.

Robert G Moulder, Nicholas D Duran, and Sidney K D'Mello. 2022. Assessing multimodal dynamics in multi-party collaborative interactions with multilevel vector autoregression. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 615–625.

Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin C. Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. "Any Other Thoughts, Hedgehog?" Linking Deliberation Chains in Collaborative Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5297–5314, Miami, Florida, USA. Association for Computational Linguistics.

Laurie Miller Nelson. 2013. Collaborative problem solving. In *Instructional-design theories and models*, pages 241–267. Routledge.

Allen Newell. 1975. A tutorial on speech understanding systems. *Speech recognition*, pages 4–54.

Derek Palmer, Yifan Zhu, Kenneth Lai, Hannah VanderHoeven, Mariah Bradford, Ibrahim Khebour, Carlos Mabrey, Jack Fitzgerald, Nikhil Krishnaswamy, Martha Palmer, and James Pustejovsky. 2025. Speech Is Not Enough: Interpreting Nonverbal Indicators of Common Knowledge and Engagement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

James Pustejovsky, Nikhil Krishnaswamy, Bruce Draper, Pradyumna Narayana, and Rahul Bangar. 2017. Creating common ground through multimodal simulations. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer.

Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672.

David Traum. 1994. A computational theory of grounding in natural language conversation.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. *Current and new directions in discourse and dialogue*, pages 325–353.

Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. 2023. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 39–49, Nancy, France. Association for Computational Linguistics.

Jingxuan Tu, Kyeongmin Rim, Bingyang Ye, Kenneth Lai, and James Pustejovsky. 2024. Dense Paraphrasing for Multimodal Dialogue Interpretation. *Frontiers in Artificial Intelligence*, 7.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Hannah VanderHoeven, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2023. Robust motion recognition using gesture phase annotation. In *International conference on human-computer interaction*, pages 592–608. Springer.

Hannah VanderHoeven, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024a. Point target detection for multimodal communication. In *International Conference on Human-Computer Interaction*, pages 356–373. Springer.

Hannah VanderHoeven, Mariah Bradford, Changsoo Jung, Ibrahim Khebour, Kenneth Lai, James Pustejovsky, Nikhil Krishnaswamy, and Nathaniel Blanchard. 2024b. Multimodal design for interactive collaborative problem-solving support. In *International Conference on Human-Computer Interaction*, pages 60–80. Springer.

Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, James Pustejovsky, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. Propositional Extraction from Natural Speech in Small Group Collaborative Tasks. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 169–180, Atlanta, Georgia, USA. International Educational Data Mining Society.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

A Technical Specifications of Individual Modules

Automatic Speech Recognition We run Whisper at `float16` precision.

Object Detection FasterRCNN was initialized with the default ResNet-50-FPN weights from TorchVision and trained 10 for epochs with batch size 32, input size 3×416×416, SGD with learning rate $1e-3$, momentum $9e-1$, and weight decay $5e-4$.

Gesture Recogntion We use hand features extracted from depth video using MediaPipe (Lugaresi et al., 2019) as inputs to the gesture recognizer. For the "near" and "far" radii for the pointing frustum of VanderHoeven et al. (2024a), we use 40mm and 70mm, respectively.

Gaze Detection In the absence of eye tracking, we use direction of participants' noses as a proxy for gaze direction. This is extracted from the body rigs recognized using the Azure Kinect SDK, which consist of directed acyclic graphs containing 32 "joints." We average both ear joints, resulting in a point roughly behind the nose, and gaze direction is calculated using the vector between this point and the nose joint. Like VanderHoeven et al. (2024a), we extend this vector out into 3D space to see which objects participants' gazes are landing on. Averaging the locations of both eyes and the nose resulted in a stable prediction that matched the direction of the participant's gaze. Because participants are always looking at something even if they aren't focusing on it (unlike intentional deixis), objects are not considered "selected" by gaze, but gaze may be used as a secondary feature.

Epistemic Move Classifier The epistemic move classifier we used is slightly modified from the one appearing in Khebour et al. (2024b). openSMILE features were normalized using min-max scaling. SMOTE (Chawla et al., 2002) was used for oversampling the data, as in Khebour et al. (2024b), but when used for discrete features can create invalid data. For example, collaborative problem solving (CPS) facets (Sun et al., 2020) used in training the model are supposed to be binary values, but SMOTE can output continuous values, so synthetic values are rounded to the nearest binary value. Finally, we also include a ReLU layer after the first linear layer for each modality. See Khebour et al. (2024b) for other model specifications, that remain unchanged.

