THESIS

# EXPLORING THE EFFECTS OF MULTIMODAL FEATURES ON A MACHINE LEARNING KNOWLEDGE TRACKER

Submitted by
Ibrahim Khebour
Department of Computer Science

In partial fulfillment of the requirements
For the Degree of Master of Science
Colorado State University
Fort Collins, Colorado
Spring 2026

Master's Committee:

Advisor: Nikhil Krishnaswamy
Co-Advisor: Nathaniel Blanchard

Christopher Peterson

ABSTRACT

EXPLORING THE EFFECTS OF MULTIMODAL FEATURES ON A MACHINE LEARNING KNOWLEDGE TRACKER

Conversations involve multiple channels of information exchange. Spoken language is the most common, but non-verbal cues such as gestures, body pose, and movements also play a role. These channels carry semantic information but are discrete and harder for machines to detect. Recent advances in multimodal Large Language Models (LLMs) show that incorporating additional modalities can improve performance, raising the question: how much do extra modalities contribute, and what are the limits of continually stacking them? Modeling the flow of conversation remains challenging for AI, particularly in natural, collaborative settings where non-verbal channels are prominent. To address this, TRACE was developed, a multimodal system that monitors shared knowledge in group tasks by tracking utterances, gestures, and actions. The system runs in real time using speech-only features, while an offline version integrates broader modalities, including problem-solving cues from speech, actions, and gestures. This thesis extends the live system by incorporating additional features. Some require training new models to process visual inputs in real time. Since components may differ from the offline version, I will conduct a comparative analysis of both systems. The evaluation will highlight cases where the live version underperforms, as some loss is expected. A comparison with the current live tracker will also measure the impact of new modalities. The Weights Task Dataset [Khebour et al., 2024a, 2023] will be used for training, testing, and evaluation of action and gesture classification. Automating this process reduces the need for manual annotation and links gestures to broader semantic context, offering substantial value for future work.

# ACKNOWLEDGMENTS

First and foremost, I would like to start by expressing my perpetual gratitude to my supervisor Dr. Nikhil Krishnaswamy for his unwavering support, valuable guidance, and insightful feedback throughout this work. I have seen first hand his brilliance and passion toward his work, and effective methodology. His open-mindedness to growth, and the idea of always learning and always trying to evolve is admirable. I will always be truly appreciative of having been given the opportunity to work with someone as inspiring.

In addition, I would like to thank Dr. Nathaniel Blanchard and Dr. Christopher Peterson for helping with their advising in this thesis. I deeply appreciate the time that they generously offered, their constructive input, and willingness to engage with my work.

Last but not least, this work would not have been made possible without the undeniably and important contributions of my labmates in different aspects of this project. And for that I insist on sharing my gratitude and acknowledgment to their respective efforts. I'd like to thank Mariah Bradford, Hannah VanderHoeven, Videep Venkatesha, Huma Jamil, Changsoo Jung, Jack Fitzgerald, Austin Youngren and Carlos Mabrey. And of course other researchers with significant contributions such as Brady Bhalla, Kenneth Lai, Richard Brutti, and Yifan Zhu.

This research was supported in part by the U.S. National Science Foundation AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 and the U.S. Defense Advanced Research Project Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program under Other Transaction award HR00112490377. Approved for public release, distribution unlimited. The views expressed herein are those of the author and should be taken to represent the views, expressed or implied, or the U.S. Government.

# DEDICATION

*To my family, and to my friends*

*for their unconditional love and support*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1
# Introduction

Collaboration is at the core of human problem-solving. Classrooms, workplaces, and research labs, are all places where groups achieve more when they combine their different perspectives and thinking processes, to build on each other's ideas. Despite its importance, collaboration has received relatively little attention in AI research, as most AI systems still treat communication as a single-channel process.

Within situated collaborative environments, due to the multiple channels people use to communicate, the potential benefit of a multimodal model should be self-evidence, where the richness of interaction requires a broader view of multimodality. However, most previous work in multimodal machine learning emphasizes the quantity of data or relies on a small set of modalities, usually images and text.

This thesis takes collaboration as the main application and asks how different modalities (spoken language, gestures, and other non-verbal cues) contribute to understanding group interactions. These signals can often determine how and what information is communicated in real time. A simple head nod can signal agreement while a pointing gesture can anchor a discussion. That is why, before designing AI systems that can support and eventually intervene in human teamwork, it is crucial to understand how these modalities work together.

This research provides a conceptual contribution, by advancing our understanding of multimodal interaction in collaborative settings, and questions the trivial assumption that "more data is always better", specifically in multimodal machine learning. It also offers practical contributions, including s a proof of concept for building richer multimodal AI pipelines that can integrate under-represented features alongside more traditional ones. Showing the scenarios when multimodality strengthens performance, and when it has the opposite effect, this works lays the foundation of how to build AI systems that are more context-aware, more interpretable, and more importantly expandable to better enhance human collaboration.

## 1.1 Motivation

The traditional focus in machine learning systems has mainly been on a single type of input, such as text, images, or audio. Real world interactions are, however, multimodal. Naturally, humans combine speech, gestures, gaze, and body movement to convey meaning.

Only recently, multimodal machine learning has gained visibility [Baltrušaitis et al., 2018]. Technological advances in data collection and data processing made both labor and cost softer for gathering diverse signals. Also, new multimodal fusion architectures made the alignment and integration of these signals within the same system possible. As a result, we get a giant wave of impactful applications, from AI assistants that combine speech and vision, to generative systems for image captioning and video interpretation. This is very promising for multimodal AI, but they highlight some open questions: When is it helpful to add modalities, and when does it create redundancy or noise?

A specific area was left with more questions: collaboration [Barron, 2000]. Multimodal machine learning is focused on individual tasks, when group interactions introduce more interesting challenges and opportunities. To sustain an effective collaboration, participants must maintain a shared common ground [Stalnaker, 2002, Clark, 1996]: an evolving record of what has been said, agreed upon, or disputed. A shared knowledge grants the group time to coordinate, avoid misunderstandings, and adapt to new developments in real time. Tracking common ground is a powerful subject to study for both human and machine understanding of collaboration. By automatically recognizing when beliefs are aligned, when there's a new suggestion, or when progress stalls, AI systems can provide significant support for teamwork. This support can highlight overlooked contributions or questions, it can prompt alternatives, and can also develop techniques to keep the group on track.

This thesis aims at tackling that challenge by studying multimodality within machine learning in collaborative problem solving environments. I aim to uncover both the strengths and the limitations of multimodal fusion, by isolating and analyzing how the signals of different features shift the tracking of common ground. The ultimate goal is to inform the design of AI systems that can enhance human-human conversations by fostering richer and more effective collaboration.

## 1.2 Research Questions and Hypotheses

This work's goal is to explore the following questions:

**RQ1:** How do additional modalities influence the performance of a multimodal machine learning model in a collaboration tracking task?

**RQ2:** What are the limitations of stacking modalities on top of each other?

**RQ3:** To what extent does the inclusion of low-resourced features impact a model's performance in a collaborative setting?

Following up on the research questions, these hypotheses are formulated:

**H1:** Adding additional modalities will improve performance of the multimodal model compared to unimodal baselines in collaboration tracking tasks.

**H2:** The main limitations to stacking modalities on top of each other are computational (training time, memory usage), rather than performance trade-off.

**H3:** Low-resource modalities will require more training time to converge compared to high-resource counterparts, but once trained, their performance contributions will be comparable.

## 1.3 Approach Overview

I will introduce a newly collected dataset of group collaboration, and present a systematic analysis of modality contributions. I design experiments that ablate specific input channels, such as speech, gestures, and other non-verbal cues, to isolate their effect on model performance. I investigate the advantages and the limitations of multimodal fusion. The goal is not merely to add more data, but to understand how each modality shapes the AI system's learning, and to inform the design of interpretable systems that can adapt the integration of the most useful channels.

To investigate these multimodal contributions in collaborative settings, this thesis shows how speech, prosody, gestures and others features affect the performance of a computational model referred to as Common Ground Tracker (CGT). This model automatically tracks shared knowledge in collaborative dialogues. And investigating which modalities matter most and why sheds light on how intelligent systems can better interpret human group interactions.

The CGT framework was initially introduced in [Khebour et al., 2024b]. It models the evolving state of common ground. Common Ground is identified as the beliefs and goals that participants mutually recognize during a collaborative problem-solving. The system integrates a move classifier, which identifies conversational acts such as Statement, Accept, and Doubt, and a set of logical closure rules that update the group's shared belief space. The CGT offers a dynamic representation of what participants agree/disagree on, which issues remain open, and how evidence for or against propositions accumulates over time.

The main focus of this research is to understand why multimodality matters. Speech conveys explicit content and intentions; prosody adds cues about emotion and certainty; gestures enrich spatial and referential grounding. These channels together can shape the

evolution of common ground.The CGT is trained under different sets of modalities. The evaluation of performance improvements alone is not enough, I will also present how some features contribute more effectively than others.

The remainder of this thesis is organized as follows: Chapter 2 reviews related work on multimodal machine learning, with a focus on collaboration and existing datasets. Chapter 3 details the dataset used in this work and its multimodal richness. Chapter 4 introduces the real-time Common Ground Tracking system. Chapter 5 presents the analysis and results. Finally, chapter 6 will conclude this thesis by taking a look at the whole picture, and suggest directions for future research.

# Chapter 2
# Related Work

The scope of this research is interdisciplinary, spanning over artificial intelligence, human-human interactions and learning sciences. This chapter will cover the wide literature that is connected to these research topics. This section reviews literature from two angles: multimodality in machine learning and collaborative problem solving, highlighting studies outside classroom settings. In the first angle, I will cover the different fusion techniques used in previous work, and the challenges often met in the process of training a multimodal ML model. The second angle will help explain the choices made for feature selection, as well as the design of the model and the system that will solve the task. This chapter will help address the research questions above as it situates the study of multimodal machine learning and collaborative problem solving within published work. Indeed, examining fusion strategies and the challenges of combining modalities will lay the groundwork for exploring how additional low-resourced modalities influence model performance.

## 2.1 Multimodality in Machine Learning

The goal of multimodal machine learning is to develop models that can process multiple input channels such as language, vision, audio, gesture and physiological signals, to improve the performance relative to single-modal AI models. Human communication and cognition are inherently multimodal, which makes this area relevant for applications in various domains like human-computer interaction, education, etc.

When designing a multimodal AI system, one of the main decisions that must be taken is at what point should the different features be fused. Three main fusion strategies are commonly used: *early fusion*, *late fusion*, or *hybrid fusion*.

Early fusion, or feature-level fusion, is when the data is joined at the beginning of the AI model. In most cases, the data is simply concatenated, but other aggregate functions can be used. The obtained tensor is used as a single input. Aggregating the features early can lead to imbalance in contribution of the final prediction. This technique is best used when data from all modalities is temporally aligned or when modalities are tightly coupled. If the different modalities do not share the same format or shape it can be difficult to find a meaningful way to represent all the features within a single one. Indeed, the new

representation will be biased towards the numerically larger data, thus losing the semantic information encapsulated in the smaller valued tensor.

Meta-Transformer [Zhang et al., 2023] is a framework that introduced a novel approach to multimodal learning by using frozen encoder capable of processing different modalities by mapping raw inputs into a shared token space making it able to learn across 12 modalities. This is a variation of early fusion using a Transformer encoder that enhanced the model's ability to generalize across diverse tasks. Another model that uses early fusion is the 4M framework [Mizrahi et al., 2023], unifying different modalities into a token sequence that a single Transformer model processes to learn from multiple modalities simultaneously. This model was developed by EPFL researchers to address the limitations of existing multimodal approaches, beating the aforementioned Meta-Transformer by training across 21 different modalities.

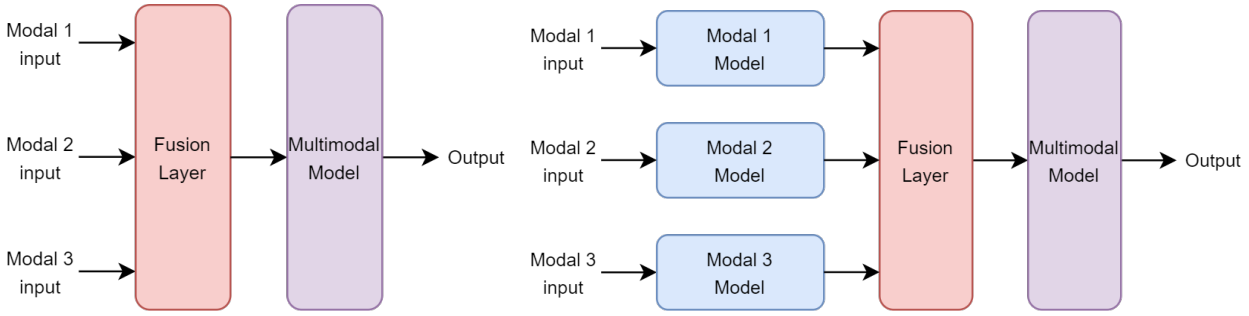
The second approach is the late fusion , also called score-level fusion, which is when each modality is separately used to generate a decision score. The scores are then aggregated to produce a final prediction. A voting system can be used for the final step, where each model votes on a prediction, then the model outputs the most voted one. This is more suited when modalities are heterogeneous or unaligned. When working with modals that have different formats, this method is highly recommended as it preserves multimodality benefits while maintaining the advantages of unimodal models. However, this technique meets its limitations faster as it doesn't have much room where multimodal fusion can improve. One of the issues that can arise is the computation time that rapidly increases relative to the earlier fusion. And the size of the neural network can cause the vanishing gradient problem.

One of the more recent models that use a late fusion adaptation is the AutoM3L [Luo et al., 2024] which leverages the capabilities of LLMs to automatically build multimodal training pipelines. The model understands modalities and selects adequate models based on user requirements, making it easier for the researcher and others to do feature engineering and hyperparameter tuning.

The hybrid fusion approach seems to be the correct equilibrium between the other two as it trains some modals separately (likely to be the modalities that are disconnected from the other ones) while the other features can be merged together. This method gives more room for smarter and specific designs for AI systems since it encourages experimenting with different fusion points for different combinations of modalities. This can serve as a form of hyperparameter tuning to optimize fusion configurations. The main advantage of this fusion process is that feature selection isn't as restrictive as the previous methods. In fact, a hybrid fusion based architecture can work with a larger pool of features, where some can share the same format, and others can have their own shapes.

Hybrid fusion was the base of a multitude of models, MultiZoo & MultiBench [Liang et al., 2023] is composed of a toolkit and a dataset that offer a classic implementation of more than 20 trivial multimodal algorithms that spread across 15 datasets, 10 modalities and 20 forecasting tasks and 6 distinct research areas. The toolkit assesses generalization, computational efficiency and modality robustness. The algorithms used in this framework offer a variety of fusion strategies including hybrid fusion. Unified-IO-2 [Lu et al., 2023] is an auto-regressive sequence model that performs joint decoding across modalities while using separate encoders for modality-specific pretraining. The model was released by Allen Institute for AI and is open source and was trained on a diverse dataset including 1 billion image-text pairs and 180 million video clips, making it a significant step toward understanding multimodality. Self-Supervised Multimodal Learning Survey [Zong et al., 2023] discusses different fusion techniques including contrastive learning methods aligning representations from different modalities and mutual information maximization strategies enhancing the shared information across modalities. Different approaches were presented, all aiming at learning robust multimodal representations all in a self-supervised context. Figure 2.1a shows an example of early fusion and late fusion architectures, while hybrid frameworks can take many different forms depending on the feature selection among other factors like model complexity.

Multimodality is a technical and design decision. This decision determines both what information is captured and how it is captured to contribute to machine learning. The above works show the pros and cons of each fusion technique discussed. This thesis does not propose a novel fusion method but rather aims at extracting the advantages of each technique as they apply to specific modalities in collaboration problem-solving environment. Understanding these design decisions provides the foundation for selecting the signals to be used to automatically track in real-world collaborative environments.




(a) Early Fusion (b) Late Fusion

Figure 2.1: Comparison of Early vs. Late Fusion approaches.

## 2.2 Domain Knowledge

### 2.2.1 Collaborative Problem Solving

Collaborative Problem Solving (CPS) is when a group of individuals get together to solve a common problem. CPS rewards complementary skills, perspectives, knowledge. It is different from individual problem solving as it emphasizes social coordination, communication and group-decision-making. In the 21st century, collaboration is a skill that has grown in importance as reflected by many frameworks such as PISA [OECD, 2019] and 21st century skills [Partnership for 21st Century Learning, 2019]. CPS is present in many theoretical frameworks, highlighting that knowledge is co-constructed through interaction and dialogue. Its part of cognition and provides a channel to analyze actions within social and cultural contexts that are goal driven. Theories reinforce the idea that effective problem solving in groups necessitates cognitive skills but also social competencies like turn taking, negotiation and shared attention. But CPS wasn't spared from the rise of AI in recent times, as we've seen a move toward automating CPS analysis from interactions, leveraging the technological advancements made in NLP, and multimodal machine learning. Verbal signals sensor technologies have also helped with the process of extracting behavioral and nonverbal cues. Indeed, researchers are now developing models that automatically recognize, predict and support CPS. Tasks like dialogue act classification can detect collaboration states from conversations, while embodied agents and virtual tutors can do knowledge tracing incorporating collaborative signals.

