












Multimodal Design for Interactive Collaborative Problem-Solving Support

Hannah VanderHoeven¹  , Mariah Bradford¹ , Changsoo Jung¹ ,
Ibrahim Khebour¹ , Kenneth Lai² , James Pustejovsky² ,
Nikhil Krishnaswamy¹ , and Nathaniel Blanchard¹ 

¹ Colorado State University, Fort Collins, CO 80523, USA

² Brandeis University, Waltham, MA 02453, USA

{hannah.vanderhoeven,nikhil.krishnaswamy,
nathaniel.blanchard}@colostate.edu

Abstract. When analyzing interactions during collaborative problem solving (CPS) tasks, many different communication modalities are likely to be present and interpretable. These modalities may include speech, gesture, action, affect, pose and object position in physical space, amongst others. As AI becomes more prominent in day-to-day use and various learning environments, such as classrooms, there is potential for it to support additional understanding into how small groups work together to complete CPS tasks. Designing interactive AI to support CPS requires creating a system that supports multiple different modalities. In this paper we discuss the importance of multimodal features to modeling CPS, how different modal channels must interact in a multimodal AI agent that supports a wide range of tasks, and design considerations that require forethought when building such a system that most effectively interacts with and aids small groups in successfully completing CPS tasks. We also outline various tool sets that can be leveraged to support each of the individual features and their integration, and various applications for such a system.

Keywords: Collaborative problem solving · Multimodal agents · HCI

1 Introduction

When humans engage in collaborative problem solving (CPS), the interaction is overwhelmingly likely to involve multiple communicative modalities simultaneously. Information may be communicated by speech, gesture, pose and interacting with objects in physical space. As artificial intelligence (AI) becomes more integrated with everyday workflows in environments such as classrooms and workspaces, there is increased potential for AI to support CPS in small groups such as project teams or workgroups in classes.

In this paper, we present a vision of AI agents whose purpose is not to automate tasks or replace human workers or teachers, but rather to augment the natural collaborative capabilities of humans and enable teams to think and reason better. That is, in a classroom context, an effective agent would not provide

the answer to a problem, but rather assist a group in discovering the answer organically, thus optimizing learning and retention. An AI agent that effectively supports CPS must be able to interpret many forms of communication to infer relevant context from a situation. However, accurate interpretations of group interaction rarely rely on any one of the above features alone (e.g., linguistic information alone may not adequately indicate which objects are the focus of attention or discussion). Therefore, many design decisions must be made in the data collection and feature extraction to make meaningful inferences about small group communication, as would be performed by an assistive AI agent in real-time. In addition, decisions need to be made about tools and common logic used for feature extraction so they can be brought together and used in a multimodal fashion. In this paper we examine design decisions made during the creation of the Weights Task Dataset—a multimodal CPS dataset that serves as our testbed, the features selected to motivate creation of a multimodal agent to support collaborative problem solving in small groups, and a vision for integration to create meaningful methods for interactive AI for a wide range of CPS tasks.

The attribution of mental states (e.g., beliefs, desires, and intentions) to interlocutors is a primary requirement of successful collaboration, but modern interactive systems, such as chatbots driven by large language models (LLMs) struggle with this capacity [46,56]. Namely they do not display fundamental characteristics of a Theory of Mind (ToM). In large part, this is due to a lack of mechanisms to interpret not just what is expressed, but *how* it is expressed, and how human interlocutors will interpret expressions in context. This requirement motivates many of the features we focus on in our proposed multimodal collaborative agent architecture. An agent endowed with the capabilities described below, both technically and theoretically, would not simply be the sum of the individual processing modules, but a gestalt system that, through the ability to process, integrate, and engage with the multitude of ways that humans may express their underlying mental states, simulates the epistemic positioning of its interlocutors toward a task *a la* [30] and thus assists them in organically achieving it.

2 Related Work

Given the interdisciplinary nature of this work, which draws on AI, human-computer interaction, linguistics, and learning sciences, among others, there is a vast background literature implicated by the various components of our research. A few key works are referenced here, and more are provided in subsequent sections.

Collaborative problem solving (CPS) is when two or more people use “their knowledge and skills to solve complex problems without predefined solutions” [50]. As such, this is a particular method of modeling interaction between users based on research in the learning sciences. Frameworks for CPS have been developed to capture relevant behaviors and different types of collaboration [1, 15, 49]. These frameworks are helpful in creating labeled data and provide

a bridge for computer scientists to operationalize and apply knowledge from the learning sciences. Previous work has successfully detected and classified these facets and showed improvements when using multimodal models [5, 48], explored technical requirements on an AI agent for tracking collaboration in small groups, such as relevant toolkits [7], and showed how adding contextual features to models improves the generalizability of the models in collaborative contexts [8]. We expand on both of these topics in this paper.

The intersection of multimodal processing (gesture, pose, and other nonverbal behavior) with interactive systems demands an increased focus on common semantic interpretations of different modal channels [42]. We draw from semantic representation schemes at various levels of abstraction (of which [3] is a seminal example), and unify them with coding schemes directed at collaborative problem solving research grounded in the learning sciences [11].

A well-designed multimodal agent is likely to leverage multiple independent communicative features to extract context from a scene to interpret the current state of a given situation. Additionally, an agent might interact with participants based on inferences drawn from the features. One such example of an interactive multimodal agent is Diana, an embodied agent that collaborates with participants as a direct participant the task itself [26, 27, 41]. Effective multimodal design is vital in creating a system like Diana, in order to understand and best interact with participants in real time. Diana models an interactive user, and while it does not leverage communicative features to understand and provide feedback on the current state of a task, it does leverage features, such as speech, gestures and facial expression, to meaningfully understand and interact with users as they work together to complete a task.