Propositional Extractor The propositional extractor from Venkatesha et al. (2024) proved to be limited by the sparsity of propositions actually expressed in the task (a total of 128) compared to the total number of propositions that could be expressed in the domain (total of 5,005). For example, while $yellow + purple + green > red$ is a possible proposition according to the combinatorics of the objects, it is extremely unlikely to ever actually be expressed during task performance (because the combination of yellow, purple, and green blocks so obviously outweigh the red block that groups never even need to try this). Meanwhile $green + purple = yellow$ is much more likely but may be sparsely represented in actual data (only occurring once in a group if at all). Therefore we improved cross-encoder performance using data aug-

mentation. We prompted GPT-4 through its API to create 10 utterances that expressed each of the 128 propositions that occurred in the actual data. The model was then trained on the original transcript utterances augmented with this set. The GPT system prompt is given below, which is followed by the specific proposition for which we generated supplementary corresponding utterances. The generated utterances were subsequently human-validated for correctness before model training.

The cross-encoder was trained to output a score $Score(u_i, p_j) = MLP([V_{CLS}, Vu_i, Vp_j, V_{u_i} \odot V_{p_j}])$ for an utterance u_i and a candidate proposition p_j over the concatenated representations of the BERT `[CLS]` token for the utterance-proposition sequence, the individual utterance and proposition, and their Hadamard product, using the same hyperparameters reported in Venkatesha et al. (2024).

Where Venkatesha et al. (2024)'s heuristic pruning left more than 137 candidate propositions, cross-encoder inference became slower than performing a vector-similarity comparison against all propositions in the vocabulary. In these cases, we back off to the cosine similarity method from Khebour et al. (2024b).

GPT System Prompt for Propositional Data Augmentation

Conversation Background: Participants are first given a balance scale to determine the weights of five colorful wooden blocks. They are told that the red block weighs 10 grams, but that they have to determine the weights of the rest of the blocks using a balance scale.

The possible weights of the blocks are 10, 20, 30, 40, 50. Propositional content in the Weights Task takes the form of a relation between a block and a weight value (e.g., $red = 10$), between two blocks (e.g., $red = blue$), or between one block and a combination of other blocks (e.g., $red < blue + green$). The possible colors are red, blue, green, purple, yellow. The possible weights are 10, 20, 30, 40, and 50. The possible relations are $=$, $!=$, $<$, $>$. Generate 10 different utterances that could be expressed by a participant while solving this task that expresses the following proposition:

B Substitution Study Design

An ablation study as technically defined requires that the system experience "graceful degradation" (Newell, 1975) when an input is removed. However in the case of multiparty dialogue, this is often not possible. Due to the nature of dialogue, any automated system will not perform at all in the absence of speech information or transcribed audio. Dense paraphrasing requires access to both gestures and objects simultaneously; viz. dense paraphrased text without either one of them is identical to the raw text (this is evident in the dependencies in Fig. 2). Thus a standard ablation study where one modality is left out entirely is not realistic. Therefore we frame our study as a "substitution" study *a la* Cohen and Howe (1988) which shows the importance of each modality by allowing TRACE to look up veridical information about that modality from the dataset instead of removing it entirely. Thus our evaluation follows extremely long-standing best practices in the field of AI.

C Performance Profiling

Table 4 shows performance statistics for a single live CGT tracking session on a consumer-grade gaming laptop, lasting approximately 5 minutes and using 1 microphone and 1 Kinect, with 3 task participants.

Hardware Specifications	
Processor	12th-gen Intel® Core™ i7-12700H, 2.70 GHz
RAM	16 GB
GPU	NVIDIA GeForce RTC 3070 Ti Laptop GPU
VRAM	8 GB
Live Performance Usage Ranges	
GPU	60.0–74.0%
VRAM	4.5–5.0/8 GB
RAM	54.0–58.0%
- Python	42.0–44.0%
- TRACE Modules	12.0–14.0%
CPU	14.0–20.0%
FPS	5–6

Table 4: Sample performance profiling.

When evaluating live performance, latency must be taken into account, however due to many factors this is difficult to assess consistently. For example, specific system hardware plays a critical role in latency, so latency time reported in one configuration may not be reliably reproduced in another configuration. The configuration reported in Table 4 represents an approximate lower bound on the hardware that will support the version of TRACE reported in this paper, and so the reported frame rate of 5–6 FPS can be taken as an approximate upper bound on the level of latency induced by processing that can be considered acceptable performance for real-time common ground tracking.