To mention some of the more recent frameworks and research advancements made towards improving collaborative problem solving, we can invoke ThinkTank [Surabhi et al., 2025], a framework designed to transform specialized AI agent systems into versatile collaborative intelligence platforms that has capacity to help in difficult problem-solving through different domains. Thinktank gives the power to organizations to leverage collaborative AI for knowledge based tasks, using frameworks built around Llama 3.1, all in a cost-effective package offering security and data privacy. Collaborative Multi-Agent, Multi-Reasoning-Path Prompting (CoMM) [Chen et al., 2024] is a multi-agent framework that improves the reasoning of LLMs based on a collaboration technique that dispatches the task across different roles each assigned to a distinct agent, together forming a problem-solving team. CoMM employs diverse reasoning paths and a variety of expertise making it robust when facing complex science problems. MetaGPT [Hong et al., 2024] is another framework that embraces human workflows into LLM-based multi-agent collaborations. It embeds Standardized Operating Procedures (SOPs) into prompt sequences making it very efficient in collaborative

software engineering benchmarks. CPS-TaskForge [Haduong et al., 2024] is a data generator capable of generating environments for different communication tasks. This is a great tool for producing CPS corpora with multiple agents, enabling researchers to dive deeper into human-human as well as human-AI collaborations.

It is important to understand the dynamics of CPS to clarify why multimodal signals such as speech, gesture, and action are central to modeling collaborative interactions. These insights directly inform the feature selection choices and the framing of the prediction tasks which I will address later in this thesis.

### 2.2.2 Common Ground Tracking

In a dialogue, the listener has to link the semantic understanding from what the speaker says, to the speaker's intent considering the context. This process is referred to as "establishing common ground" between speakers [Grice, 1975, Clark and Brennan, 1991, Stalnaker, 2002, Asher, 1998, Traum and Larsson, 2003]. Common Ground refers to the set of common beliefs among participants in a Human-Human Interaction (HHI) [Markowska et al., Traum, 1994, Hadley et al., 2022], as well as H-Computer Interaction (HCI) [Krishnaswamy and Pustejovsky, 2019, Ohmer et al., 2022] and Human-Robot Interaction (HRI) [Kruijff et al., 2010, Fischer, 2011, Scheutz et al., 2011]. [Del Tredici et al., 2022] have recently employed the notion of common ground operationally to identify and select relevant information for conversational QA system design. [Stewart et al., 2021] and [Bradford et al., 2023] both study human-human collaboration through the lens of an AI agent. Dialogue state tracking (DST) focuses on determining the current dialogue state or belief state of users during their conversations [Budzianowski et al., 2018, Liao et al., 2021, Jacqmin et al., 2022]. Present DST models can be divided into three categories: fixed ontology [Henderson et al., 2014, Mrkšić et al., 2017, Chen et al., 2020], open vocabulary [Gao et al., 2019, Hosseini-Asl et al., 2022, Wu et al., 2019], and hybrid approaches [Goel et al., 2019, Zhang et al., 2019, Heck et al., 2020]. Lately, pretrained language models have gained popularity for modeling slot relationships, while Graph Attention Networks (GATs) have been employed to represent the hierarchical structure of DST, allowing for the integration of semantic compositionality, cross-domain knowledge sharing, and co-reference resolution.

The exploration of nonverbal behavior's role in multimodal communication has historically been a focus of research in HCI, but it has recently garnered renewed attention in the Computational Linguistics field and the broader AI community. Gestures present a variety of distinct dimensions in communication, including references to specific situations, indications of precise spatial locations, and expressions of manner and orientation [Rohrer et al., 2020,

Kopp and Wachsmuth, 2010, Kong et al., 2015, Kendon, 1997, 2004, McNeill, 2019]. Gesture Abstract Meaning Representation (GAMR) [Brutti et al., 2022] addresses gestures that carry the same propositional content and intentionality as speech acts. A gesture can carry meaning independently or can amplify the meaning conveyed by spoken language [Goldin-Meadow, 2005, Krishnaswamy and Pustejovsky, 2020]. An essential aspect of multimodal dialogue is human action, which not only conveys deictic and bridging information but can also enact enduring changes in the world, thereby influencing the common ground [Tam et al., 2023]. Significant efforts have been made to enable action recognition from video [Sigurdsson et al., 2016] [Gu et al., 2018] [Li et al., 2020], along with annotating particular semantic roles [Sadhu et al., 2021].

In [Di Maro et al., 2021], the authors utilize dynamic belief sets represented as graphs, a method I do not directly employ. Nevertheless, this approach aligns theoretically and computationally with the one presented in this thesis since the outcome (post-condition) of a public announcement or observed action can serve as preconditions for transforming evidenced propositions into strong beliefs, which provides a clear interpretation of common ground tracking as a graph.

[Alikhani and Stone, 2020] propose a tutorial that focuses on grounding human-human communication, dialogue systems and multimodal interactive systems. The authors suggest creative ways that conversational agents might seek all while supporting their understanding with evidence. The paper presents how humans establish and maintain common ground during face-to-face communication, how dialogue systems like chatbots use grounding mechanisms to improve understanding, and how multimodal systems can integrate grounding strategies in their designs.

[Udagawa and Aizawa, 2021] address the difficulty with sustaining shared understanding between agents in temporally dynamic environments. The authors introduce Dynamic-OneCommon, a new task in which agents with randomly shifting views of a shared scene continuously refer to the same entity as it moves. A large dataset is collected of human dialogues rich in spatio-temporal language and they analyze the strategies that participants use to maintain grounding, such as motion tracking and referring back to previous agreements. Neural networks were proven to initiate common ground, but they struggle to maintain it under dynamic conditions.

Reflect [Zhou et al., 2022] is another dataset targeting the role of common ground in dialogues. It's composed of 600 dialogue contexts manually annotated by the authors using five types of inferred shared knowledge or beliefs, then they collect 9000 human responses grounded in these inferences. Analysis showed that less than half of original dataset responses meet high quality criteria (sensible, specific, interesting), whereas Reflect's responses exceed

the halfway point. They showed a 30% response quality improvement when they fine-tune BlenderBot or prompt GPT-3 to reflect before replying. They find that inference-aware generations are more engaging and create a contextually richer dialogue compared to reflexive replies.

The design of a Common Ground Tracker is a core contribution of this thesis. Reviewing existing and similar approaches to multimodal representations of interactions establishes the theoretical basis for developing a model that captures shared understanding in collaborative settings. Now that the domain of interest has been broken down, let us examine the datasets available for studying collaborative interactions and common ground.

## 2.3 Pre-existing Datasets

After laying the theoretical foundations of collaborative problem solving and common ground tracking, the next step is to look into available data resources for studying these phenomena. In this section I will introduce pre-existing datasets that influenced this work. However, these datasets either lack the integration of verbal and physical modalities, or are limited in capturing genuine collaborative dynamics.

The HCRC Map Task Corpus [Anderson et al., 1991] is a dataset that focuses on studying natural spontaneous conversations. Its motivation is understanding how people interact and collaborate when completing tasks. In the Map Task, a first person describes a route to a second person who possesses a different map, causing a spontaneous friction among them for engaging them in clarification, negotiation, and mistake correction. The study has been the object of analysis of various aspects of language, highlighting the way people refer to different things, how they align their ideas, and how they use nonverbal cues. The dataset includes conversations with transcriptions, audio recordings, and annotations, making it a valuable resource for dialogue studies, researching communication, and computational models of interaction.

[Liu et al., 2017] propose a novel method for human-human interaction recognition by focusing on spatial relationships and motion trends derived from skeletal data. A feature descriptor is introduced, to capture how different joints move relative to one another, both within a single person and across two interacting individuals. Semantic trends such as upward or forward movement are used to represent motion, then they are encoded into histograms that reflect interaction patterns across time. The authors use a custom kernel function to compare them to assess similarity across different sequences. The approach is evaluated on a newly collected RGB-D dataset, as well as on the SBU Interaction Dataset, and is more accurate compared to existing methods. Combining semantic motion features with spatial

configurations proved especially effective in distinguishing nuanced human interactions.

Another dataset that's closely related to the one proposed here is the one published by [Van Gemeren et al., 2016]. The paper a novel spatio-temportal deformable part model designed for offline identification of fine-grained interactions between two people in video sequences. The model is evaluated on a newly assembled interaction dataset and established benchmarks. The ShakeFive2 dataset is composed of 94 RGB-D video clips recorded in a controlled environment. Each video features two individuals interacting and is accompanied by skeleton joint annotations on a frame-level obtained using Kinect2. The videos present small variations in viewpoint. They are clustered into five close proximity interaction classes: fist bump, hand shake, high five, hug and pass object, where the last category describes a small orange object being passed from one person to the other.

All in all, the datasets mentioned above are very valuable in their respective domains and offer different perspectives on human interaction. The HCRC Map Task Corpus focuses on verbal coordination and linguistic grounding in task-oriented dialogues, exploring how common ground is negotiated through speech and clarification. In contrast to that, the datasets published by [Liu et al., 2017] and [Van Gemeren et al., 2016] prioritize spatial and motion-based cues over verbal communication, by focusing on physical interaction captured via RGB-D and skeletal data. Both use temporal and joint spatial coordinates to model specific dyadic interactions. However, they do not take into account the multimodal and communicative interchange between physical action and spoken language. They also either constrained in terms of modality coverage, or validity in capturing genuine natural collaborative dynamics. The next chapter will build on these weaknesses to present a dataset that meets the demands of this thesis.

# Chapter 3
# Dataset

In this chapter, I will introduce the dataset used to perform the multimodal analysis as well as answer the problem of common ground tracking. Modeling human-human interactions in collaboration environments requires more than speech alone. Focusing just on the speech channel would omit important information that might affect the Human-Human interaction. The channels that make the backbone of knowledge sharing includes a multitude of modalities, among which are speech, gestures, physical grounding actions, gaze, joint attention, level of engagement. These additional channels can enhance AI agents' understanding of human-human interactions but they require processing to extract meaningful semantic information. I will present a multimodal dataset of a situated and shared collaborative task that comprises the channels mentioned above annotated to encode the different aspects of the involvement of participants in the task.

This dataset will be the bridge to answer the research questions as it offers a variety of modalities. By offering both well-established and low-resourced modalities, it enables the analysis of how these channels individually and jointly contribute to model performance, what limitations arise from stacking them, and how less conventional or low-resourced features influence learning in collaborative contexts.

The collection of this dataset, as well as its annotation was a collective work with other researchers; Mariah Bradford, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski and Corbyn Terpstra.

## 3.1 The Weights Task Dataset

By integrating task-driven dialogue, gesture and physical manipulation, the Weights Task Dataset offers a multimodal, collaborative multiparty setting in which verbal and nonverbal modalities jointly contribute to problem solving. This places the dataset as a novelty resource for studying multimodal common ground tracking among other interesting tasks.

At a circular table, triads accomplish the Weights Task (explained in the next paragraph). Participants and task equipment are captured by a webcam. 3 Kinect Azure cameras record RGB-D footage from various perspectives. Task equipment include six blocks of various

weights, sizes, and colors, a balancing scale, a worksheet, and a computer with a survey which participants must complete. Participants were chosen among Colorado State University students who spoke English and were at least eighteen years old. Informed consent was acquired with knowledge. The gender and ethnicity breakdown is shown in Table 3.1.

Figure 3.1: P3 adjusts the scale while P2 seeks clarification, highlighting the need for multimodal understanding.

Table 3.1: Demographic Breakdown of Participants in the Weights Task Dataset

| **Male** | **Female** | **Caucasian** | **Non-Hispanic** | **Hispanic / Latino or Asian** |
|---|---|---|---|---|
| 80% | 20% | 60% | 10% | 30% |

A balance scale is provided to the participants so they may infer the weights of the first five block handed to them after being provided the weight of the red block (10g). Each block is placed on the worksheet in the cell that corresponds to its weight once it has been determined. Participants are then given a new block and are asked to determine its weight using the pattern seen in the original block weights, without the use of a scale. Ultimately, participants must deduce the weight of the subsequent hypothetical block in the set and

articulate their reasoning. The block weights follow the Fibonacci sequence (10g, 10g, 20g, 30g, 50g, ...).

Table 3.2: Descriptive Statistics for Participants and Video Lengths

| **Measure** | **AVG.** | **SD** | **MIN** | **MAX** |
|---|---|---|---|---|
| Participant age (yrs.) | 24.58 | 4.58 | 19 | 35 |
| Video length (mins.) | 17.00 | 7.00 | 9 | 34 |

Following each phase, groups submit their responses via the survey form. The dataset comprises 10 videos (approximately 170 minutes). Table 3.2 presents descriptive statistics of the data. Figure 3.1 illustrates participants interacting with the objects on the table from the viewpoint of the primary Kinect. Figure 3.2 displays various annotations (explained in the coming section).

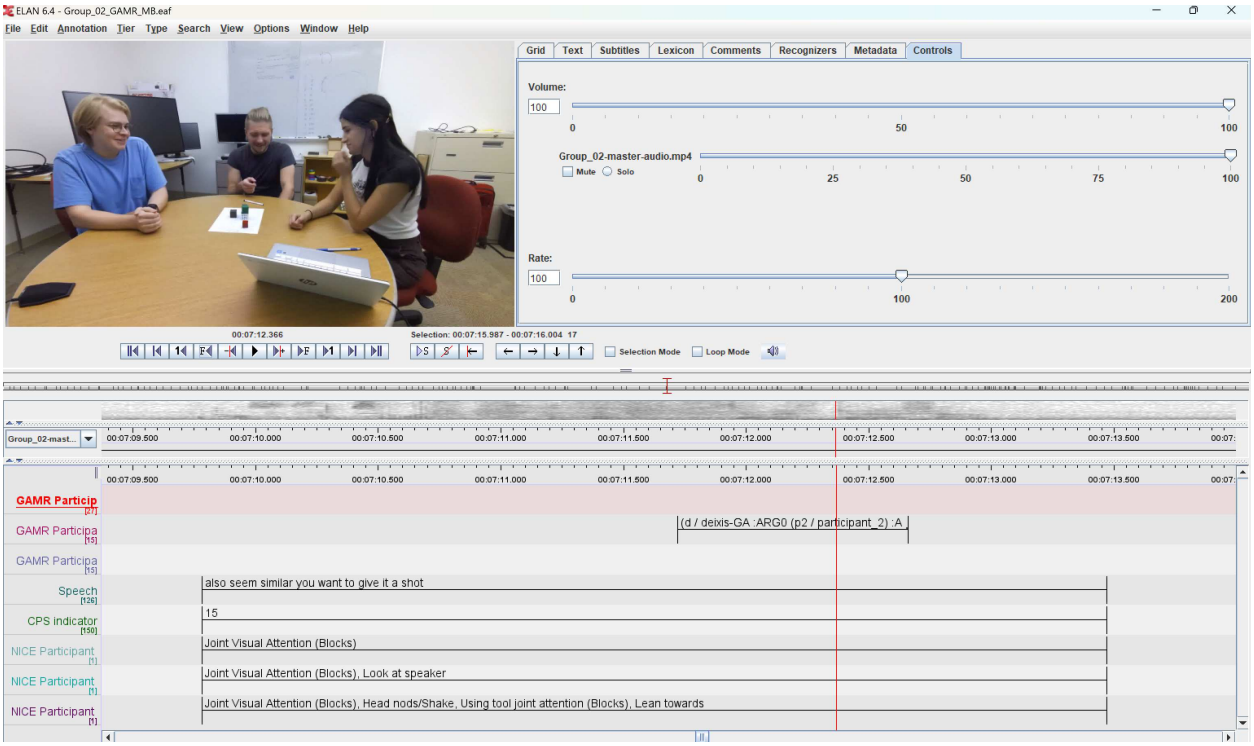



Figure 3.2: Multichannel (GAMR, NICE, speech transcription, and CPS) annotation "score" using ELAN [Brugman and Russel, 2004].

Having introduced the general structure and collection process of the dataset, the following sections describe the annotation procedures and the various modalities that compose it.

### 3.1.1 Annotation Process

By convention, participants are identified numerically from left (P1) to right (P3). Camera and microphone positioning are kept constant and the cameras calibrated using the standard Kinect SDK calibration procedure at the start of each session.