3 Weights Task Dataset

We ground our agent design in a dataset that represents human-human interactions that display characteristics we want such an agent to augment and amplify. The Weights Task Dataset (WTD) is a collection of audio-visual recordings where triads collaborate to identify the weights of various colored blocks, and the pattern describing block weights (an instance of the Fibonacci series) using a balance scale [23]. This dataset consists of 10 groups in which participants interact with each other and physical objects in their environment to complete this shared situated task. The interactions display reflective reasoning, consensus-building behaviors, and exposition of shared and contradictory beliefs during the course of executing actions toward the task goal. The nature of the task requires not just speech to communicate, but also gesture, action, and nonverbal behavior. Therefore, an agent in this and similar tasks would need to integrate inputs from diverse modal channels to interact effectively with the group. Figure 1 shows two example stills from the WTD, with the corresponding speech included. As a participant (Participant 2)¹ picks up and places a block on the scale, they make two statements: “By touch feels lighter” and “And that looks like it might be about even”. Without the additional context provided by interpreting situated

¹ Participants are conventionally indexed 1–3 from left to right in the video frame.

action and the locations of objects, it is impossible to infer which block they are referring too, and whether or not their statements are correct. Detecting gestures and actions—in this case a grasp—and tracking object locations would provide the context needed to detect that the speaker is referring to the green block.



Fig. 1. Stills from the Weight Task Dataset.

The dataset has been annotated, partially or completely, with most of the individual features discussed in Sect. 5. Details on annotation procedures, including adjudication and inter-annotator agreement, are given in [23].

4 Collaborative Problem Solving

Collaborative problem solving (CPS) is a form of collaboration wherein small groups work together to solve a nonroutine task with no set plan, where the quality of the solution can be evaluated by the team members as the task proceeds, and there is a differentiation of roles but interdependence within the team [15]. Previous work has designed several frameworks for identifying characteristics of CPS. For example, Hesse et al. developed a framework to assess students for CPS skills [18] which assesses students’ social and cognitive skills over the entirety of the task rather than marking specific events as occurrences of displayed skills. Andrews-Todd and Forsyth similarly broke CPS down into cognitive and social dimensions [1] but at a level that is specific to a simulated circuitboard task and therefore not easily generalizable to other settings. A general CPS support agent requires a framework that is not hyperspecific to actions only relevant to one domain, but which can also ground CPS skill indicators to specific events as the task unfolds. To this end, we use the framework developed by Sun et al. [49]. In this framework, CPS was formalized into hierarchical levels; 19 indicators that include moves such as proposing a correct solution or interrupting others, and three facets which are *Constructing shared understanding*, *Negotiation/Coordination* and *Maintaining team function*. These indicators allow us to identify specific collaborative moves; for example, in Fig. 1, the participants are discussing the results of weighing two blocks, where *discussing results* is an indicator enumerated in the Sun et al. CPS framework.

5 Features

5.1 Speech

Speech is a critical method of communication seen in group work. Participants use this modality to share their understanding, ask questions, discuss results, plan, and more. This is an explicit method of communication, making it a foundational starting point for any agent tracking group work. Previous studies demonstrated that speech is a meaningful feature for a model tracking group states [5, 48]. When combined with other features, utterances can help add context to how a participant is interacting with the space around them. Existing automatic speech recognition (ASR) tools, such as Google ASR and Whisper ASR [43, 59] can segment and transcribe audio into utterances of speech. Use of automatic or manual segmentation is an important design decision in integrating speech with other channels, as it affects the fidelity of downstream inference based on speech [51]. For real-time support, an agent will have to rely on an ASR system. Speech must be automatically diarized and transcribed for a system to work in real time, so these methods are a necessary component for an agent to have.

There is a high proportion of demonstrative terms and anaphors (“this”, “that”, “it”, etc.) implicated in dialogue during situated shared tasks. Automatically interpreting them usually involves recourse to another modality such as deictic gesture. As an interpretational technique, *Dense Paraphrasing* is a linguistically-motivated textual enrichment strategy that explicitly realizes the otherwise elided compositional operations inherent in the meaning of the language. This broadly involves three kinds of interpretive processes: (i) recognizing the diverse variability in linguistic forms that can be associated with the same underlying semantic representation (paraphrases); (ii) identifying semantic factors or variables that accompany or are presupposed by the lexical semantics of the words present in the text, through dropped, hidden or shadow arguments; and (iii) interpreting or computing the dynamic changes that actions, events, and other communicative modalities impose on objects 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 .

5.2 Acoustics

Acoustic features convey additional meaning in language. From turn-taking to posing a question, the way someone presents their statements provides additional information to others. Acoustic information allows us to understand sarcasm, perceive tone, recognize high energy, and more. For an agent, acoustics

(cadence, prosody, etc.) can help classify the sentiment of statements. Prior work has shown that acoustics are useful features for a model classifying a group’s state [5,48]. One system for automatically extracting acoustic features is openSMILE [13]. This allows users to retrieve the acoustic information within a detected segment. There are also existing feature sets which extract acoustic features relevant to human voice such as the GeMAPS set [12]. This is a condensed set which captures the most impactful information for sentiment in acoustics using a minimal amount of features, making it fast and lightweight—ideal for a live system. These acoustic features would be automatically extracted by an agent to more accurately process human language, including being able to distinguish questions from statements based on tone and cadence.

5.3 Gesture and Pose

Gestures and body language may also be important communicative modalities. Gesture is frequently used to disambiguate language and has complementary strengths (for instance, deictic gesture is naturally suited to indicating locations, while spoken language is more felicitous for indicating nominal qualities such as color). Pose and body language in a group context, meanwhile, can be an indicator of engagement with the team or lack thereof, focus of attention, etc. However, the interpretation of gesture and body pose may be subjective, and conditioned upon personality, background, culture, etc. [24]. Therefore, in a computational context, some form of structured representation language is required to make the continuous discrete and the intractable tractable.

The two representational schemes we build on here are **Gesture Abstract Meaning Representation** (GAMR) for gesture [6] and **Nonverbal Interactions in Collaborative-Learning Environments** (NICE) for other kinds of nonverbal behavior [10]. Annotations using these frameworks serve as output sets against which gesture and pose recognition models can be trained.

GAMR. Gesture AMR (GAMR) is a formalism intended to encode the meaning of gesture in multimodal interactions between agents. It is an extension to Abstract Meaning Representation (AMR), adopting both the annotated graph structure and the predicate-argument representation of that formalism [3]. Gesture AMR was developed to encode how gesture packages meaning both independently of and in interaction with speech; and how the meaning of gesture is temporally and contextually determined.