#### Segmentation and Automatic Speech Recognition

For AI agents to be active within the real world, they need to accurately understand the environment and exchanges that are happening around them. In the case where the agent is deployed in a classroom setting, it must have the capability to distinguish different speakers, and when a speech starts and ends. This is referred to as speech segmentation and speech diarization. In the context of the WTD, it can be very challenging to identify distinct discourse components, as these are real-life exchanges that include interruptions, people speaking at random times and orders, and other difficult scenarios. This problem in itself can spiral into a more complex challenge, but since it's not the main purpose of this thesis, we will avoid delving too much into details.

Automatic Speech Recognition (ASR) is the automatic process of determining the transcription of a spoken utterance. Several options exist in terms of ASR models, where each segments audio using a different strategy. [Terpstra et al., 2023] proposed a deeper study on a few of these options, mainly Whisper [Radford et al., 2022] and Google ASR [Velikovich et al., 2018]. It compared these models to the oracle – or manual – segmentation of the audio. Although it revealed limitations in automatic solutions, it also highlighted their advantages in situations where the automation of this process is necessary. These models are also used to transcribe the speech within the audio segmentation.

#### Collaborative Problems Solving (CPS) Facets

CPS coding is conducted at the utterance level utilizing the framework established by [Sun et al., 2020b]. Annotators viewed the video and assigned multiple labels to each utterance based on its content, context, and its position within the conversational sequence. Two annotators annotated the videos (with a Kappa score $\kappa = 0.62$), and an expert, who received comprehensive training in the framework, adjudicated the results. The CPS data is provided in .csv files.

### Gesture Abstract Meaning Representation (GAMR)

Participants' gestures are annotated utilizing the GAMR framework [Brutti et al., 2022]. The majority of WTD gestures are *deictic*, signifying reference to an object or a location. *Iconic gestures* depict attributes of an action or object. The significance of *emblematic gestures* is determined by cultural convention. GAMR was annotated by two annotators who were trained by the authors of the framework (SMATCH F1-score = 0.75). This data is presented in PENMAN notation within .eaf files.

### Nonverbal Indicators of Collaborative-Learning Environments (NICE)

The NICE coding scheme [Dey et al., 2023] captures nonverbal behaviors when people are working together in groups, such as the direction of gaze, posture (e.g., leaning toward or away from the activity area), and usage of tools (including pointing at or to the tool, as well as directly manipulating it). NICE was annotated by an author of the framework over Groups 1-3 and Group 5. This data is presented in .xlsx format. Annotating this modality is time-intensive, and other works were prioritized, but it is still part of the dataset, and offers valuable insight about it.

### Azure Kinect Data

Joint positions and orientations were extracted from each frame of the raw RGBD data. For all 32 joints on each body detected by Microsoft's body tracking SDK. This information (which is available in .JSON files) can be used to analyze body pose and gesture correlation to other modalities, or alone to classify gestures.

The raw data was recorded in .mkv format, including the depth channel, which is too large to include in the distributable dataset. We converted the RGBD videos to .mp4 and extracted the skeleton data from Azure depth channel.

Together, these multimodal annotations form the backbone of the analyses conducted in later chapters, enabling the examination of both verbal and nonverbal contributions to collaborative understanding.

The main purpose of collecting this data was to examine multimodal indicators of CPS. Its rich multichannel nature also makes it suitable for other research directions. It can benefit researches in education and learning sciences to create activities that encourage group collaboration and learning. This dataset ca be useful in natural language processing tasks like the evaluation of speech recognition fidelity (e.g., [Terpstra et al., 2023], which compared the effects of different segmentation methods) and for studying interactive behavior

and communication, such as modeling the evolution of group common ground over time, a la [Clark and Carlson, 1981]. AI researchers will benefit from the rich multimodality it offer, for instance, object and action detectors or gesture recognition algorithms (e.g., [VanderHoeven et al., 2023]) can be developed and trained using Kinect data. By evaluating important multimodal features of collaborative group interaction in context, the various modalities can operate as signals to an interactive AI agent that supports facilitators and scales up collaborative group activities (cf. [Bradford et al., 2023]).

Early fusion might benefit from synchronized Kinect along audio streams; hybrid fusion can leverage partially aligned modalities like gestures and speech; late fusion could combine CPS annotations with NICE features.

While the Weights Task Dataset provides a rich multimodal setting, it also comes with practical limitations that influence model generalization and evaluation. This dataset was collected as a starting point to build a functional proof of concept, it only contains 10 videos, making it easy for a ML model to overfit, especially when it presents some data imbalance issues. Indeed, since there's only the Weights Task to be completed by the participants, it can be observed that certain gestures might be more frequent than others. Another problem is the diversity in educational background of the participants, as they have all been recruited from CSU's Computer Science Department. Additional limitations may emerge as the dataset evolves, but for now, this dataset will be more than enough to answer the research questions proposed above.

# Chapter 4
# Real-Time Common Ground Tracking

When examining interactions in collaborative problem solving tasks, a variety of interpretable communication modalities are likely to be present. These modalities may encompass speech, gestures, actions, emotions, body posture, and the positioning of objects in physical space, among others. As AI increasingly becomes prevalent in everyday applications and diverse learning settings, such as educational institutions, there exists an opportunity for it to enhance our understanding of how small groups collaborate to work on CPS tasks. The design of interactive AI to facilitate CPS necessitates the development of a system that accommodates multiple modalities. In this chapter, I address the significance of the chosen multimodal features in modeling CPS, the necessity for different modal channels to interact within a multimodal AI agent that can assist with a broad spectrum of tasks, and the design considerations that must be contemplated when constructing such a system to effectively engage with and support small groups in successfully completing CPS tasks. Additionally, I present various tool sets that can be utilized to enhance each individual feature and their integration, along with potential applications for such a system. The design and implementation of this system were the result of a collaborative effort between the SIGNAL and VISION laboratories at Colorado State University. The efforts was distributed as would suggest the authors' orders in the publications that were produced; [Bradford et al., 2023], [Venkatesha et al., 2024], [Khebour et al., 2024b], [VanderHoeven et al., 2024], [VanderHoeven et al., 2025] and [Venkatesha et al., 2025].

This chapter addresses the research questions in chapter 1 by proposing a model designed to use multimodal signals. To better understand how these signals contribute to shared understanding, the following sections first examine the role of nonverbal indicators before introducing the mechanisms through which these modalities are fused within the system.

## 4.1 Nonverbal Indicators

While verbal communication is pivotal in collaborative problem solving (CPS), a significant portion of the communicative intent during group interactions is transmitted through nonverbal means. Gestures, body posture, eye contact, facial expressions, and the manipulation of objects within a shared physical environment frequently provide essential insights regarding

attention, intention, uncertainty, disagreement, or mutual understanding—sometimes even more consistently than verbal language. In group collaboration, such nonverbal indicators are necessary to synchronize actions, negotiate meanings and establishing common objectives among coworkers.

AI systems are slowly starting to integrate into collaborative environments, as suggested the literature review in chapter 2. There is still opportunity to expand this angle of research, creating agents that do not merely analyze speech in isolation rather interpret the comprehensive spectrum of communicative signals, specifically nonverbal ones. In this chapter, I propose an advanced AI agent that analyzes human communication, promoting natural group dynamics by acknowledging the nuanced nonverbal signals that are fundamental to effective collaboration. Instead of delivering direct solutions, this agent can observe how individuals communicate—through gestures, movements, and interactions with objects—and aid in fostering productive, self-directed problem solving within the group.

For an AI system to effectively interpret signals, it is crucial to incorporate information from various modalities and to reduce the possible confusion that may arise when signals are analyzed independently. Nonverbal behaviors are vital in augmenting verbal communication, as they emphasize important messages and reflect changes in group focus. Therefore, it is vital to make deliberate design choices concerning the selection, extraction, and integration of these various modalities. Such decisions influence how an AI agent perceives and represents the changing knowledge and emotions of participants, ultimately promoting a more human-like comprehension of group dynamics. This study uses the Weights Task Dataset in conjunction with its multimodal system to illustrate this goal: the objective is to create agents capable of engaging in collaborative tasks by comprehending both the verbal exchanges of participants and their physical and social interactions.

## 4.2 Common Ground Structure

In this discourse, I examine the framework of a multi-participant, task-oriented dialogue that includes communication via diverse content-generating modalities, such as language and gestures, in addition to mutually interpretable non-verbal actions (e.g., behaviors) [Kruijff et al., 2010, Pustejovsky and Krishnaswamy, 2021]. To enable this, it is crucial to create a data structure that reflects the common ground within this context, which can be dynamically updated during the conversation.A variant of the Dialogue Game Board (DGB) is used, as outlined in [Ginzburg, 2012]. Given the fluid and dynamic characteristics of co-interactive dialogue and situated actions, the approaches of [van Benthem et al., 2014] and [Pacuit, 2017] inspired the implementation of an *evidence-based model* of belief. In this model, the commit-

ments to propositions that describe situations or facts are not simply binary; rather, they are graded, allowing them to either weaken or strengthen based on the available *evidence* as the dialogue unfolds.

Let the minimal structure of a task-oriented interaction as a sequence, $D$, of dialogue steps, such that each move in the dialogue changes its situation or state.

Let $P = \{p_1, p_2, p_3\}$, be the participants in our dialogue. From any situation $s_k$, we define a $D$ move, $m_i$, as $m_i = (p_j, C_j, s_{k+1})$: participant $p_j$ performs a communicative act $C_j$, bringing the multimodal dialogue into situation $s_{k+1}$. The $D$ can be defined as the sequence of these moves: $D = m_1, \ldots, m_n$.

The objective is to monitor the situational content arising from each move: the collection of propositions that reflects the present state of the world, the ongoing advancement towards a goal, or the condition of a task. Furthermore, it encompasses the current inquiries being discussed and the beliefs held within the dialogue.

In light of these factors, three elements essential for monitoring shared understanding in conversation are recognized: a basic static framework representing levels of belief; a data structure that differentiates the components of the agents' shared understanding that are being monitored; and a dynamic process that refreshes this structure when new information and evidence become accessible to the agents. We will examine each of these components sequentially below.

[Pacuit, 2017] presents a model for evidence-based belief, wherein agents gather evidence supporting a proposition, $\varphi$, and can ultimately come to believe $\varphi$. We utilize a streamlined version of the evidence-based Dynamic Epistemic Logic (EB-DEL) as articulated in [van Benthem et al., 2014] and [Pacuit, 2017]. We characterize a model as a tuple, $\mathcal{M} = (W, E, V)$, where

(1) a. $W$ is a non-empty *set of worlds*;
    b. $E$ is an *evidence relation*;
    c. $V$ is a *valuation function*.

Let $E(w)$ signify the set $\{X \mid wEX\}$, which includes the worlds accessible to $w$ via the evidencing relation, $E$. The evidence-based epistemic language, $\mathcal{L}$, comprises the set of formulas produced by the following grammar:

(2) $p \mid \neg\varphi \mid \varphi \wedge \psi \mid [E]\varphi \mid [B]\varphi \mid [A]\varphi$

As shown in (2), the language $\mathcal{L}$ can be built from six types of expressions. $p$ means that a proposition can be atomic which represent simple facts about the world. More complex statements can be formed. A proposition ($\varphi$) can be negated ($\neg\varphi$) or can be a conjunction

$(\varphi \wedge \psi)$. $[E]\varphi$ expresses an agent having evidence for $\varphi$, while $[B]\varphi$ expresses the agent expressing belief in that proposition, whereas $[A]\varphi$ expresses that all the agents have evidence for or believe $\varphi$.

We differentiate the scenario in which an agent possesses "evidence in favor of" a proposition $\varphi$, denoted as $\varphi$. Given that an agent may hold evidence for propositions that present contradictory information, it can contemplate both $[E]\varphi$ and $[E]\neg\varphi$. This situation reflects an agent having multiple neighborhoods, $X$, each evidenced in a distinct manner by $w$. Nevertheless, we should consider the set of non-contradictory worlds as a unique subset of $X$, which possesses what [van Benthem and Pacuit, 2011] describes as the *finite intersection property (fip)*. This property enables us to identify a neighborhood of accessible worlds characterized by non-contradictory propositional content. When this condition is met, we assert that an agent holds *belief* in a proposition, $[B]\varphi$. Ultimately, the universal modality is regarded as "knowledge" of a proposition, $[A]\varphi$.

Capturing situational state information within a task-oriented dialogue is essential for accurately representing the current common ground and for anticipating future dialogue actions [Traum and Larsson, 2003, Schlangen and Skantze, 2011, Zhang et al., 2020b, Jacqmin et al., 2022]. For the purposes of this discussion, we utilize the concept of a *Dialogue Game Board* [Ginzburg, 1996, 2012], which has been adapted to account for the varying levels of evidence related to the propositions being discussed. A Common Ground Structure, denoted as $cgs$, is defined as a triple, $(QB, EB, FB)$, defined as such:

(3) a. Questions Under Discussion (QBank): set of topics or unknowns that need to be answered to solve the task;
b. Evidence (EBank): set of propositions for which there is some evidence they are true;
c. Facts (FBank): set of propositions believed as true by all participants.

The process commences with a collection of unknowns known as the "Questions under Discussion" (QUDs). In this implementation, we construct a finite model that encompasses a finite model of questions. For every object within the domain pertinent to the task, questions are formulated for each relationship involved in the task concerning that object. For instance, in the Weights Task, the objective is to ascertain the weights of five distinct blocks, followed by determining the algebraic relationship among them, specifically the Fibonacci sequence. The weight of each block varies from 10 to 50 grams, in increments of 10 grams. Therefore, for each block in $B$, where $B = \{red, blue, yellow, green, purple\}$, we have five potential values, articulated as yes/no questions. Consequently, the initialization of the QBank yields the following set:

(4) QBank = $\{Eq(r,10)?, \ldots, Eq(r,50)?, \ldots, Eq(p,10)?, \ldots Eq(p,50)?\}$

At the beginning of the discussion, both EBank and FBank start as empty sets, since no task-related propositions have been recognized or accepted as commonly evident.

## 4.3 Closure Rules

In light of the epistemic logic discussed previously, we present the mechanisms that modify the information state during a dialogue. Building on the work of [Plaza, 1989] and the later advancements in Public Announcement Logic [Baltag et al., 2016], we propose a novel operator for the model, known as the announcement operator, !. Public announcements are declarations made to all agents, and following such an announcement, every agent is aware that the statement has been made and that it holds true.

If $[!\varphi]$ represents the act of announcing $\varphi$, then $[!\varphi]\psi$ means "after $\varphi$ is announced, then $\psi$ is believed to be the case."

In order to distinguish evidence for $\varphi$ from belief in $\varphi$, we relativize the impact of a statement to the context within which it is uttered. Let us interpret $[!\varphi]\psi$ as follows.

(5) a. *Update with Evidence*:
$[!\varphi][E]\psi$: Given the announcement of $\varphi$, there is evidence for $\psi$;

b. *Update with Belief*:
$[E]\varphi \rightarrow [!\varphi][B]\psi$: Belief in $\varphi$ is conditionalized on $\varphi$'s announcement in the prior context of evidence for $\varphi$.

Semantically, an update represents the state of affairs after an announcement. This entails transforming the current model by removing all states where the announced formula is false. With evidence distinguished from belief/knowledge, we also update the evidence function, where $[!\varphi]$:

(6) a. Updates the worlds: $W' = W \cap \varphi$
b. Updates the Evidence function: $E'(w) = E(w) \cap \varphi$
c. $(M, w) \models \varphi$ implies $(M|_\varphi, w) \models [E]\psi$

This update fundamentally alters the foundational evidence sets. The announcement is regarded as a form of direct evidence. Therefore, in order to acknowledge that the announcement of $\varphi$ transforms into evidence rather than mere belief, the evidence sets for each agent are constrained (or revised) to represent the scenarios in which $\varphi$ holds true. Consequently, the belief function will inherently modify in response to the updated evidence sets.

Operationally, once (5a) is executed, the model is adjusted to reflect evidencing neighborhoods where $\varphi$ holds true. The proposed approach identifies 3 types of dialogue moves that concern the task of common ground tracking with the WTD; STATEMENT, ACCEPT, and DOUBT. A STATEMENT is the announcement of an evidence $\varphi$. An Accept is an agreement with evidence $\varphi$. A DOUBT is a disagreement with evidence for $\varphi$. These move dialogues will be the tools to update the common ground structure. As mentioned above, the structure initializes with EBank and FBank as empty while the QBank contains all possible questions for the participants to answer. By the end of the task the FBank must contain all correct propositions, while both QBank and EBank are empty. The updates are as such: A STATEMENT will take a proposition from the QBank to the EBank; when a participant announces a proposition, then there's evidence supporting it. When a dialogue has an ACCEPT move type, the proposition announced is moved from the EBank to the FBank. If an utterance has a DOUBT, then the proposition is stripped of the evidence that was backing it, this moves a proposition from the FBank or the EBank back to the QBank.

## 4.4 Speech Transcription

Speech is essential for effective communication in group settings, allowing participants to share insights, ask questions, discuss results, and formulate strategies. It lays a strong groundwork for managing group activities. Research indicates that speech is a key component in frameworks that examine group dynamics [Bradford et al., 2023, Stewart et al., 2021]. When combined with other elements, spoken language can offer important context about how participants interact with one another. Advanced Automatic Speech Recognition (ASR) technologies, like Google ASR and Whisper ASR [Radford et al., 2022, Velikovich et al., 2018], can accurately segment and transcribe audio into clear speech. Choosing between automatic and manual segmentation is a crucial decision, as it greatly affects the precision of the interpretations made from the speech data [Terpstra et al., 2023]. For real-time assistance, an agent relies on an ASR system that must effectively diarize and transcribe speech to function properly. As will be explained further more in the next sections, the system being studied runs in real-time, so the need to find a reliable, fast, and accurate ASR and speech transcriber model is present. Faster-Whisper was a good solution for this specific deman, as it is a fast, lightweight inference wrapper around OpenAI's Whisper model.

## 4.5 Gestures

Gestures are very important modalities in communication. They are mostly used as complementary to speech to remove ambiguity. Some gestures referred to as deictic are used to indicate locations when speech is not enough or when it requires too many words to translate specific details. In collaboration, gestures are commonly interpreted as indicators of engagement, and it shows through the weights task dataset. Particularly, pointing gestures are very present, often used along demonstratives words like "this" "that". Ignoring this modality can create misunderstandings as participants will not be able to establish common ground. But a gesture can have multiple meanings, and their interpretation is subjective, which is why a structure representation is required to maintain semantic fidelity especially on a computational level.

Gesture Abstract Meaning Representation, or **GAMR**, is a method to encode the meaning of gestures in interactions where multiple agents use different modes of communication. It builds on Abstract Meaning Representation (AMR), using its graph structure and how it represents actions and their parts. Gesture AMR was created to show how gestures carry meaning on their own and also work together with speech. It also explains how the meaning of a gesture changes over time and depends on the situation.

Gesture AMR identifies four main types of referential gestures: *iconic*, *deictic*, *metaphoric*, and *emblematic*. These categories are based on research from several studies [Kendon, 2004, Kong et al., 2015, Mather, 2005, McNeill, 1992]. Since our data is focused on gestures that happen during task-based activities, most gestures show the physical features of things or actions, like the shape of an object or how an action is done. Like the findings in another study [Brutti et al., 2022], metaphorical gestures are not very common in this kind of setting.

GAMR uses a system that can note gestures that fit into one or more of these categories. This allows for more detailed descriptions of different gestures that people might use when working on various CPS tasks. The inset shows how a "gesture unit" is structured, including both pointing (deixis) and symbolic (iconic) parts. In this system, `ARG0` is the person making the gesture, `ARG1` is what the gesture is about, and `ARG2` is who the gesture is meant for. These parts exist for each gesture subsection in the annotation.

For a gesture recognition task, multiple solutions are available [Fan et al., 2021, Hara et al., 2018, Narayana et al., 2018, Tong et al., 2022], but most of these solutions fail at the same challenge, recognizing gestures at unusual angles, or gestures that are too far from the camera, and fine-tuning the state of the art models can be useless as these models have been heavily trained on data that doesn't address these issues. This calls for a solution more adapted for our dataset, and that can be flexible to minor changes in its training data to

answer some of the more specific issues that the weights task dataset and that the task at hand require.

Table 4.1 shows the distribution patterns across gesture type, ARG0, and ARG1 in the WTD. These patterns strongly justify the inclusion of GAMR as input for the move classifier. Indeed, gestures not only carry strong semantic meaning, but they also show a correlation with conversational moves. For instance, deictic gestures (deixis-GA) – typically pointing gestures – overwhelmingly co-occur with STATEMENTS (≈92%), which suggests that participants often use pointing to anchor verbal references to objects or locations during propositional moves. Conversely, iconic and emblematic gestures show more balanced distributions between Accept and Statement, showing that they accompany agreement-related utterances. Similarly, ARG1 values reveal strong grounding in object- and action-related semantics: entities like blocks, colors, or "put" dominate Statements, mapping onto task-oriented actions. Additionally, the occurrence of head gestures ("yes"/"no") corresponds closely to Accept/Doubt categories, reinforcing the multimodal coupling between gesture and move type. Overall, the clear, intuitional and systematic alignment between gesture semantics, and conversational grounding behaviors underscores the value of incorporating GAMR representations to capture the multimodal regularities that a text-only model would likely miss.

From past findings relative to gesture semantics, we see a tradition of modeling a gesture into phases, mainly pre-stroke, post-stroke and the stroke [Arnheim, 1994, Kendon, 1980, Lascarides and Stone, 2009]. Based on this, I use a gesture recognition pipeline previously developed by [VanderHoeven et al., 2023], with the goal to streamline the detection of complex gestures, for a faster deployment in real time. This pipeline is built using hand detection tools like MediaPipe [Zhang et al., 2020a] for joint location detection, which extracts 21 joints in 3D coordinates from an individual's hands in a frame. The pipeline has three main parts. The first is a static classification model that identifies the basic shape of a gesture during any "hold" phase. The second is a movement segmentation algorithm that watches how hands move over time and divides a video into parts based on changes in movement. The third part breaks down phases using the results from the first two steps to find parts of the video that are in a "hold" phase. The start and end of these hold parts are marked as "key frames", and these frames are the most meaningful frames of a gesture.

Each static classification model can be trained for different gestures that are useful for CPS tasks, making the system flexible and detailed. This pipeline was used to detect various types of complex gestures, including small, subtle hand movements called *microgestures* [Wolf et al., 2011, VanderHoeven et al., 2023, Kandoi et al., 2023], as well as deictic gestures [VanderHoeven et al., 2024].

Figure 4.1 shows how this recognition model is used to detect pointing in the WTD. In the figure, participant 1 points at the blocks on the scale. A pointing frustum is built around the vector extended out from the pointer's index. This structure narrows down the blue block as the target of interest, which can be confirmed by checking with the GAMR annotations at the same time interval.

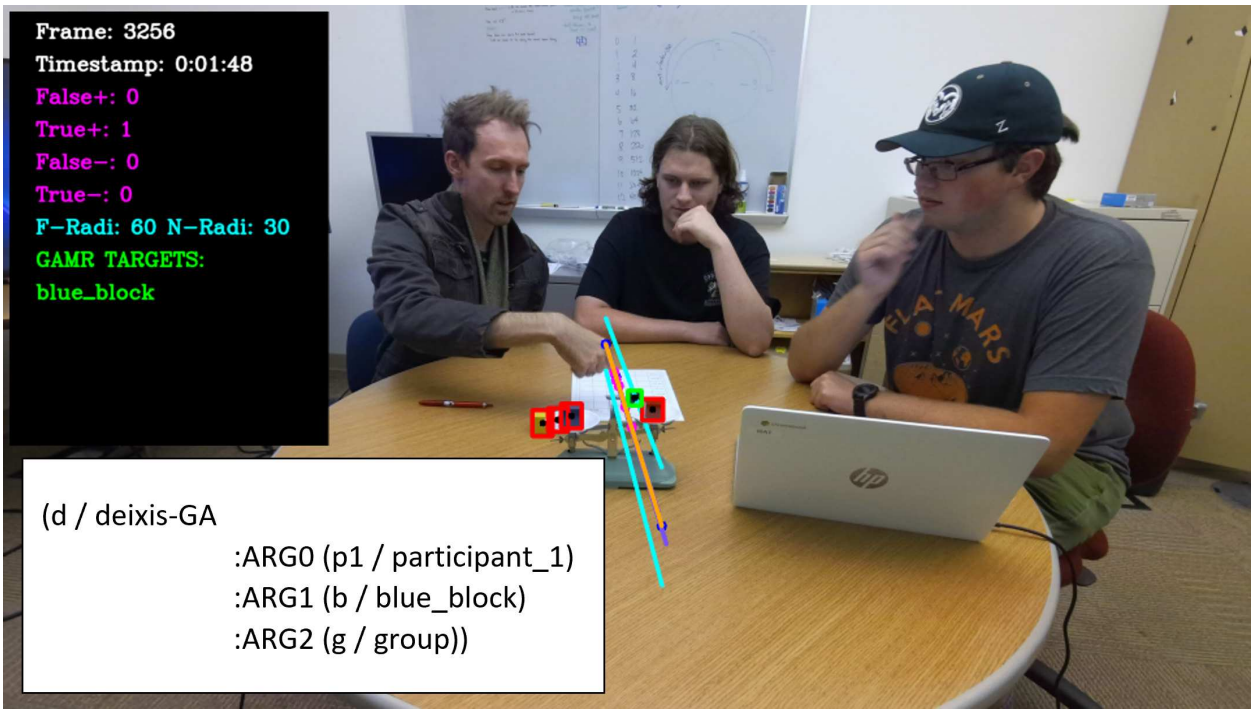



Figure 4.1: Group 1 deixis with GAMR example (reproduced from [VanderHoeven et al., 2024])

In the original CGT task, gesture features were encoded using $k$-sparse representations. We replace these with embeddings generated by an attention-based graph encoder-decoder architecture (see Fig. 4.3). The gesture type itself serves as the root node, while the argument values act as leaf nodes (see Figure 4.2). Each leaf node is connected to the root node through bidirectional edges, allowing the leaf nodes to learn not only from the root but also from their neighboring nodes. This bidirectional connectivity ensures that the embeddings effectively capture both local and global dependencies within the graph, enhancing the representation of gesture semantics.

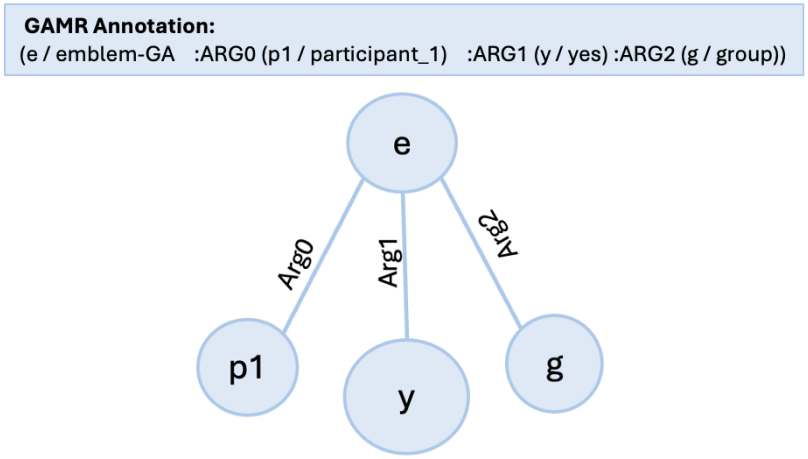



Figure 4.2: A GAMR annotation represented as a structured semantic graph.

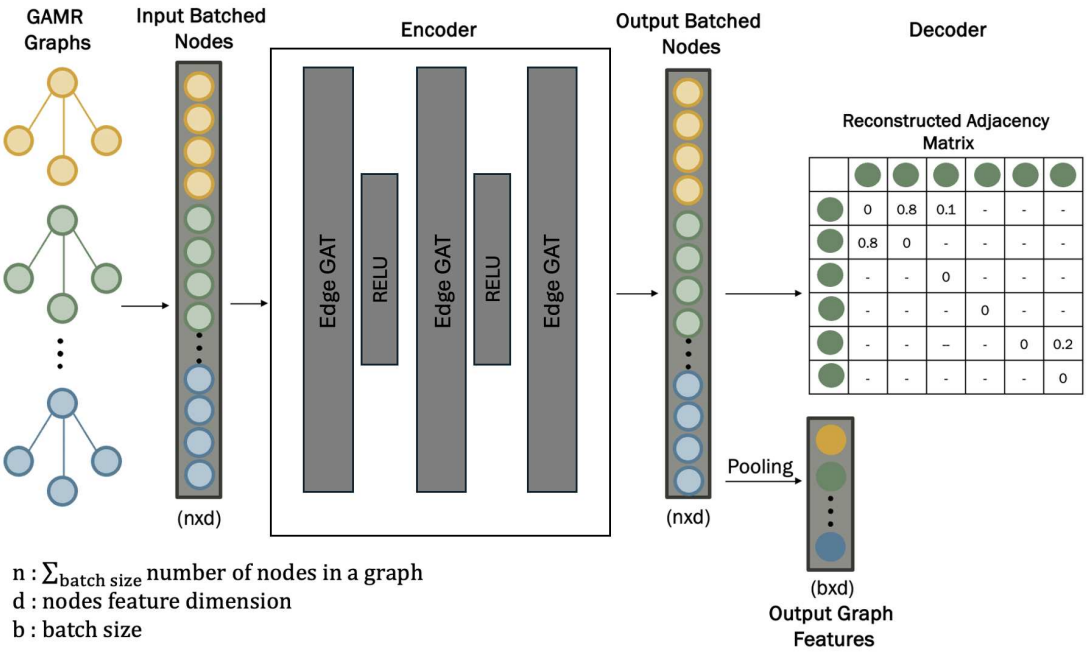



Figure 4.3: Attention based graph encoder-decoder architecture.

We adopt the attention-based message passing mechanism, EdgeGAT, from [Zhang and Ji, 2021] to construct the encoder. For each node in the graph, attention scores are computed for all neighboring nodes by concatenating node and edge features, passing them through a fully connected layer, and applying Leaky ReLU followed by a softmax function. The neighborhood information is then aggregated using these attention scores, normalized, and

combined with the original node feature, weighted by a parameter $\lambda$.

The encoder consists of three layers of EdgeGAT, each followed by ReLU activation except for the last. The model processes nodes in batches while retaining their graph membership. This ensures that node embeddings are computed jointly but still associated with their respective graphs, allowing for meaningful graph-level representations.

The decoder reconstructs the adjacency matrix $A$ from the learned node embeddings, where the reconstructed adjacency matrix is given by:

$$\hat{A} = \sigma(ZZ^T), \tag{4.1}$$

where $Z$ is the matrix of node embeddings from the final EdgeGAT layer, and $\sigma(\cdot)$ is the sigmoid activation function.

The model is trained using leave-one-out cross-validation with an edge-based loss formulation. We treat observed edges as positive examples and randomly sample non-existing edges as negative examples. The reconstruction loss is defined as the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{|E^+|} \sum_{(i,j)\in E^+} \log \hat{A}_{ij} - \frac{1}{|E^-|} \sum_{(i,j)\in E^-} \log(1 - \hat{A}_{ij}). \tag{4.2}$$

Here, $E^+$ represents the set of positive edges (existing connections), and $E^-$ denotes the set of sampled negative edges (non-existent connections).

During evaluation, we obtain the GAMR feature representation by aggregating node embeddings via average pooling:

$$g = \frac{1}{|V|} \sum_{i\in V} h_i, \tag{4.3}$$

where $V$ denotes the set of nodes in the GAMR graph and $h_i$ denotes the embedding of the $i$-th node. This pooled graph-level representation serves as the final feature vector for downstream multimodal learning tasks.

## 4.6 Object Detection

In the context of the Weights Task, knowing where the objects, or in this case the blocks, are is evidently important. For the participants, being able to see which blocks are available, which ones are on the scale and which are on the table makes a big difference in understanding the task progression. For the common ground tracker, the objects' positions is relevant to

not only track which objects are in action, but also which ones are being pointed at by the participants.

The location of an object in 3D space can be figured out by predicting its 6DOF pose, which means knowing how it translates and rotates in all three orthogonal directions [Hu et al., 2020, Labbé et al., 2020, Wang et al., 2021]. Being able to track the object's position over time helps understand how it interacts with its surroundings, especially if those interactions change how the object is arranged or behaves. Usually, this information can be gathered from regular video images, but in this case, we also use depth information from Azure Kinect devices, to get better and more accurate results.