GAMR GESTURE UNIT

```
(g / gesture-unit
 :op1 (i / icon-GA
       :ARGO (g2 / gesturer)
       :ARG1 (b / block)
       :ARG2 (a / addressee))
 :op2 (d / deixis-GA
       :ARGO g2
       :ARG1 (l / location)
       :ARG2 a))
```

GAMR includes schemata to annotate gestures that fall into one or more of these categories, thus providing granularity when representing a variety of gestures that might be used to communicate in various CPS tasks. The inset shows the structure for a “gesture unit” including both deixis and iconic components. ARGO denotes the gesturer, ARG1 the semantic content of the gesture and ARG2 is the addressee or intended recipient; these fields exist for each gesture subsection in the annotation.

NICE. The NICE coding scheme captures nonverbal behaviors when people are working together in groups. There is a subtle, yet important distinction between a silent individual who is nonverbally participating in their group and a silent individual who mostly works by themselves and neither verbally nor nonverbally participates. Additionally, different nonverbal cues can occur concurrently, and in clusters [63].

NICE captures multiple modalities that indicate collaborative learning and engagement, such as the direction of gaze (where are they looking?), posture (are they 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). Eye gaze could be indicative of where attention is directed [47, 52], whether it is jointly on group work or on other interlocutors. Head movements (such as nodding in agreement or shaking in disagreement) are captured as an indication that the person is paying attention [38], as are leaning forwards to look at the joint activity or participating in the same [4]. The coding scheme also captures contrasting behaviors that would imply lower collaboration or attention, such as looking at or doing their own work (instead of the joint work) or outside the activity area, leaning away, “fiddling” (idly interacting with non-task-related objects or interacting with task-related objects in non-task-related ways) [16], etc. Additionally, the NICE coding scheme captures four emotions (*positive emotion*, *negative emotion*, *boredom*, and *confusion/concentration*) as they are working together, based on observational cues, which provide indications of learning [54].

NICE is designed to be calibrated to the task in that the vocabulary of objects must be pre-specified to match the perceptible object space of the common ground that evolves between participants as the task unfolds.

Gesture AMR distinguishes four general types of referential gestures: *iconic*, *deictic*, *metaphoric*, and *emblematic* [22, 25, 34, 35]. Because our data focuses on gestures in a task-based setting, most depictions of entities and events appear to reflect their concrete properties, such as the shape of an object or the manner of an action. Similar to the interactions reported on in [6], metaphoric gestures do not appear with any frequency.

Gesture Recognition. There are many possible solutions to gesture recognition [14, 17, 36, 53], but nearly all suffer from difficulties in recognizing gestures at unusual angles or that may be far from camera, and deep learning approaches come with a high training and data overhead, making them difficult to adapt to new environments and situations. This difficulty demands a more lightweight robust solution that can rapidly be deployed under novel circumstances, potentially on everyday hardware.

From the gesture semantics community comes a tradition of modeling gesture in terms of preparatory, “stroke” (including pre- and post-stroke “hold”) and subsequent recovery *phases* of gestures [2, 21, 31]. Based on this, we previously developed a gesture recognition pipeline, with the goal to streamline the detection of complex gestures, for eventual deployment in real time [57]. This pipeline uses hand detection tools, like MediaPipe [62], to detect joint locations of individual hands in a frame (21 joints in 3D coordinates). The pipeline is made up of three major stages: a static classification model that recognizes the general static shape of a gesture in any of the “hold” phases, a movement segmentation algorithm that tracks the movement of hands over time and breaks a video stream down into segments based on changes in motion patterns, and a phase breakdown process that uses the results of the previous steps to identify segments with frames in a “hold” phase. The start and end frames of these hold segments are recorded as “key frames,” or the frames that comprise the most semantic significance of any given gesture. Individual static classification models can be trained for a variety of different relevant gestures for CPS tasks, providing flexibility and granularity for a larger multimodal system. We have leveraged this pipeline to successfully detect multiple different kinds complex gesture, ranging from subtle small hand movements (*microgestures* [20, 57, 61]), to deictic gestures [58].

Figure 2 shows an example of our recognition method applied to pointing detection in the WTD. In the frame we can see that Participant 1 is pointing at the blocks on the scale. A pointing frustum built around the vector extended out from the participant’s index finger has further narrowed down the blue block as a target of interest (see [58] for details). Through a comparison to the GAMR annotations at the same intervals, we can see that the blue block is in fact the intended target. Combining target detection using deixis with other features like speech in a multimodal system can further disambiguate the intended subjects of action during collaborative problem solving.

Pose Detection. Similar adaptability concerns inform approaches to pose detection in multimodal CPS scenarios. In this case, important features are more likely to be associated with gross body motion than fine-grained joint positions on the hand. Using the depth channel from Azure Kinect recordings, the positions and orientations of 32 joints on the body can be extracted, in a similar manner to hand detection with MediaPipe. To classify instances of non-gestural nonverbal behaviors, along the lines of those captured by the NICE coding, joint features need to be tracked over time and converted to nonverbal behavior labels.



Fig. 2. Group 1 deixis with GAMR example (reproduced from [58])

Suitable approaches to this task may include processing the concatenated raw joint positions through a sliding window of fixed size to accumulate descriptive features of the motion over time, and then training a neural classifier to fit the relationship between joint positions and NICE codes.

Because the raw joint positions are anchored to the physical location of the different participants, the individual bodies may either need to be segmented and processed individually with distinct models for each person, or transformed into a normalized space before feature processing and classification.

5.4 Actions

ACTION ANNOTATION

```

(p / put-ACT
:ARG0 (p1 / participant-1)
:ARG1 (gb / green-block)
:ARG2 (o / on
:op1 (rs / right-scale)))

```

Actions in context provide important information that situates other modalities within the environment. For instance, the subject of an action may be the antecedent of a subsequent demonstrative even if it is never explicitly labeled in dialogue.

This motivates the use of a rigorously-defined interaction semantics for action tracking. Annotation of all task-specific actions engaged in by the participants, like GAMR, follows an AMR-style syntax which introduces the notion of annotating actions in the style of speech and gesture. This involves making reference to a taxonomy of action classes, adapted from relevant predicates from PropBank [39], that are interpreted as VoxML programs [40]. For example, an action of putting a green block on the right side of a scale would be assigned the annotation shown. The argument structure given for the Action AMR annotation of this event follows that of the corresponding PropBank predicate, in this case, *put-01*.