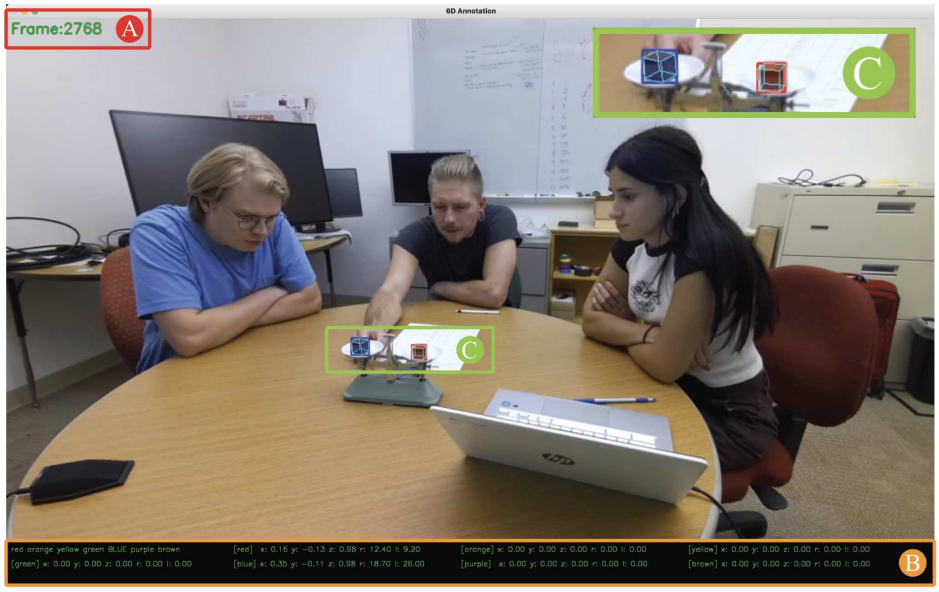



Figure 4.4: 6DOF Pose Annotation Tool on WTD. *A* shows the current frame number, *B* shows the position and rotation information for each object of interest, and *C* (expanded in inset) shows annotated 2D and 3D bounding boxes.

In common ground tracking, 6D pose estimation meets some difficulties in visual feature extraction. The camera placement is one of them. Indeed, to record enough information the Kinects must be placed further away from the blocks than they typically are in 6D pose estimation tasks and datasets [Rennie et al., 2016, Tyree et al., 2022]. This bigger distance is required to capture the participants and their interactions. The tradeoff is that the objects appear too small for not only the annotation process but also for a typical object detection model, as shows the figure 4.4.

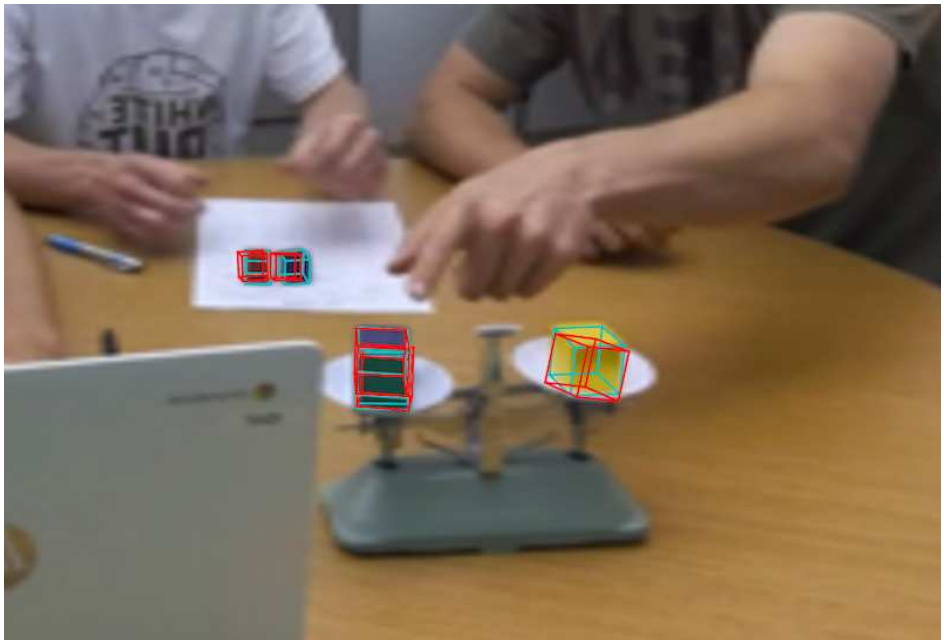


Figure 4.5: Ground truth object bounding boxes (blue) and predicted bounding boxes (red). Deixis is used to select a spatial region containing one or more objects, which may be further disambiguated by contemporaneous speech or prior context.

Figure 4.5 illustrates the convergence of pointing and object detection. In the context of automated object selection utilizing deixis, an end-to-end solution necessitates the simultaneous automatic detection of objects within the scene alongside gesture recognition, as opposed to relying on pre-annotated bounding boxes (as depicted in Figure 4.1). Leading methodologies for such tasks are generally comprised of multiple modules, which are trained using a mix of real images and 3D renderings of the objects of interest across various orientations. A conventional approach may initiate with a convolutional neural network (CNN) designed to extract spatial information and visual features of the objects. These visual features are subsequently employed to predict the poses of the objects in the following module. After this, a 'refinement step' occurs, wherein the module estimates the object pose, and these estimates are utilized to render images of the objects, which are then juxtaposed with the actual training images. Errors are backpropagated until the renderings and real images converge within a suitably small epsilon.

The Common Ground Tracker's object detector is built using a FasterRCNN ResNet-50-FPN model [Lin et al., 2017]. Faster-R-CNN produces feature maps using a backbone network, in this case the ResNet-50 feature pyramid network. Subsequently, the region proposal network (RPN) which is a smaller convolutional network, traverses the feature maps to create bounding-box predictions. These RPN region predictions are then input into a Fast R-CNN [Girshick, 2015] detection network, which ultimately yields the predicted

bounding boxes. Alongside the bounding boxes, the model also provides confidence scores during inference for each predicted bounding box, and we select the bounding boxes with the highest confidence scores for each class. Given the RPN's capability to swiftly generate region proposals, Faster R-CNN is an appropriate model for real-time object detection.

This model was trained on block bounding box annotations from The Weights Task Dataset. This model was trained for 10 epochs, with a batch size of 32, input size of 3x416x416, using a Stochastic Gradient Descent, a learning rate of 1e-3, a momentum of 9e-1 and a weight decay of 5e-4.

After careful examination of the model's output, it was clear that it had some difficulties with certain scenarios. Issues with object occlusions, diversity in block positions and in light conditions, were mainly the focus of the next upgrade. In the WTD, participants were seen handling blocks in certain ways that occluded these objects enough for the object detector to fail to at predicting accurate boundin boxes. Blocks can easily be occluded by participants as well as by other blocks or the scale on the table. The object detector also showed weakness with frames were objects where shadowed when participants reached across the table. The color of the blocks get darker and the model was too sensitive to a variation of the lights. In the WTD, the objects usually start at the center of the work zone, which in this case is the table. But throughout the task, the objects can be moved around, and if they're placed at the edge of the table, the object detection model loses its positions.

Additional data was collected to augment the WTD to improve the object detection model's performance. The new data was collected on two separate occasions for different objectives. The initial set was aimed at assessing light conditions. It was collected under varying light conditions: full illumination, partial illumination, and natural light (Figure 4.6). Table 4.2 illustrates the quantity of frames collected under the different lighting scenarios. While the next set focused on occlusions arising from various interactions. This set, captured four distinct gestures to examine occlusions during interactions. The four gestures included those predominantly utilized in the collaborative task: pinching the center of the block, pinching the top of the block, covering the top of the block, and placing the blocks on palms. Table 4.3 presents the number of samples corresponding to the various gestures. Both sets comprise frames depicting the blocks positioned on the table alongside the scale, mirroring the actual Weights Task. The blocks-on-scale scenario presented numerous instances of occlusion caused by the horizontal and vertical stacking of blocks. Furthermore, we altered the placements of the blocks and scale in each instance to ensure diverse object positions within every frame. The supplementary data was annotated utilizing SAM2 within the CVAT annotation tool.

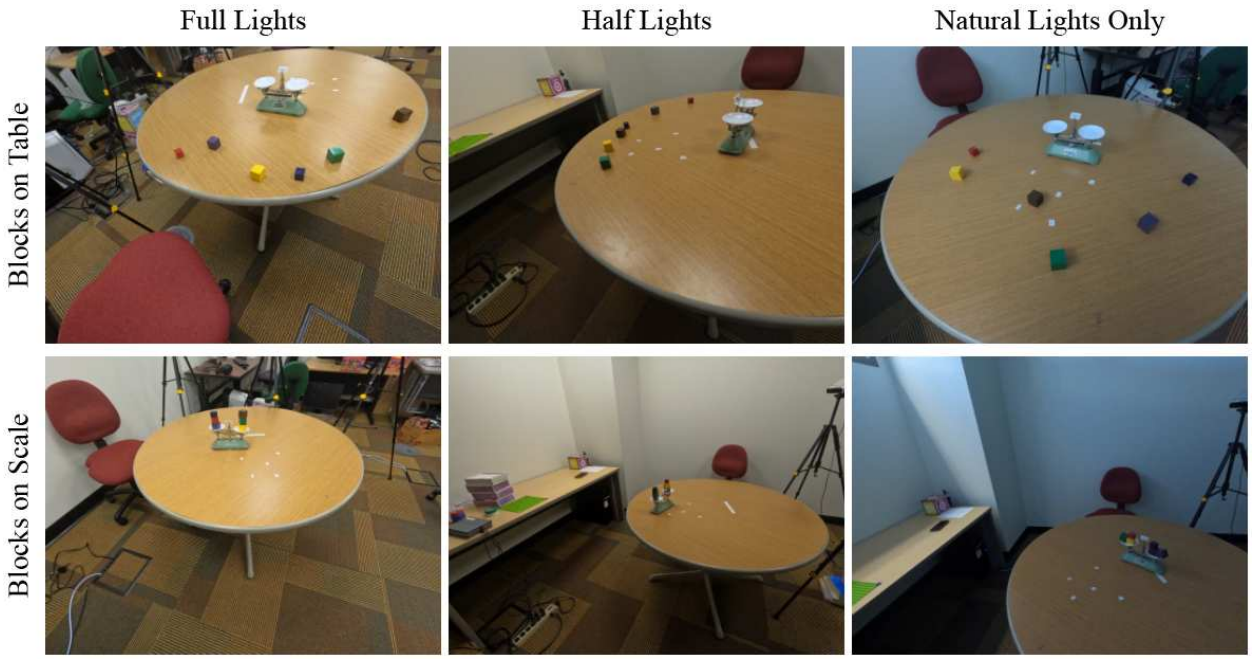



Figure 4.6: Additional data collection in variant light conditions.

With the new augmented WTD, the faster-RCNN model's training process had to be updated. The object detector was first trained on ImageNet-1K [Deng et al., 2009], then fine-tuned on 300,000 frames from the WTD, 5,428 frames from a private demo of the Weights Task, and on the newly captured frames from different light conditions, gestures, and block placements, mentioned above.

## 4.7 Dense Paraphrase

A significant number of demonstrative expressions and anaphoric references (such as "this", "that", "it", etc.) are involved in conversations during contextually shared tasks. The automatic interpretation of these terms often necessitates the use of an additional modality, like deictic gestures. As a method of interpretation, *Dense Paraphrasing* serves as a linguistically grounded strategy for textual enhancement that makes explicit the compositional operations that are typically omitted in the meaning of language. This approach generally encompasses three types of interpretive activities: (i) acknowledging the various linguistic forms that can correspond to the same fundamental semantic representation (paraphrases); (ii) pinpointing semantic elements or variables that are either present or assumed by the lexical semantics of the words in the text, which may include omitted, concealed, or implicit arguments; and (iii) analyzing or calculating the dynamic alterations that actions, events, and other communicative modalities exert on the objects mentioned in the text.

More formally, given the pair, $(S, P)$, where $S$ is a source expression (e.g., a textual narrative, image caption, or a speech transcription), and $P$ is a linguistic expression, we say $P$ is a valid *dense paraphrase* of $S$ if: $P$ is a lexeme, phrase, or sentence that eliminates any contextual ambiguity that may be present in $S$, but that also makes explicit the underlying semantics that is not (usually) expressed in the economy of sentence structure, e.g., default or hidden arguments, dropped objects or adjuncts. $P$ is both meaning-preserving (consistent) and ampliative (informative) with respect to $S$.

## 4.8 Prosody

Prosody is observed using what is referenced to as prosodic features. These non-linguistic features are extracted from the speech of each group's audio recording which is processed using openSMILE. These features relate to frequency, amplitude and balance of the speaker's voice. The extended feature set predefined by [Eyben et al., 2015]. This feature set is minimalist but very effective in capturing significant information. In total, this framework proposes 88 prosodic features for each utterance, including information like loudness and spectral flux.

## 4.9 Proposition Extractor

Propositions include the semantic content of an utterance that is relevant to defining the state of the task. For instance, one of the participants may have a relatively lengthy utterance that intends to propose a solution such as "I think the blue block weighs the same as the red block is it's also equal to 10 grams", in this case the proposition expressed is $blue = 10$. A key problem in propositional extraction from natural dialogues is that people may have radically different ways of expressing the same underlying semantic content—they use filler words, disfluencies, and have different idiosyncracies and preferences for expressing certain content. This problem was previously addressed by [Venkatesha et al., 2024], and we use their system as our propositional extractor. We also use the information of presence or absence of a proposition as a feature for one of our model's components (the move classifier—see below) as we observed a high correlation between the feature and the type of move an utterance contains (STATEMENT, ACCEPT, DOUBT).

Propositions are primarily extracted from speech transcriptions, as the semantic content mostly resides in the words spoken. However, like CPS facets, non-verbal features play a role. Due to the situated nature of the Weights Task, a lot of information is expressed using aligned speech and gestures—specifically demonstrative pronouns and deictic gestures

(pointing). For instance, *green* = 20 might be expressed by the utterance "I think *that one's* 20 grams" while pointing to the green block. In this case, the transcribed speech alone will not enable the model to recognize which block the participant is referring to. Thus we use a *dense paraphrasing* procedure [Tu et al., 2024] which decontextualizes the reference by rewriting it with explicit information from other modal channels. Under this transformation, "I think *that one's* 20 grams" plus pointing at the green block gets rewritten to "I think *[green block]'s* 20 grams."

## 4.10 CPS

Collaborative Problem Solving (CPS) facets are a way of representing different dimensions of a group's interaction as they contribute to successful problem-solving in a team setting. We use the framework by Sun et al. [Sun et al., 2020a]. Specifically, we used as features the CPS *facets*, or the highest level of this framework's hierarchy. These facets include *constructing shared knowledge*, *negotiation/coordination*, and *maintaining team function*. Successful exhibition of these facets in the course of an interaction are assumed to facilitate the exchange of information, align team efforts, and ensure that all members' opinions are considered, thus enhancing team function, encouraging collective understanding, and enabling teams to tackle complex tasks which would have been challenging for individuals to solve alone. A crucial aspect of CPS is that certain facets may be expressed non-verbally. Intonation, facial expressions and body language all play significant roles in conveying intent, emotion, agreement, confusion or other significant indicators of the current state of the collaboration. These non-verbal cues support verbal communication, making interactions richer, which improves group performance. In this work, we use speech and prosody as input for a random forest model, as first reported in [Bradford et al., 2023], to infer the presence or absence of CPS facets from an utterance.

## 4.11 Move Classifier

The move classifier is a multimodal LSTM-based model, intended to capture contextual information that conditions the sequence of cognitive states in a dialogue. Each utterance, including a prior context of $w = 3$ previous utterances, was processed through a linear layer (512 units) followed by ReLU activation and an LSTM block of 512 units. The final hidden states of the LSTM block for each modality of interest were concatenated and passed through a 512-unit linear layer, *tanh*, another 512-unit linear layer, and SiLU before the classification layer. Fig. 4.7 shows this architecture.

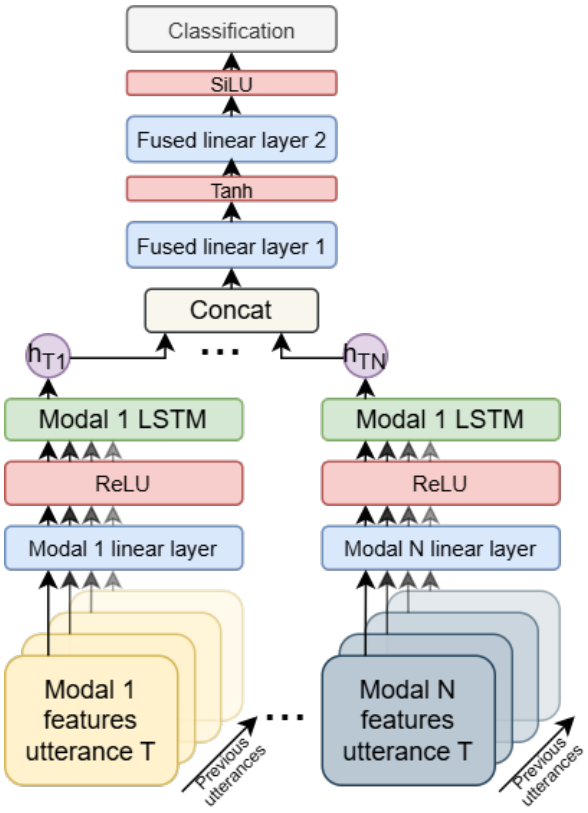



Figure 4.7: Move classifier architecture.

I optimized for the detection of *STATEMENT*, *ACCEPT*, and *DOUBT*. To alleviate imbalance during training, we augmented the data with SMOTE [Chawla et al., 2002]. I trained using Kaiming initialization with a uniform distribution [He et al., 2015]. All layers except the classification layer are trained using a triplet loss with a margin of 1 [Balntas et al., 2016] for 200 epochs and a learning rate of $10^{-4}$. Subsequently the entire model was trained using cross-entropy loss and a learning rate of $10^{-3}$ for 100 epochs, and for 200 further epochs with a learning rate of $10^{-4}$. Evaluation was done using a 10-group-fold validation. Meaning that 10 instances of the model were created, each trained over 9 groups from the dataset, and evaluated on the remaining group which is different for each instance.

The move classifier can be used with many modalities, including, speech transcription BERT embeddings, prosodic features, GAMR annotations, actions annotations and presence/absence of a proposition for the utterance of interest. [Khebour et al., 2024b] has shown how some groups of participants have been more expressive than others using different communication channels like gestures and body pose. This supports the diversification of input modalities for the move classifier, and also for the tracker.

## 4.12 TRACE

Trace is a modular framework that integrates features from speech, acoustic, RGB, and depth channels to analyze the linguistic and nonverbal behaviors of task participants, thereby modeling their shared task-relevant beliefs. All feature modules define an output *interface* or a class that represents the data type produced by a module. Additionally, modules specify zero or more input interfaces that are necessary for generating the output. For instance, the Propositional Extraction module requires solely text input, whereas the Dense Paraphrasing module necessitates text, gesture, and object inputs (Fig. 4.8). Trace facilitates modules in designating their input interfaces as dependencies, ensuring that the contents of the required output interface are automatically transmitted to the dependent input interface. Consequently, the entire system, or any system constructed using trace, can be organized as a directed graph, with features represented as vertices and edges linking a module to all of its dependencies. This architecture permits the interchange of various multimodal processing modules, enabling the creation of different variants of the system. The most important feature that TRACE adds on previous works is its real time processing. Running a real-time multimodal tracker presents several notable challenges. Several machine learning models, particularly those employed in NLP or CV, are high consumers in resources, and thus are generally optimized for offline inference. Which creates the first challenge, as some of the models used in the TRACE system can be limited by standard hardware and can lead to potential lag or frame drop, especially if they are ran in parallel and continuously. Secondly, multimodal inputs are received asynchronously and at different frequencies (for instance, audio at 44.1kHz, video at 30fps, and skeletal tracking at lower rates), which complicates real-time alignment and fusion. Thirdly, real-time systems must adeptly manage missing or noisy data—such as when gesture tracking temporarily fails or speech becomes unclear due to overlapping dialogue or background noise. TRACE answers to these challenges through meticulous modular optimization, which allows for selective computation, close integration between modules, and pipeline designs that reduce latency offering a consistent output rates. The system processes inputs at 6 FPS demonstrating the robustness of TRACE's design.

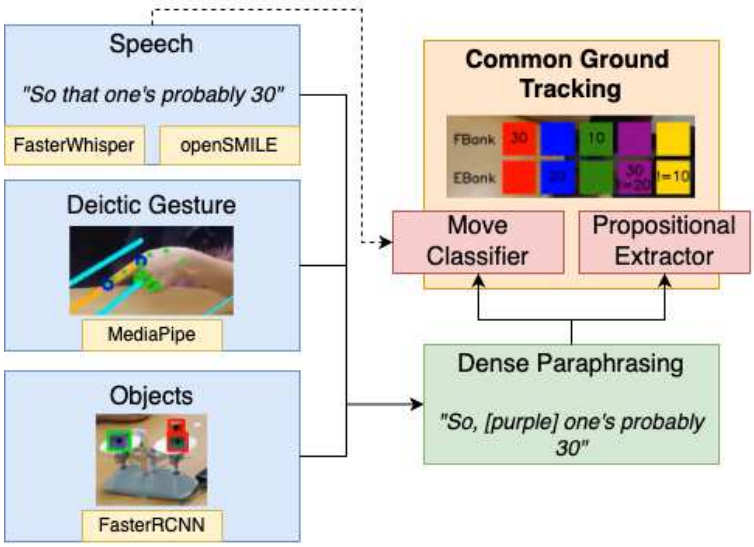



Figure 4.8: High-level schematic of information flow in real-time multimodal common ground tracking. We combine signals from speech, gesture, and objects in the environment to determine the task-relevant content being discussed, and the epistemic positioning expressed in each utterance. Logical closure rules unify these outputs into the set of common QUDs (QBank—not displayed for space reasons), pieces of evidence (EBank), and facts (FBank).

Table 4.1: Distributions of move types (STATEMENT, ACCEPT, DOUBT) across gesture types, ARG0, and ARG1. Values in parentheses indicate percentages.

| **Gesture Type** | **Accept** | **Doubt** | **Statement** |
|---|---|---|---|
| and | 1 (12.5%) | 1 (12.5%) | 6 (75.0%) |
| deixis-GA | 25 (6.7%) | 5 (1.4%) | 341 (91.9%) |
| emblem-GA | 12 (28.6%) | 0 (0.0%) | 30 (71.4%) |
| icon-GA | 4 (44.4%) | 1 (11.1%) | 4 (44.4%) |

| **ARG0** | **Accept** | **Doubt** | **Statement** |
|---|---|---|---|
| participant_1 | 20 (18.2%) | 2 (1.8%) | 88 (80.0%) |
| participant_2 | 10 (6.1%) | 3 (1.8%) | 150 (92.0%) |
| participant_3 | 12 (7.6%) | 2 (1.3%) | 143 (91.1%) |

| **ARG1** | **Accept** | **Doubt** | **Statement** |
|---|---|---|---|
| agree-01 | 1 (50.0%) | 0 (0.0%) | 1 (50.0%) |
| block | 11 (16.2%) | 1 (1.5%) | 56 (82.4%) |
| blocks | 1 (3.6%) | 1 (3.6%) | 26 (92.9%) |
| blue_block | 2 (6.7%) | 1 (3.3%) | 27 (90.0%) |
| brown_block | 0 (0.0%) | 0 (0.0%) | 13 (100.0%) |
| computer | 1 (11.1%) | 0 (0.0%) | 8 (88.9%) |
| cup | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| even | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| green_block | 1 (1.9%) | 0 (0.0%) | 51 (98.1%) |
| i-don't-know | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| i-dont-know | 1 (25.0%) | 0 (0.0%) | 3 (75.0%) |
| laptop | 0 (0.0%) | 0 (0.0%) | 3 (100.0%) |
| list | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| location | 0 (0.0%) | 1 (20.0%) | 4 (80.0%) |
| maybe | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| move-01 | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| mystery_block | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| no | 0 (0.0%) | 0 (0.0%) | 2 (100.0%) |
| paper | 2 (3.9%) | 0 (0.0%) | 49 (96.1%) |
| participant_1 | 0 (0.0%) | 1 (100.0%) | 0 (0.0%) |
| participant_2 | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| participant_3 | 0 (0.0%) | 0 (0.0%) | 3 (100.0%) |
| phone | 1 (25.0%) | 0 (0.0%) | 3 (75.0%) |
| purple_block | 5 (10.9%) | 1 (2.2%) | 40 (87.0%) |
| put | 2 (100.0%) | 0 (0.0%) | 0 (0.0%) |
| range | 2 (100.0%) | 0 (0.0%) | 0 (0.0%) |
| red_block | 2 (11.8%) | 0 (0.0%) | 15 (88.2%) |
| researcher | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| scale | 0 (0.0%) | 1 (12.5%) | 7 (87.5%) |
| stop-01 | 0 (0.0%) | 0 (0.0%) | 1 (100.0%) |
| worksheet | 0 (0.0%) | 0 (0.0%) | 2 (100.0%) |
| yellow_block | 0 (0.0%) | 0 (0.0%) | 38 (100.0%) |
| yes | 10 (34.5%) | 0 (0.0%) | 19 (65.5%) |

Table 4.2: Number of Frames on Various Light Conditions

| **Light Condition** | **Blocks on Table** | **Blocks on Scale** |
|---|---|---|
| Full Lights | 53 | 56 |
| Half Lights | 53 | 78 |
| Natural Light Only | 60 | 49 |

Table 4.3: Number of Frames on Various Gestures

| **Gesture** | **Blocks on Table** | **Blocks on Scale** |
|---|---|---|
| Pinch the center of block | 15 | 12 |
| Pinch the top of block | 20 | 9 |
| Cover the top of block | 14 | 14 |
| Put the blocks on palm | 13 | - |

# Chapter 5
# Multimodal Analysis

This chapter presents a multimodal evaluation of TRACE. I will test the limitations of this system by trying different combinations of input modals. One thing worth highlighting from the start of this chapter, is that the goal of this work is not to enhance the system's performance. The goal is rather to analyze the effects of different features, mainly the non-verbal ones, in knowledge tracking. I will enumerate the experimental design adapted for this analysis, as well as the evaluation process adopted and the results found after the fact. By systematically varying the combination of modalities and examining their performance across interaction phases, this analysis directly addresses the three research questions outlined in Chapter 1.

## 5.1 Experiments

The TRACE system offers a multitude of modalities. To better understand the effects of their presence within the live tracker, I have specifically designed a set of experiments that should provide strong insights about the limitations of multimodal machine learning in common ground tracking.

TRACE strictly needs speech as a modality since without it, it becomes impossible to keep up with the conversation held among the group members, and thus their shared knowledge. That is why all the experiments will run with speech even if not mentioned.

However, the speech transcripts can be automatically extracted from the voice recordings, as well as manually. The term *Ground truth Speech* will refer to switching the speech transcription into using the manually annotated data, as opposed to the automatically transcribed speech using FasterWhisper.

The same way I can switch the speech channel from using automatic to manual data, I can do the same for both gestures and objects. Indeed, the ground truth objects data consist of using manually annotated coordinates of the bounding boxes for the objects, namely the blocks in the WTD. The ground truth gestures data replaces the automatically extracted GAMR annotations obtained using MediaPipe's hand landmarks, the block positions, and the gesture detector, by GAMR data annotated by experts. Exploiting these ground truth data will help track the system's limitations, and how error can propagate from one end to

another.

Various components can be turned on and off in order to realize the experiments of interest. The first component is the dense paraphrase model. As a reminder, this component takes ambiguous words off a sentence and replaces them with appropriate wording while maintaining semantic fidelity. Turning off this model may imply a weaker understanding of the context, as some utterances would be left without enough information to comprehend the speaker's intention. This model uses speech transcripts, blocks positions, and hand landmarks for gesture recognition and point tracking. It's output is linked to the move classifier and the propositional extractor, making it a very important component of the TRACE pipeline. That is why, testing the sensibility of the system relative to the Dense Paraphraser provides a deeper grasp of the effects of the available feature set.

Other modalities can be removed and reinstated back into the system for a better analysis. These modalities are prosody, CPS, proposition presence, and GAMR annotations. These modalities are non-verbal, and are employed to fill in the gaps that text transcripts leaves behind. Prosody contains information relative to the voice tone and other cues. CPS showed correlation with move types, and so did presence of proposition, which is why they will be parts of the experiments.

The list of experiments consists of 9 tests where each uses a different combination of input channels to investigate how the model reacts. This list can be found in table 5.1, with the components used for each experiment.

Table 5.1: Experiments with additional modalities and evaluation features (speech transcripts feature is always included and so is not shown here, e.g., Experiment 1 is an automatic speech transcription-only baseline).

| **Experiment No.** | **Dense Paraphrase** | **Ground Truth Speech** | **Ground Truth Gesture** | **Ground Truth Object** | **Prosody** | **CPS** | **Proposition** | **GAMR** |
|---|---|---|---|---|---|---|---|---|
| **1** | | | | | | | | |
| **2** | | X | | | | | | |
| **3** | X | X | | | | | | |
| **4** | X | | | | | | | |
| **5** | X | | X | X | | | | |
| **6** | X | X | | | X | X | X | |
| **7** | X | | | | X | X | X | |
| **8** | X | | X | | X | X | X | X |
| **9** | X | | | | X | X | X | X |

## 5.2 Evaluation process

The Sørensen-Dice Coefficient (DSC) is used as the primary evaluation metric. DSC is an IoU-style (Intersection over Union) metric that normalizes for the sizes of the sets being compared. This is also the primary metric used in [Khebour et al., 2024b].

DSC, also referred to as the Dice coefficient, is widely used as a similarity metric to compare the overlap between two sets. It's defined as $\text{Dice}(A, B) = \frac{2|A \cap B|}{|A|+|B|}$. As can be observed, this coefficient gives more importance to the overlap in small sets, making it effective for tasks involving short texts or sparse features [Schütze et al., 2008]. It's bounded between 0 and 1, facilitating its interpretability and its adoption in different applications [Dice, 1945]. The coefficient is very efficient for binary or categorical data comparisons, such as in image segmentation and token-based similarity tasks [Zou et al., 2004]. It's equivalent to the F1-score in binary classification scenarios [Van Rijsbergen, 1979] as it balances precision and recall.However, the coefficient shows a few limitations. It does not account for true negatives, which may lead to misleading evaluations in imbalanced datasets [Powers, 2020]. It is very sensitive to small perturbations in small sets, which causes large changes in similarity with only slight differences [Bilenko and Mooney, 2003]. Additionally, it fails to capture structural or sequential relationships, making it less reliable in applications that depend on syntax or graph structures.

The test dataset is composed of 4 videos out of the 10 that exist in the Weights Task Dataset. These videos (Groups 1, 2, 4, and 5) contained ground truth annotations for all the relevant modalities, enabling a complete suite of experiments. I use a leave-one-group-out experimental format where models were trained over all but one group and evaluated on the remaining group. I calculate the average DSC for each one of the 4 test groups, then I compute the average across those 4 values.

The evaluation will be the comparison of different set from the Common Ground Structure, over the predicted and the ground truth. I use the Common Ground Annotation method (CGA) to obtain the ground truth data. The statement IDs from the common ground annotation are used to systematically align them with the propositions expressed in the utterance. During the move prediction, the alignment is referred to to retrieve the propositional content to be linked to the move during the common ground structure update. Since annotators had access to the video channel and all other modalities when annotating the propositions expressed, this method is a multimodally-informed method of propositional extraction. Then the closure rules are used to finalize the construction of the ground truth.

The results analysis, in the following section, focuses the comparison on F $\cup$ E. This set is more expressing, as the move classifier limits TRACE's performance. Indeed, the model

does not perform very well when it comes to distinguish doubts and accepts from statements. This is due to data imbalance in the WTD. This limitation causes the tracker to blur the FBANK-EBANK border.

## 5.3 Results Analysis

Let's start this section by going through the results of the object detector's performance on the augmented WTD compared to the base version. A test set consisting of 1,786 frames from a separate private demo of the Weights Task is used to measure the performance of the `fasterrcnn_resnet50_fpn`. For each fine-tuning stage, Table 5.2 shows the global mean average precision (mAP), mAP at an intersection over union (IoU) threshold of 0.5 ($mAP_{50}$), mAP at an IoU threshold of 0.75 ($mAP_{75}$), and mean average recall for the model's top 10 predictions based on the confidence scores ($mAR_{10}$). The mAP metric takes the mean of the average precision for each class at IoU thresholds $\in \{0.5, 0.55, \ldots, 0.95\}$. The base fine-tuning stage represents the model weights after being fine-tuned on the WTD and the private demo data, and the following stages (light conditions, gestures, and light conditions + gestures) are further fine-tuned using the base stage as a starting point.

Table 5.2: Faster R-CNN Fine-tuning Performance

| **Fine-tuning Stage** | **mAP** | **$mAP_{50}$** | **$mAP_{75}$** | **$mAR_{10}$** |
|---|---|---|---|---|
| Base | 0.3843 | 0.7107 | 0.3698 | 0.4385 |
| Light Conditions | 0.4626 | **0.7578** | 0.4976 | 0.5725 |
| Gestures | 0.4477 | 0.7565 | 0.4467 | 0.5767 |
| Light Conditions + Gestures | **0.5100** | 0.7472 | **0.5652** | **0.6337** |

The table 5.2 shows how, despite the small size of the new frames added to address the issues of prior version, there's is a very evident increase in performance. In fact, only the mean Precision with an IoU threshold of 50 isn't at its maximum with all the augmented data. But even in that case the additional frames for different light conditions improves compared to the base WTD. This is very important because it decreases the error rate propagated by the object detector, a good news since most of the experiments in table 5.1 are dependent of this model.