The results of actions performed over objects accommodate downstream reasoning, such that in a context or configuration \mathcal{C} , the execution of an action or program π results in state \mathcal{R} ($\mathcal{C} \rightarrow [\pi]\mathcal{R}$ according to [41, 42]) which can further indicate what a participant may be thinking or reasoning about. For instance, in Fig. 1, P2 placing the green block on the scale is also an indication of *intent* and of what P2 believes the likely results will be based on the *affordances* of the objects involved: in this example, namely, that the scale will end up balanced.

5.5 Facial Expression

Facial expression is an informative modality for an agent tracking group work, as it can indicate both level of engagement (similar to body pose) and also participants' attitudes towards individual events or even each other. Recent work has shown improvement in facial expression recognition through improvements in deep learning [32, 33]. However, there is a great diversity of ways to interpret different facial expressions, often depending on context. For our purposes we care about the relations of specific expressions to collaborative problem solving. Toward identifying these affects, D'mello and Graesser defined patterns in affective states specific to learning [11]. This allows us to narrow down into expressions that will be important for an agent to track. Recent work focusing on these affective states has been able to recognize facial expressions representing these states [45]. This modality allows an agent to identify the learner's affective condition. For example, an agent may detect that a learner is confused and offer clarification. Previous work also showed eye gaze to be an informative feature for classifying group member involvement [37]. While eye gaze detection is still limited, it could be an informative channel for an agent when detecting participation.

5.6 Physical Space

Fundamentally, an agent will be unable to meaningfully interact with users and support groups in a physical context without the ability to track the movement of objects in space and make inferences regarding the relationships between them. Mechanistically, this requires common calibration settings to allow data extracted using various tools to be used together (e.g., gesture landmarks exist in the same space as the object locations). The use of 6DOF object pose estimation [9] to extract object locations admits challenges when deployed in group work scenarios, particularly those in classroom environments, where interactable objects are likely to be small, moved a lot, and subject to partial or complete occlusions. One way to address this is with a model that estimates the object mask to predict the position of an object and then crops the masked image to estimate the its rotation.

6DOF Pose. The position of an object in 3D Cartesian space can be estimated through prediction of the object's 6DOF pose (6 degrees of freedom, comprising translation and rotation in all three orthogonal dimensions [19, 29, 60]). By

detecting the positions of objects, a system may couple that with their sizes and properties (see Sect. 5.4) to make inferences about the physical relationships between objects. Another advantage is that 6DOF pose estimation allows objects to be tracked over time, and thus allows tracking the context of an interaction with an object that may result in a change to its state or configuration. This information is typically extractable from an RGB pixel stream, but in our usages we also leverage the additional benefits of a depth channel, for instance by using Azure Kinects, for greater accuracy.

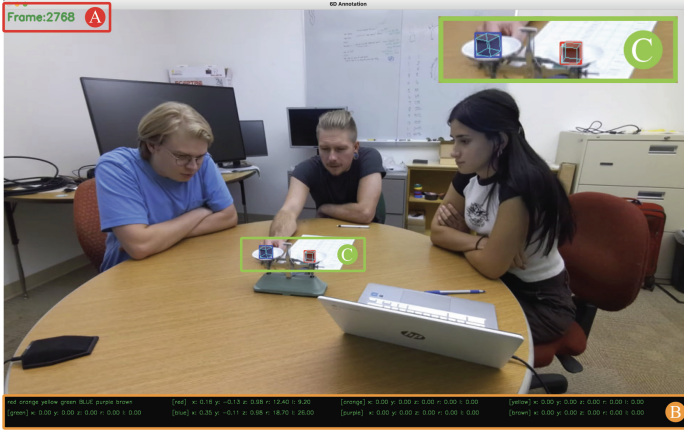


Fig. 3. 6DOF Pose Annotation Tool on CPS Data. *A* shows the current frame number, *B* shows the position and rotation information for each object of interest, and *C* (expanded in inset) shows annotated 2D and 3D bounding boxes.

The challenge of CPS tasks for 6DOF pose estimation is the difficulty of visual feature extraction. To capture a sufficient amount of the scene in which a CPS task typically unfolds, the camera must be placed further away from the relevant objects that it typically has been in other 6DOF pose estimation tasks and datasets [44, 55]. This is naturally required to capture the participants and how they interact with each other using modalities such as gesture and pose (see Sect. 5.3), but also renders the relevant objects very small in the frame, making them difficult to annotate for training, and to capture at inference time. Figure 3 shows the challenge of 6DOF pose annotation in the WTD.

Figure 4 shows the convergence of pointing and object detection. When performing automated object selection using deixis, an end-to-end solution would require that the objects be automatically detected within the scene at the same time that gestures are also recognized, rather than using pre-annotated bounding boxes (as in Fig. 2). Current state of the art approaches to tasks of this kind are typically composed of multiple modules, trained over a combination of real images and 3D renderings of the objects of interest in a variety of orientations. A typical solution may start with a convolutional neural network (CNN)

to extract spatial information and visual features of objects. Visual features are used to predict the poses of objects in the next module. Following this, a “refinement step” takes place, in which the module estimates object pose and those estimates are used to render images of the objects, which are then compared to the real training images. Error is backpropagated until the renderings and real images are within an appropriately small epsilon.

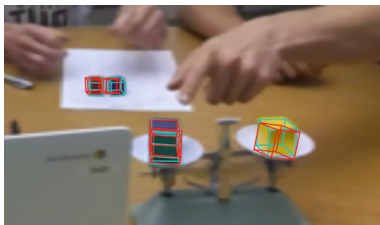


Fig. 4. Ground truth object bounding boxes (blue) and predicted bounding boxes (red). Deixis is used to select a spatial region containing one or more objects, which may be further disambiguated by contemporaneous speech or prior context. (Color figure online)

One or more objects may be detected within a region singled out by deixis, indicating the object or set of objects that are the likely foci of attention. This may be further disambiguated by linguistic information, such as nominal descriptors or previous discussion of objects that have been acted upon.