Table 5.3: Experimental results averaged across test groups. F ∪ E denotes the union of FBANK and EBANK Khebour et al. [2024b] and this serves as a proxy for extraction of the correct propositional content even if the level of evidence assigned to it is incorrect. Bold shows which feature set performed best for each bank.

| | **Exp 1** | **Exp 2** | **Exp 3** | **Exp 4** | **Exp 5** | **Exp 6** | **Exp 7** | **Exp 8** | **Exp 9** |
|---|---|---|---|---|---|---|---|---|---|
| Average QBANK DSC | 0.583 | 0.592 | **0.608** | 0.575 | 0.583 | 0.560 | 0.559 | 0.563 | 0.515 |
| Average EBANK DSC | 0.189 | 0.159 | 0.168 | 0.179 | 0.208 | 0.127 | **0.213** | 0.170 | 0.160 |
| Average FBANK DSC | 0.057 | 0.069 | 0.146 | 0.065 | 0.046 | 0.082 | 0.100 | **0.166** | 0.129 |
| Average F ∪ E DSC | 0.397 | 0.477 | **0.514** | 0.373 | 0.443 | 0.429 | 0.411 | 0.378 | 0.324 |

The results of the experiments in Table 5.1 can be found in Table 5.3, and the best results across all experiments from that table are compared to the CGT–the offline version–using the average DSC over the test set. Experiment 1 provides a baseline using only automatically-transcribed speech. Experiment 2 shows the maximum utility of speech alone, as it uses the ground truth data. Experiments 1 and 2 show that with automatic transcriptions (Experiment 1), the models can get almost .40 DSC for F ∪ E, and that is over 83% of the potential of speech when using ground truth.

Experiment 3, which uses ground truth speech and dense paraphrasing with automatically-detected gestures and objects, shows the maximum performance on QBANK and FBANK using that feature set and the utility of dense paraphrasing. However, Experiment 4 shows that noise in the automatic speech recognition does have an impact, and reduces performance evaluated on F ∪ E by about 27%.

With Experiment 5, we see that if we remove the ground truth of the speech and replace it with the ground truth for both the objects and gestures features, the model's performance slightly drops, thus proving that speech remains the most important feature, as the context and information it encompasses are far greater than what the pointing gestures and objects do alone.

Experiment 6 adds more non-verbal features linked to the move classifier, but with ground truth speech, while Experiment 7 uses the same features without the ground truth data speech. Here, performance is similar between the two, showing the impact that features like CPS facets, propositions, and prosody can have even with noise introduced by automatic extraction methods, in that they largely allow the model to close the gap with the ground truth induced by the automatic speech transcriptions. This shows a faster path for model optimization; while several works have shown that increasing training data size in AI models is crucial for performance increase, this result suggests that in a task such as CGT, increasing the number of available modalities may also help.

In Experiments 8 and 9 we introduce the GAMR representation into the move classifier.

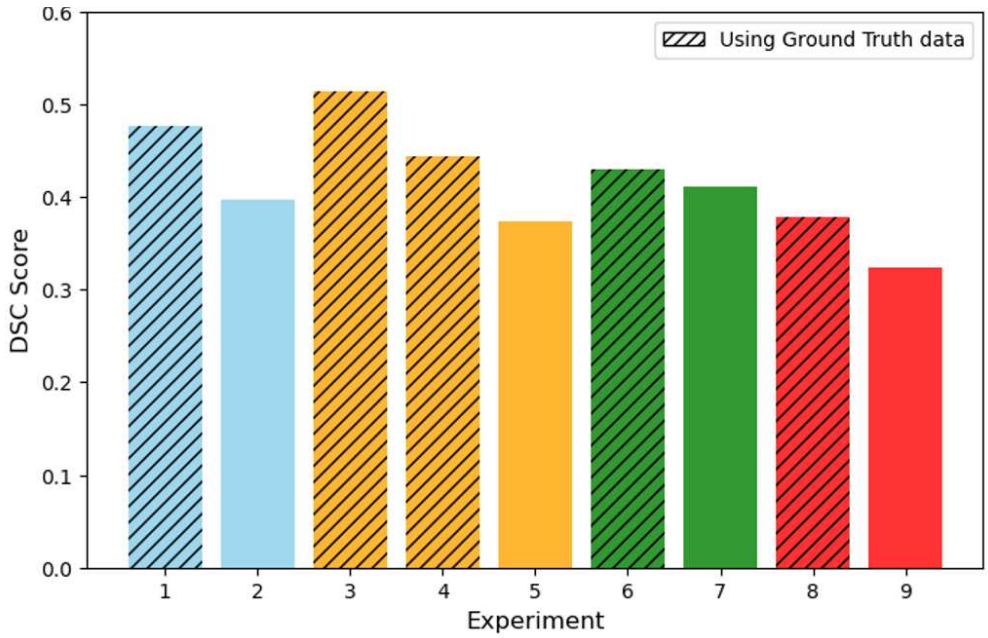
Here, we see a decrease in performance compared to the previous experiments. We also see that the model's limits using ground truth data (Experiment 8), takes a hit as well. When we average the DSC values from Experiment 9 and compare them to those in Experiment 8, we see that the model operates at 88% of its capacity, which is higher than the 75% capacity of the model at Experiment 2, but lower than the capacity found at Experiment 7. This can be explained by the sparsity of the GAMR annotations. In fact, real-time TRACE only extracts 8 GAMR annotations from all 4 test groups, compared to the 326 GAMR features we find in annotations of the same 4 groups; all of these are at the disposal of the offline version, but not the real-time version.

There's a noticeable decrease in performance when we compare Experiment 1 with Experiment 9, even though we added more modalities. This is very different from the results from Khebour et al. [Khebour et al., 2024b]. In that version, we see a mix of trends, but the decreases are not as big as with TRACE in real-time. In offline CGT the general trend across the test groups and all 4 values of DSC, is a drop of 1% when more modalities are used, whereas the live model shows an 8% drop. This is explained by the sparsity of the additional modalities in the move classifier. These features also go deeper into the model compared to the dense paraphrase, the outputs of which impact many other components of TRACE, further indicating the data sparsity issue. The best performing experiment that did not use any ground truth data is Experiment 7, which uses automatic speech transcriptions, dense paraphrases with ASR transcripts, automatically detected pointing gestures, and automatically-detected objects, as well as prosodic, CPS, and propositional features. This indicates a plausible best feature set for future work in real-time common ground extraction. GAMR features may also still have utility in a less-sparse data condition. The figures in Fig 5.1 presents a more visual summary of the results of the experiments.

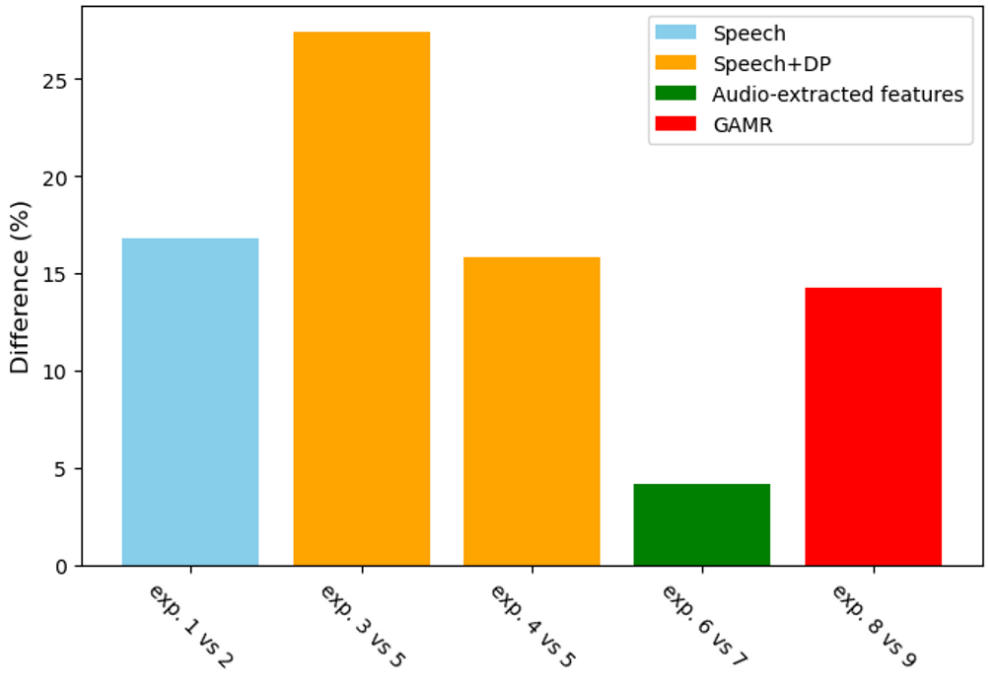
Table 5.4: Average DSC over test groups comparing CGT with TRACE. The last row shows TRACE results reported in Table 5.3

| **Modalities** | **QBank** | **EBank** | **FBank** | **F ∪ E** |
|---|---|---|---|---|
| CGT-All modalities | 0.714 | 0.535 | 0.313 | 0.851 |
| CGT-Speech only | 0.725 | 0.551 | 0.184 | 0.928 |
| TRACE-best performance | 0.608 | 0.213 | 0.166 | 0.514 |

Fig 5.2 shows the Kernel Density Estimation of the experiments in Tab 5.3. The KDE plots show more instability along the experiments using more modalities as we see the bell thickening from experiment 6 until 9. The estimations also suggest optimal stability is reached with experiment 4, where the model only uses dense paraphrase and speech.




(a) Summary of results in table 5.3 with hatched bars use the ground truth input.




(b) Difference in percent between the experiments using the same feature set.

Figure 5.1: Summary figures for the results analysis presented in this section.

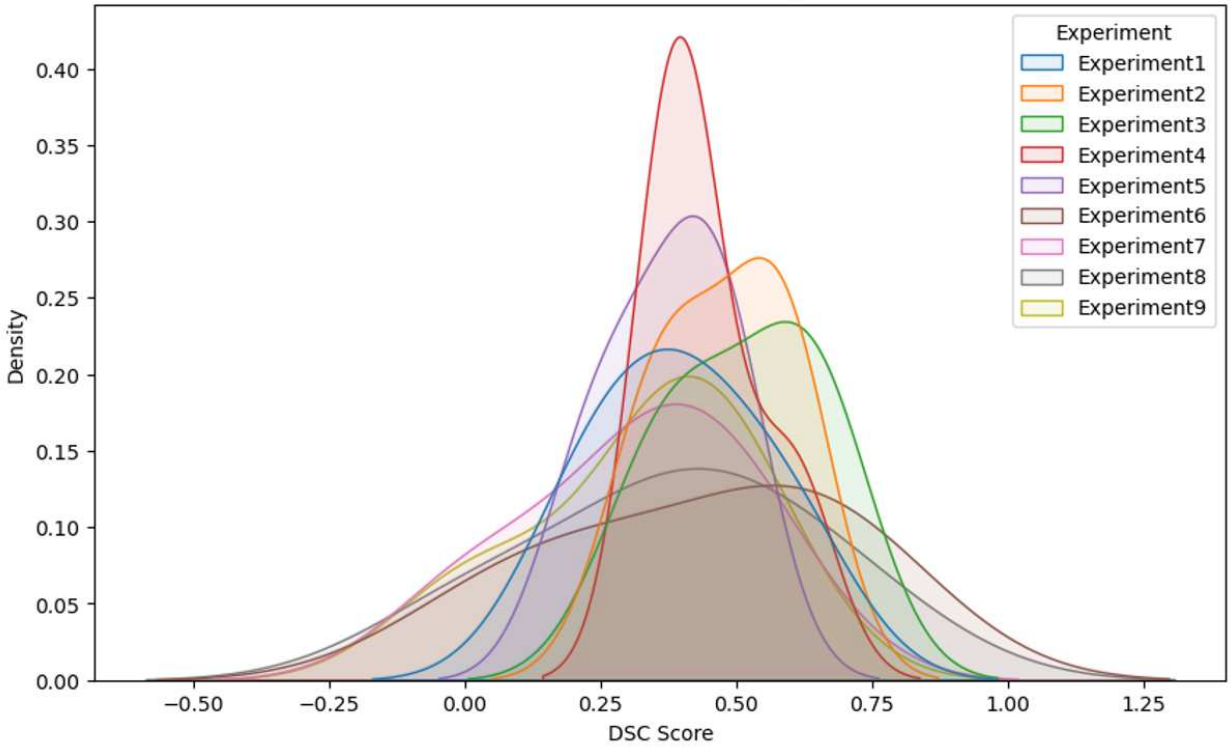



Figure 5.2: KDE plots computed using DSC results for each one of the four test videos.

# Chapter 6
# Discussion and Future Work

## 6.1 Results Discussion

This thesis examined the integration of multiple modalities into a real-time multimodal common ground tracker (TRACE) and analyzed how the inclusion of additional, potentially low-resourced modalities influences model performance in collaborative problem-solving contexts. Through the systematic addition and evaluation of visual and audio features, this work provided empirical evidence about the benefits and limitations of multimodal fusion in human–AI collaboration tracking tasks.

*RQ1: How do additional modalities influence the performance of a multimodal machine learning model?*

RQ1 explores how a multimodal ML model in a collaborative setting can be affected by different modalities. The results were more nuanced than expected, highlighting how the community's knowledge in multimodality might have been overestimated. Indeed, the addition of audio-based features improved the model's performance, however the visually extracted modalities (such as gesture and object-based cues) often caused a decrease in accuracy. This performance drop can largely be blamed on the difference in the temporal resolution of these modalities. As the TRACE system was designed to run in real-time, visual data had to be down-sampled substantially to maintain computational feasibility. The sparsity in the visual signals reduced the model's ability to extract consistent, meaningful features, leading to a degraded performance. In contrast audio modalities, characterized by the continuity and the high-frequency of their signals, retained a certain consistency that better aligned with TRACE's dynamic inference and learning process.

*H1: Adding additional modalities will improve performance of the multimodal model compared to unimodal baselines in collaboration tracking tasks.*

To relate back to the hypotheses from the first chapter, the findings only partially support H1. Although adding audio modalities did improve performance over unimodal baselines, the inclusion of visual modalities had the opposite effect by reducing accuracy, showing that not all added modalities necessarily enhance model performance.

*RQ2: What are the limitations of stacking modalities on top of each other?*

RQ2 focused on understanding the limitations of stacking multiple modalities together. There was a clear trade-off between model complexity and informative gain shown by the experimental results. Increasing the number of modalities can boost representational capac-

ity, but doing so comes with greater computing overhead, synchronization problems, and possible noise.

The thesis demonstrated that the model approaches a performance plateau after a certain point, indicating the presence of an upper limit dictated by the complementarity of its features as well as the model's architecture. This supports the notion that theoretical and empirical arguments, not only the availability of data, should serve as the basis for multimodality. Suboptimal outcomes may arise from an indiscriminate mix of modalities that dilutes the signal-to-noise ratio.

Therefore, the contribution here is to define a realistic boundary for building multimodal models: in order to achieve consistent, temporally aligned input streams and achieve effective performance gains, modalities must be balanced in addition to being stacked.

*H2: The main limitations to stacking modalities on top of each other are computational (training time, memory usage), rather than performance trade-off.*

H2 was not confirmed. The results demonstrated that the limitations of stacking modalities extend beyond computational costs. Performance trade-offs due to noise, redundancy, and lack of complementarity were more significant than expected, indicating that multimodal fusion requires careful theoretical justification rather than simply scaling input sources.

*RQ3: To what extent does the inclusion of low-resourced features impact a model's performance in a collaborative setting?*

The third research question examined how low-resourced or sparsely represented modalities impact performance. The experiments involving the GAMR gesture embedding features highlighted a key challenge in integrating such modalities. The sparseness of gesture features led to reduced utility in real-time inference, emphasizing that low-resourced features can only contribute meaningfully when they are sufficiently continuous or densely represented.

However, these characteristics could be useful for offline or hybrid systems with less stringent computing requirements. Additionally, they might encode supplementary data that could improve comprehension of group dynamics in cooperative tasks, like affect, attention, or engagement indications. Therefore, the results encourage a more focused investigation into the ways in which low-resourced characteristics might be improved for multimodal learning systems by being redefined, enhanced, or temporally smoothed.

*H3: Low-resource modalities will require more training time to converge compared to high-resource counterparts, but once trained their performance contributions will be comparable.*

Finally, H3 was only weakly supported. Low-resource modalities such as gesture features did not reach comparable contributions to higher-resource counterparts in real-time conditions, even with extended training. However, their potential value in offline or hybrid systems suggests that the hypothesis holds in less time-constrained settings, but not in the

real-time collaborative context studied here.