6 Multimodal Fusion

A key foundational challenge of signal fusion is one of aligning the different modal channels. For instance, a non-linguistic feature could align with more than one utterance. One common strategy to handle such instances is mapping non-linguistic inputs to the utterances that they share the greatest temporal overlap with. A potential problem with this strategy may arise if, for instance, a very brief gesture begins at the end of one utterance *A*, but lasts long enough to overlap more with the next utterance *B*. Another possible issue is when ASR models detect speech where none exists, causing non-linguistic inputs to end up aligned with a “hallucinated” transcription. Alternate mapping strategies could involve choosing the overlapping utterance based on length, semantic qualities, or distance metrics between feature representations, such as a GAMR annotation and an utterance’s AMR.

With the level of feature diversity presented and the wide variation within each, to arrive at a representation usable by an AI system, a deep learning solution is *de rigueur*. Design choices in fusion algorithms will primarily revolve around the step at which the fusion of the different feature types takes place, of which there are 3 primary classes: *early fusion*, *late fusion*, or *hybrid fusion*.

In *early fusion*, the data is fused at the start of the learning algorithm, such as through concatenation, then processed as a single input. This may lead to imbalance in feature contribution to the final output. If the different modalities differ in format or size (e.g., input dimensions or one-hot vs. real-valued vectors), the output will be biased toward the numerically richer type of data,

regardless of semantic contribution. *Late fusion* trains on each modality separately in unimodal submodules, and then merges those outputs. This method can better handle imbalance in feature input size, as the submodules’ output sizes are controllable and specifiable. This may result in a larger neural network with associated potential issues, such as the vanishing gradient problem. *Hybrid fusion* mixes the previous two: some modalities are trained separately, but if two or more modalities have a certain connection (e.g., overlaps between gesture and action), or if they have the same format and sizes, they may be handled together. Fusion may be as simple as concatenating the features, depending on the stage at which it is performed, or could involve more complex methods such as learning attention weights between queries of one modality, and keys/values of another modality. Figures 5a and 5b show high-level schematic diagrams of early and late fusion, respectively. Hybrid fusion combines the two, in that the some modalities may be processed through individual submodules while others are input directly to the fusion layer.

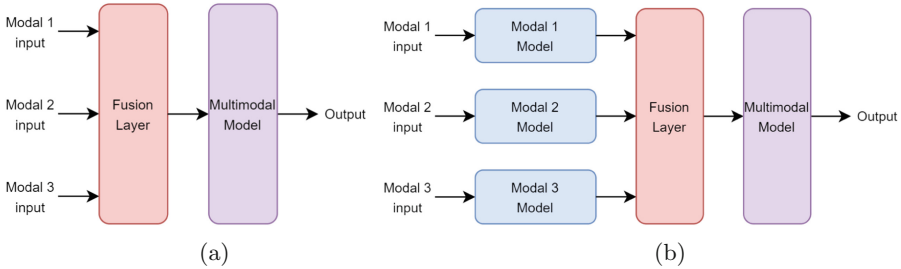


Fig. 5. (a) Early fusion schematic diagram. (b) Late fusion schematic diagram.

Finally, the joint representation is considered as a standard feature vector by the prediction head. The model may leverage linear or non-linear layers either as a discrete classification head or the joint representation may be fed into another module, such as a pre-trained language model for output generation (see below).

7 Behavior Generation

Our focus in this paper is on processing and integrating multimodal channels to enable an agent to process human task behavior. To actually act upon its inferences, however, the agent also needs to produce naturalistic output that is coherent, on-topic, supports the task goal, and scales beyond deterministic and trivial use cases. Such an agent would itself be multimodal, to symmetrically replicate the same modes of interaction as human group members. Examples of previous multimodal agents (e.g., Diana from Sect. 2) interact directly with the user using speech, gesture, and action. This contrasts with our aims here, which is to build an agent that supports group dynamics rather than performing the task itself, but shows the importance of a multimodal generation capability.

The advent of generative AI models, of which large language models (LLMs) such as ChatGPT/GPT-4 are noteworthy exemplars, provides a partial solution to such a problem, with their facility in generating language that on occasion appears indistinguishable from human writing or speech. A key challenge is that off-the-shelf generative AI systems tend to 1) generate longer outputs than the sentence fragments actually used by people during collaborative interaction and 2) display pronounced weaknesses in problems pertaining to situational and multimodal reasoning. Open-weight LLMs can be tuned on task dialogue samples to replicate more realistic dialogue structure. That is, rather than an LLM that produces fully informative, complete sentences, generated utterances should actually be more fragmented and ambiguous, and dependent on multimodal information that can then be validated against the environment.

Likewise, to generate multimodal information such as gestures and actions, the underlying model needs to be trained to insert non-linguistic “tokens” into the output. This may involve fine-tuning the autoregressive mechanism over specialized datasets that include naturalistic multimodal ensembles, for example of multimodal referring expressions (e.g., [28]). Multimodal information like gesture and action annotations can be represented as special tokens, such as unique non-human-interpretable identifiers added to the model’s vocabulary, with their representations projected into the LLM’s space to facilitate generation in context. Outputted special tokens, when encountered by the agent’s linguistic parser, would be skipped by the agent’s linguistic renderer (e.g., text-to-speech system), and routed instead to a gesture and action manager to make the agent execute behaviors that the language must then be grounded to. These behaviors could be animated gestures and actions in a simulated environment, or, with the right hardware support, deployed on a physical robot co-situated with the group.

8 Evaluation

With a new type of interactive agent comes a need to develop appropriate evaluation metrics to gauge success or failure of the agent design. Here we propose a number of quantitative and qualitative metrics that may be considered.

1. Agent-augmented team vs. non-augmented team task completion rate. Agent inferences and corresponding outputs can be correlated to group factors like task completion time/rate and time to resolve uncertainties;
2. Agent adaptation to human behavior. Adaptations to improve group collaboration can be assessed with respect to which agent moves prompt increased collaboration in the group members. This can be assessed over time to see how the adaptation of agent outputs prompts more positive CPS from humans;
3. Common ground among agent-augmented team vs. non-augmented team. A proxy for the ToM capabilities of an agent (human or artificial) is its ability to build a consensus with its interlocutors and act according to it. Given a task and a set of relevant propositions, CPS skills displayed by a group (with

or without the agent) can be correlated to beliefs or intentions the group members correctly attribute to each other;

4. Decision making quality. While this is largely a subjective metric, an agent-supported group should be more transparent, reflective, and deliberative, while drawing on the perspectives of everyone involved and the collective knowledge of the group, resulting in decisions supported by all parties and the maximum spectrum of available evidence.