Overall, by showing how modality characteristics—particularly frequency and continuity—affect the design and effectiveness of multimodal systems, this thesis adds to the expanding corpus of work at the nexus of artificial intelligence and human–computer interaction. It offers a framework for critically evaluating which modalities to use and how best to combine them for collaborative analysis in real time.

## 6.2 Implications

The findings of this study have a number of ramifications for both the practical and research fields. They start by emphasizing how crucial modality selection and synchronization are for real-time systems. Multimodal architecture designers need to take into account how a modality interacts with other modality across time in addition to the information it offers. Compared to semantically rich but sparse modality (such as gestures or object data), a continuous but low-level modality (such as audio) can frequently contribute more to temporal inference.

Second, this study emphasizes how important multimodal balance is. The best configuration depends on matching data frequency, informativeness, and computing feasibility; adding modalities does not ensure better performance. These results go beyond cooperation tracking and could influence the development of multimodal perception models for communication analysis, social robotics, and education.

Finally, the study also shows that current multimodal fusion techniques do not involve any learning related to the relations between modalities. The fusion approach used as well as the other existing techniques fail at aligning modalities the way humans do. They are efficient mathematical ways of using multiple modalities within the same system, but they did not provide sufficient proof that the model understands how the modalities align temporally but also semantically.

## 6.3 Limitations

This study gave insightful information, however, these results remain limited by a number of constrains in terms of generalizability. The video recordings from WTD were carried out in a controlled environment, with no back noise, solving a particular task which is different from real life scenarios. Additionally, due to computational constraints, visual modalities had to be down-sampled, which limited the investigation of higher-frequency vision-based properties.

Other models may show varying sensitivity to modality imbalance; the architecture employed also reflects a particular multimodal fusion implementation. Lastly, the methodology limited the interpretive depth of the results by concentrating on the common ground—the participants' shared beliefs—without going as far as individual belief modeling.

## 6.4 Future Work

Future work can be inspired from the discovering of this thesis. One promising direction can be the investigation of more continuous visual features such as eye gaze tracking and body pose estimation, which could offer temporally stable signals along with the audio modalities. Also, refining the gesture features to be more continuous could help overcome some of the limitations discussed, but the solution has to be computationally light to avoid down-sampling the input.

Theoretically, expanding the scope from common to individual belief modeling can offer deeper insights into cognitive and social processes underpinning collaboration. Watching how personal beliefs evolve in time during a collaboration could also deepen our understanding on how to create a machine capable of understanding human-human communication.

Another potentially interesting avenue to explore in the future is hybrid online-offline architectures so that the system can process modalities at their natural frequencies without the real-time constraints. These systems could closely assess the contribution of each modality, and guide the development of adaptive real-time resource distribution to strategically prioritize modalities based on context.

These findings confirm that while multimodality is crucial for rich human–AI understanding, not all modalities contribute equally under real-time constraints. Effective multimodal systems thus depend not only on the quantity but on the temporal and structural compatibility of the modalities being combined. However, it is still notable that TRACE unlocked more potential in its capabilities when more modalities were given. These results showed that a well designed multimodal system can outperform unimodal models.

# Bibliography


Malihe Alikhani and Matthew Stone. Achieving common ground in multi-modal dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–15, 2020.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. The hcrc map task corpus. *Language and speech*, 34(4):351–366, 1991.

Rudolf Arnheim. Hand and Mind: What Gestures Reveal about Thought by David McNeill. *Leonardo*, 27(4):358–358, 1994.

Nicholas Asher. Common ground, corrections and coordination. *Journal of Semantics*, 1998.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.

Alexandru Baltag, Lawrence S Moss, and Sławomir Solecki. *The logic of public announcements, common knowledge, and private suspicions*. Springer, 2016.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Brigid Barron. Achieving coordination in collaborative problem-solving groups. *The journal of the learning sciences*, 9(4):403–436, 2000.

Mikhail Bilenko and Raymond J Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48, 2003.

Mariah Bradford, Ibrahim Khebour, Nathaniel Blanchard, and Nikhil Krishnaswamy. Automatic detection of collaborative states in small groups using multimodal features. In *Proceedings of the 24th International Conference on Artificial Intelligence in Education*, 2023.

Hennie Brugman and Albert Russel. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. Abstract meaning representation for gesture. In *Proceedings of the thirteenth language resources and evaluation conference*, 2022.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

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

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:213466962.

Pei Chen, Boran Han, and Shuai Zhang. Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving, 2024. URL https://arxiv.org/abs/2404.17729.

Herbert H Clark. *Using language*. Cambridge university press, 1996.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren Resnick, Levine B., M. John, Stephanie Teasley, and D, editors, *Perspectives on Socially Shared Cognition*, pages 13–1991. American Psychological Association, 1991.

Herbert H Clark and Thomas B Carlson. Context for comprehension. *Attention and performance IX*, 313:30, 1981.

Marco Del Tredici, Xiaoyu Shen, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. From rewriting to remembering: Common ground for conversational qa models. *arXiv preprint arXiv:2204.03930*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Indrani Dey, Sadhana Puntambekar, Ruofan Li, Diane Gengler, Rachel Dickler, Leanne M Hirshfield, Charis Clevenger, Sierra Rose, Mariah Bradford, and Nikhil Krishnaswamy. The NICE framework: analyzing students' nonverbal interactions during collaborative learning. In *Pre-conference workshop on Collaboration Analytics at 13th International Learning Analytics and Knowledge Conference (LAK 2023)*. Society for Learning Analytics Research, 2023.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. Cutting melted butter? common ground inconsistencies management in dialogue systems using graph databases. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):157–190, 2021.

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

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

Kerstin Fischer. How people talk with robots: Designing dialog to reduce user uncertainty. *AI Magazine*, 32(4):31–38, 2011.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273, Stockholm, Sweden, September 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5932. URL `https://aclanthology.org/W19-5932`.

Jonathan Ginzburg. Interrogatives: Questions, facts and dialogue. *The handbook of contemporary semantic theory. Blackwell, Oxford*, pages 359–423, 1996.

Jonathan Ginzburg. *The interactive stance: Meaning for conversation*. Oxford University Press, 2012.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. HyST: A Hybrid Approach for Flexible and Accurate Dialogue State Tracking. In *Proc. Interspeech 2019*, pages 1458–1462, 2019. doi: 10.21437/Interspeech.2019-1863.

Susan Goldin-Meadow. *Hearing gesture: How our hands help us think*. Harvard University Press, 2005.

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

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions, 2018.

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

Nikita Haduong, Irene Wang, Bo-Ru Lu, Prithviraj Ammanabrolu, and Noah A. Smith. Cps-taskforge: Generating collaborative problem solving environments for diverse communication tasks, 2024. URL `https://arxiv.org/abs/2408.08853`.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. Trippy: A triple copy strategy for value independent neural dialog state tracking. *arXiv preprint arXiv:2005.02877*, 2020.

Matthew Henderson, Blaise Thomson, and Steve Young. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, Philadelphia, PA,

U.S.A., June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4340. URL https://aclanthology.org/W14-4340.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework, 2024. URL https://arxiv.org/abs/2308.00352.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue, 2022.

Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Léo Jacqmin, Lina M Rojas-Barahona, and Benoit Favre. " do you follow me?": A survey of recent approaches in dialogue state tracking. *arXiv preprint arXiv:2207.14627*, 2022.

Chirag Kandoi, Changsoo Jung, Sheikh Mannan, Hannah VanderHoeven, Quincy Meisman, Nikhil Krishnaswamy, and Nathaniel Blanchard. Intentional microgesture recognition for extended human-computer interaction. In *Human-Computer Interaction*. Springer, 2023.

Adam Kendon. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, 25(1980):207–227, 1980.

Adam Kendon. Gesture. *Annual Review of Anthropology*, 26(1):109–128, 1997. doi: 10.1146/annurev.anthro.26.1.109. URL https://doi.org/10.1146/annurev.anthro.26.1.109.

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

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, et al. The weights task dataset: a multimodal dataset of collaboration in a situated task. *J. Open Human. Data*, 10(10.5334), 2023.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, et al. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of open humanities data*, 10, 2024a.

Ibrahim Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard Brutti, Christopher Tam, Jingxuan Tu, Benjamin Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, et al. Common ground tracking in multimodal dialogue. *arXiv preprint arXiv:2403.17284*, 2024b.

Anthony Pak-Hin Kong, Sam-Po Law, Connie Ching-Yin Kwan, Christy Lai, and Vivian Lam. A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a database of speech and gesture (dosage). *Journal of nonverbal behavior*, 39:93–111, 2015.

Stefan Kopp and Ipke Wachsmuth. Gesture in embodied communication and human-computer interaction. *Gesture in Embodied Communication and Human-Computer Interaction*, 5934, 2010.

Nikhil Krishnaswamy and James Pustejovsky. Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pages 44–51, 2019.

Nikhil Krishnaswamy and James Pustejovsky. A formal analysis of multimodal referring strategies under common ground. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5919–5927, 2020.

Geert-Jan M Kruijff, Pierre Lison, Trevor Benjamin, Henrik Jacobsson, Hendrik Zender, Ivana Kruijff-Korbayová, and Nick Hawes. Situated dialogue processing for human-robot interaction. In *Cognitive systems*, pages 311–364. Springer, 2010.

Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020.

Alex Lascarides and Matthew Stone. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449, 2009.

Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset, 2020.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Arav Agarwal, Yun Cheng, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multizoo multibench: A standardized toolkit for multimodal deep learning, 2023. URL `https://arxiv.org/abs/2306.16413`.

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

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

Bangli Liu, Haibin Cai, Xiaofei Ji, and Honghai Liu. Human-human interaction recognition based on spatial and motion trend feature. In *2017 ieee international conference on image processing (icip)*, pages 4547–4551. IEEE, 2017.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action, 2023. URL `https://arxiv.org/abs/2312.17172`.

Daqin Luo, Chengjian Feng, Yuxuan Nong, and Yiqing Shen. Autom3l: An automated multimodal machine learning framework with large language models, 2024. URL `https://arxiv.org/abs/2408.00665`.

Magdalena Markowska, Adil Soubki, Gary Mar, Seyed Abolghasem Mirroshandel, Owen Rambow, and Anita Wasilewska. Formal representation of common ground in dialogue.

Susan M Mather. Ethnographic research on the use of visually based regulators for teachers and interpreters. *Attitudes, innuendo, and regulators*, pages 136–161, 2005.

David McNeill. Hand and mind. *Advances in Visual Semiotics*, 351, 1992.

David McNeill. *Gesture and thought*. University of Chicago press, 2019.

David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling, 2023. URL `https://arxiv.org/abs/2312.06647`.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324, 2017.

Pradyumna Narayana, Ross Beveridge, and Bruce A Draper. Gesture recognition: Focus on the hands. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5235–5244, 2018.

OECD. *PISA 2018 Results (Volume I): What Students Know and Can Do*. OECD Publishing, Paris, 2019. doi: 10.1787/5f07c754-en.

Xenia Ohmer, Marko Duda, and Elia Bruni. Emergence of hierarchical reference systems in multi-agent communication. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5689–5706, 2022.

Eric Pacuit. *Neighborhood semantics for modal logic*. Springer, 2017.

Partnership for 21st Century Learning. Framework for 21st century learning. `http://static.battelleforkids.org/documents/p21/P21_Framework_Brief.pdf`, 2019. Accessed: 2025-10-10.

Jan Plaza. Logics of public communications. In *Proceedings 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

James Pustejovsky and Nikhil Krishnaswamy. Embodied human computer interaction. *KI-Künstliche Intelligenz*, 35(3-4):307–327, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. 2022.

Colin Rennie, Rahul Shome, Kostas E Bekris, and Alberto F De Souza. A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place. *IEEE Robotics and Automation Letters*, 1(2):1179–1185, 2016.

Patrick Louis Rohrer, Ingrid Vilà-Giménez, Júlia Florit-Pons, Núria Esteve-Gibert, Ada Ren, Stefanie Shattuck-Hufnagel, and Pilar Prieto. The multimodal multidimensional (m3d) labelling scheme for the annotation of audiovisual corpora. *Gesture and Speech in Interaction (GESPIN)*, 2020.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding, 2021.

Matthias Scheutz, Rehj Cantrell, and Paul Schermerhorn. Toward humanlike task-based dialogue processing for human robot interaction. *Ai Magazine*, 32(4):77–84, 2011.

David Schlangen and Gabriel Skantze. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111, 2011.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding, 2016.

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

Angela EB Stewart, Zachary Keirn, and Sidney K D'Mello. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction*, 31 (4):713–751, 2021.

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

Chen Sun, Valerie J. Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672, 2020b. URL `https://www.sciencedirect.com/science/article/pii/S0360131519302258`.

Praneet Sai Madhu Surabhi, Dheeraj Reddy Mudireddy, and Jian Tao. Thinktank: A framework for generalizing domain-specific ai agent systems into universal collaborative intelligence platforms. *arXiv preprint arXiv:2506.02931*, 2025.

Christopher Tam, Richard Brutti, Kenneth Lai, and James Pustejovsky. Annotating situated actions in dialogue. In *Proceedings of the 4th International Workshop on Designing Meaning Representation*, 2023.

Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, Nikhil Krishnaswamy, and Nathaniel Blanchard. How good is automatic segmentation as a multimodal discourse annotation aid? *arXiv preprint arXiv:2305.17350*, 2023.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

David Traum. A computational theory of grounding in natural language conversation. *PhD thesis, University of Rochester*, 1994.

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

Jingxuan Tu, Kyeongmin Rim, Bingyang Ye, Kenneth Lai, and James Pustejovsky. Dense paraphrasing for multimodal dialogue interpretation. *Frontiers in artificial intelligence*, 7: 1479905, 2024.

Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *IROS*, 2022.

Takuma Udagawa and Akiko Aizawa. Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics*, 9:995–1011, 2021.

Johan van Benthem and Eric Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99:61–92, 2011.

Johan van Benthem, David Fernández-Duque, and Eric Pacuit. Evidence and plausibility in neighborhood structures. *Annals of Pure and Applied Logic*, 165(1):106–133, 2014.

Coert Van Gemeren, Ronald Poppe, and Remco C Veltkamp. Spatio-temporal detection of fine-grained dyadic human interactions. In *Human Behavior Understanding: 7th International Workshop, HBU 2016, Amsterdam, The Netherlands, October 16, 2016, Proceedings 7*, pages 116–133. Springer, 2016.

C Van Rijsbergen. Information retrieval: theory and practice. In *Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems*, volume 79, pages 1–14, 1979.

Hannah VanderHoeven, Nathaniel Blanchard, and Nikhil Krishnaswamy. Robust motion recognition using gesture phase annotation. In *International Conference on Human-Computer Interaction*, pages 592–608. Springer, 2023.

Hannah VanderHoeven, Nathaniel Blanchard, and Nikhil Krishnaswamy. Point target detection for multimodal communication. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management.* Springer, 2024.

Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, et al. Trace: Real-time multimodal common ground tracking in situated collaborative dialogues. *arXiv preprint arXiv:2503.09511*, 2025.

Leonid Velikovich, Ian Williams, Justin Scheiner, Petar Aleksic, Pedro Moreno, and Michael Riley. Semantic lattice processing in contextual automatic speech recognition for google assistant. pages 2222–2226, 2018. URL `https://www.isca-speech.org/archive/Interspeech_2018/pdfs/2453.pdf`.

Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, James Pustejovsky, Nathaniel Blanchard, and Nikhil Krishnaswamy. Propositional extraction from natural speech in small group collaborative tasks. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 169–180, 2024.

Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, Hannah VanderHoeven, Brady Bhalla, Austin Youngren, James Pustejovsky, et al. Propositional extraction from collaborative naturalistic dialogues. *Journal of educational data mining*, 17(1):183–216, 2025.

Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, June 2021.

Katrin Wolf, Anja Naumann, Michael Rohs, and Jörg Müller. A taxonomy of microinteractions: Defining microgestures based on ergonomic and scenario-dependent requirements. In *13th International Conference on Human-Computer Interaction (INTERACT)*, number Part I, pages 559–575. Springer, 2011.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. *arXiv preprint arXiv:1905.08743*, 2019.

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020a.

Jian-Guo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S Yu, Richard Socher, and Caiming Xiong. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. *arXiv preprint arXiv:1910.03544*, 2019.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning, 2023. URL `https://arxiv.org/abs/2307.10802`.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63 (10):2011–2027, 2020b.

Zixuan Zhang and Heng Ji. Abstract meaning representation guided graph encoding and decoding for joint information extraction. In *Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies (NAACL-HLT2021)*, 2021.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. Reflect, not reflex: Inference-based common ground improves dialogue response quality. *arXiv preprint arXiv:2211.09267*, 2022.

Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008*, 2023.

Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare MC Tempany, Michael R Kaus, Steven J Haker, William M Wells III, Ferenc A Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic radiology*, 11(2):178–189, 2004.