9 Applications

There are many different applicative use cases for multimodal systems that interpret and return feedback based on small group communication. A well-designed AI agent would collect and interpret enough context from a situation to aid in the problem solving process. For example, it could point out information the group members may not have considered, realign priorities based on team needs, and incorporate collaborator knowledge on the fly through a representation of objectives, subgoaling, changing plans, uncertainty, etc. Using the WTD as a specific example, the collaborative AI agent could maintain a model of the Fibonacci sequence as the goal and would be able to interpret the group's general understanding as they work to discover said pattern, based on how they communicate and the CPS skills they display. The agent would aid the participants by helping them reach the correct conclusion organically, and could also learn based on participants' task behavior, further tuning its feedback to the group to support their learning in an optimal way. A similar feedback cycle could be applied to any collaborative task, thus empowering teams to think and reason better in a wide range of different tasks and circumstances.

An agent might also take a more direct interactive approach to aiding small groups. One way would be to model the shared and individual beliefs of group members. This would enable it to raise questions about unspoken or unresolved conflicts or intervene in cases of groupthink for which there is no evidence. This is an important capability to prevent groups from making critical errors. As the agent gathers data on communication style and the beliefs of individuals, it would directly intervene and steer the conversation to correct misapprehensions or encourage more productive solutions to relevant subgoals. In the WTD, this might occur when one or more individuals believe a block is the incorrect weight. The agent could step in and ask participants to try weighing a specific combination of blocks. It could then analyze both if the correct conclusion was drawn from the action, and which CPS skills were displayed during the subdialogue. This interaction style could also be applied to a variety of CPS tasks providing a more direct teaching style for team support.

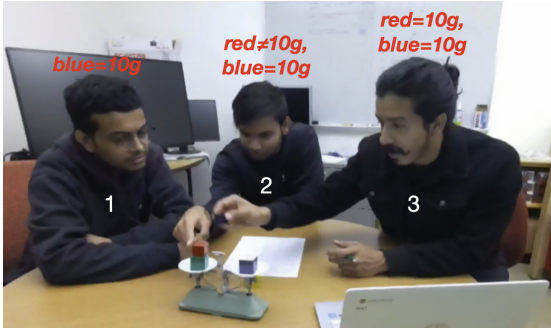


Fig. 6. Still of Group 10 from the Weights Task. The group is working to discover the weight of the blue block, based on inferences about the red block. Red text over each participant shows beliefs each apparently holds at this point, based on their prior utterances and actions. (Color figure online)

block weighs 10g, but P2 and P3 disagree about the red block. P3 *acts* based on his belief but P2 makes an inference based on his (expressed in speech).

Both of the aforementioned agent styles would be valid ways to help the group succeed. In the Weights Task example, in a more organic approach the agent would continually collect and process dialogue moves and analyze how the group is working towards the intended goal; for example, by performing inferences over recognized gestures and actions to understand what blocks are being interacted with, object recognition of their locations on the scale, and linguistic understanding of the group's statements to each other, the agent could verify that the current thought process is generally directed toward discovering the correct pattern even if there are specific inaccuracies at the current time. Using a more direct approach, the agent could reason over the inferences of each individual up to the current point in the task and intervene with corrective measures to help drive valid inferences toward the discovery of the overall pattern.

10 Conclusion and Future Work

Small groups engaged in CPS are likely to engage in various forms of multimodal communication, often simultaneously. In this paper we presented a design for an AI agent intended to support small group learning and collaboration by interacting with or providing feedback to individuals participating in various CPS tasks, specifically in a classroom setting. Potential communicative modalities such a group could utilize include speech, gestures, pose, and interaction with objects in physical space. To create such a system, tracking and interpreting various multimodal features is vital to gather the requisite level of context about how individuals interact with each other at any given time while completing the task. This leads to a set of deliberately-chosen requirements on individual components to maintain tractability, generalizability, and robustness.

Figure 6 shows a still from the WTD where a group is working to determine the weight of the blue block, based on previous inferences about the red blocks. In this scene multiple different modalities can be leveraged by an agent to determine the validity of the group's thought process, including but not limited to, what participants are saying, which blocks they are pointing at or grasping, and the location of the blocks relative to the scale. For instance, the whole group believes the blue

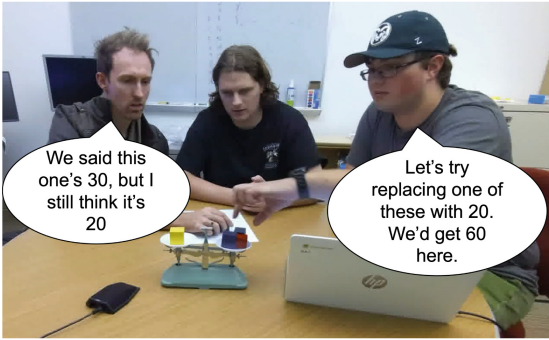


Fig. 7. Still of Group 1 from the Weights Task. Potential communicative modalities in the scene include but are not limited to, *speech, gesture, pose, action, gaze, and acoustics*.

blue blocks. This could be used further determine if the overall statement made by P3 is true, thus helping the agent understand the current progress toward successful task completion. P2 is not speaking, gesturing or interacting with the blocks, however the general direction of their gaze and forward-leaning pose indicate they remain engaged in the task. P1’s statement “but I think it’s still 20,” combined with certain cadence or prosodic patterns could signal P1’s confusion or hesitation about the group’s current trajectory (cf. [5]). Leveraging the combined modalities, such an agent would be able to maintain a relatively detailed model of what is currently happening in the scene, from specific action-level occurrences to the general level of contribution of each participant.

Additional design considerations concerning the exact methods that will bring features together, in addition to how the participants will interact with the agent, remain to be determined. A complete system should integrate implementations of 6DOF object recognition, action detection, expression recognition, and pose estimation, and would open up the wide variation of AI-assisted CPS to experimentation and evaluation, both from a human factors and multimodal fusion perspective. For instance, emergent properties of group behavior with AI assistance could be correlated to CPS skills displayed and other modalities predictive thereof, to determine the chain of events from individual behaviors to dialogue moves to CPS indicators to outcomes, which could be communicated back to the participants in real time.

Acknowledgments. This work was partially supported by the National Science Foundation under subcontracts to Colorado State University and Brandeis University on award DRL 2019805. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government. All errors and mistakes are, of course, the responsibilities of the authors.

Figure 7 shows an example of a single interaction in the Weights Task. In this situation, there are many inferences an AI agent could draw about the current state of the task. For instance, P1 and P3 are speaking and gesturing. P3 says “one of these,” while performing what appears to be a *grasp* gesture (cf. [57]). With hand and object detection to localize the blocks in the working space, an agent could infer that the group is speaking about the red or

References

1. Andrews-Todd, J., Forsyth, C.M.: Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Comput. Hum. Behav.* **104**, 105759 (2020). <https://doi.org/10.1016/j.chb.2018.10.025>
2. Arnheim, R.: Hand and mind: what gestures reveal about thought by David McNeill. *Leonardo* **27**(4), 358 (1994)
3. Banarescu, L., et al.: Abstract meaning representation for sembanking. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 178–186 (2013)
4. Barron, B.: When smart groups fail. *J. Learn. Sci.* **12**(3), 307–359 (2003)
5. Bradford, M., Khebour, I., Blanchard, N., Krishnaswamy, N.: Automatic detection of collaborative states in small groups using multimodal features. In: AIED (2023)
6. Brutti, R., Donatelli, L., Lai, K., Pustejovsky, J.: Abstract meaning representation for gesture, pp. 1576–1583, June 2022. <https://aclanthology.org/2022.lrec-1.169>
7. Castillon, I., Venkatesha, V., VanderHoeven, H., Bradford, M., Krishnaswamy, N., Blanchard, N.: Multimodal features for group dynamic-aware agents. In: Interdisciplinary Approaches to Getting AI Experts and Education Stakeholders Talking Workshop at AIED. International AIED Society (2022)
8. Chejara, P., Prieto, L.P., Rodriguez-Triana, M.J., Kasepalu, R., Ruiz-Calleja, A., Shankar, S.K.: How to build more generalizable models for collaboration quality? Lessons learned from exploring multi-context audio-log datasets using multimodal learning analytics. In: LAK2023, pp. 111–121. Association for Computing Machinery, New York, NY, USA, March 2023. <https://doi.org/10.1145/3576050.3576144>
9. Cunico, F., Carletti, M., Cristani, M., Masci, F., Conigliaro, D.: 6D pose estimation for industrial applications, pp. 374–384, September 2019. https://doi.org/10.1007/978-3-030-30754-7_37
10. Dey, I., et al.: The NICE framework: analyzing students’ nonverbal interactions during collaborative learning. In: Pre-Conference Workshop on Collaboration Analytics at LAK 2023. SOLAR (2023)
11. D’Mello, S., Graesser, A.: Dynamics of affective states during complex learning. *Learn. Instr.* **22**(2), 145–157 (2012)
12. Eyben, F., et al.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **7**(2), 190–202 (2016). <https://doi.org/10.1109/TAFFC.2015.2457417>
13. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462. Association for Computing Machinery, New York, NY, USA, October 2010. <https://doi.org/10.1145/1873951.1874246>
14. Fan, H., et al.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6824–6835 (2021)
15. Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the science of collaborative problem solving. *Psychol. Sci. Pub. Interest* **19**(2), 59–92 (2018). <https://doi.org/10.1177/1529100618808244>
16. de Haas, M., Vogt, P., Krahmer, E.: When preschoolers interact with an educational robot, does robot feedback influence engagement? *Multimodal Technol. Interact.* **5**(12), 77 (2021)
17. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6546–6555 (2018)

18. Hesse, F., Care, E., Buder, J., Sassenberg, K., Griffin, P.: A framework for teachable collaborative problem solving skills. In: Griffin, P., Care, E. (eds.) *Assessment and Teaching of 21st Century Skills*. EAIA, pp. 37–56. Springer, Dordrecht (2015). https://doi.org/10.1007/978-94-017-9395-7_2
19. Hu, Y., Fua, P., Wang, W., Salzmann, M.: Single-stage 6D object pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020
20. Kandoi, C., et al.: Intentional microgesture recognition for extended human-computer interaction. In: Kurosu, M., Hashizume, A. (eds.) *HCII 2023*. LNCS, vol. 14011, pp. 499–518. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35596-7_32
21. Kendon, A.: Gesticulation and speech: two aspects of the process of utterance. In: *The Relationship of Verbal and Nonverbal Communication*, vol. 25, pp. 207–227 (1980)
22. Kendon, A.: *Gesture: Visible Action as Utterance*. Cambridge University Press (2004)
23. Khebour, I., et al.: When text and speech are not enough: a multimodal dataset of collaboration in a situated task (2024)
24. Kita, S.: Pointing: a foundational building block of human communication. In: *Pointing: Where Language, Culture, and Cognition Meet*, pp. 1–8 (2003)
25. Kong, A.P.H., Law, S.P., Kwan, C.C.Y., Lai, C., Lam, V.: A coding system with independent annotations of gesture forms and functions during verbal communication: development of a database of speech and gesture (dosage). *J. Nonverbal Behav.* **39**, 93–111 (2015)
26. Krishnaswamy, N., et al.: Diana’s world: a situated multimodal interactive agent. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13618–13619 (2020)
27. Krishnaswamy, N., et al.: Communicating and acting: understanding gesture in simulation semantics. In: *IWCS 2017-12th International Conference on Computational Semantics-Short papers* (2017)
28. Krishnaswamy, N., Pustejovsky, J.: Generating a novel dataset of multimodal referring expressions. In: *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pp. 44–51 (2019)
29. Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: CosyPose: consistent multi-view multi-object 6D pose estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12362, pp. 574–591. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_34
30. Lai, K., et al.: Modeling theory of mind in multimodal HCI. In: *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*. Springer (2024)
31. Lascarides, A., Stone, M.: A formal semantic analysis of gesture. *J. Semant.* **26**(4), 393–449 (2009)
32. Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z.: Attention mechanism-based CNN for facial expression recognition. *Neurocomputing* **411**, 340–350 (2020). <https://doi.org/10.1016/j.neucom.2020.06.014>
33. Li, S., Deng, W.: Deep facial expression recognition: a survey. *IEEE Trans. Affect. Comput.* **13**(3), 1195–1215 (2022). <https://doi.org/10.1109/TAFFC.2020.2981446>
34. Mather, S.M.: Ethnographic research on the use of visually based regulators for teachers and interpreters. In: *Attitudes, Innuendo, and Regulators*, pp. 136–161 (2005)

35. McNeill, D.: Hand and mind. In: *Advances in Visual Semiotics*, vol. 351 (1992)
36. Narayana, P., Beveridge, R., Draper, B.A.: Gesture recognition: locus on the hands. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5235–5244 (2018)
37. Oertel, C., Salvi, G.: A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In: *Proceedings of the 15th ACM on International Conference on Multimodal Interaction - ICMI 2013*, pp. 99–106. ACM Press, Sydney, Australia (2013). <https://doi.org/10.1145/2522848.2522865>
38. Ogden, L.: Collaborative tasks, collaborative children: an analysis of reciprocity during peer interaction at key stage 1. *Br. Edu. Res. J.* **26**(2), 211–226 (2000)
39. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. *Comput. Linguist.* **31**(1), 71–106 (2005)
40. Pustejovsky, J., Krishnaswamy, N.: VoxML: a visualization modeling language. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 4606–4613. European Language Resources Association (ELRA), Portorož, Slovenia, May 2016. <https://aclanthology.org/L16-1730>
41. Pustejovsky, J., Krishnaswamy, N.: Embodied human computer interaction. *KI-Künstliche Intelligenz* **35**(3–4), 307–327 (2021)
42. Pustejovsky, J., Krishnaswamy, N.: Multimodal semantics for affordances and actions. In: Kurosu, M. (ed.) *HCI 2022. LNCS*, vol. 13302, pp. 137–160. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05311-5_9
43. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022)
44. Rennie, C., Shome, R., Bekris, K.E., De Souza, A.F.: A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place. *IEEE Rob. Autom. Lett.* **1**(2), 1179–1185 (2016)
45. Ruan, X., Palansuriya, C., Constantin, A.: Affective dynamic based technique for facial emotion recognition (FER) to support intelligent tutors in education. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) *AIED*, vol. 13916, pp. 774–779. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-36272-9_70
46. Sap, M., LeBras, R., Fried, D., Choi, Y.: Neural theory-of-mind? On the limits of social intelligence in large LMS. arXiv preprint [arXiv:2210.13312](https://arxiv.org/abs/2210.13312) (2022)
47. Schneider, B., Pea, R.: Does seeing one another’s gaze affect group dialogue? A computational approach. *J. Learn. Anal.* **2**(2), 107–133 (2015)
48. Stewart, A.E.B., Keirn, Z., D’Mello, S.K.: Multimodal modeling of collaborative problem-solving facets in triads. *User Model. User-Adap. Inter.* **31**(4), 713–751 (2021). <https://doi.org/10.1007/s11257-021-09290-y>
49. Sun, C., Shute, V.J., Stewart, A., Yonehiro, J., Duran, N., D’Mello, S.: Towards a generalized competency model of collaborative problem solving. *Comput. Educ.* **143**, 103672 (2020). <https://www.sciencedirect.com/science/article/pii/S0360131519302258>
50. Sun, C., et al.: The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Comput. Hum. Behav.* **128**, 107120 (2022)
51. Terpstra, C., Khebour, I., Bradford, M., Wisniewski, B., Krishnaswamy, N., Blanchard, N.: How good is automatic segmentation as a multimodal discourse annotation aid? (2023)
52. Tomasello, M., et al.: Joint attention as social cognition. In: *Joint Attention: Its Origins and Role in Development*, vol. 103130, pp. 103–130 (1995)

53. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: *Advances in Neural Information Processing Systems*, vol. 35, pp. 10078–10093 (2022)
54. Törmänen, T., Järvenoja, H., Mänty, K.: Exploring groups' affective states during collaborative learning: what triggers activating affect on a group level? *Educ. Tech. Res. Dev.* **69**(5), 2523–2545 (2021)
55. Tyree, S., et al.: 6-DoF pose estimation of household objects for robotic manipulation: an accessible dataset and benchmark. In: *IROS* (2022)
56. Ullman, T.: Large language models fail on trivial alterations to theory-of-mind tasks. arXiv preprint [arXiv:2302.08399](https://arxiv.org/abs/2302.08399) (2023)
57. VanderHoeven, H., Blanchard, N., Krishnaswamy, N.: Robust motion recognition using gesture phase annotation. In: Duffy, V.G. (ed.) *HCI 2023*. LNCS, vol. 14028, pp. 592–608. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35741-1_42
58. VanderHoeven, H., Blanchard, N., Krishnaswamy, N.: Point target detection for multimodal communication. In: *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*. Springer (2024)
59. Velikovich, L., Williams, I., Scheiner, J., Aleksic, P., Moreno, P., Riley, M.: Semantic lattice processing in contextual automatic speech recognition for google assistant, pp. 2222–2226 (2018). https://www.isca-speech.org/archive/Interspeech_2018/pdfs/2453.pdf
60. Wang, G., Manhardt, F., Tombari, F., Ji, X.: 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)*, pp. 16611–16621, June 2021
61. Wolf, K., Naumann, A., Rohs, M., Müller, J.: A taxonomy of microinteractions: defining microgestures based on ergonomic and scenario-dependent requirements. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) *INTERACT 2011*. LNCS, vol. 6946, pp. 559–575. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23774-4_45
62. Zhang, F., et al.: MediaPipe hands: on-device real-time hand tracking. arXiv preprint [arXiv:2006.10214](https://arxiv.org/abs/2006.10214) (2020)
63. Zoric, G., Smid, K., Pandzic, I.S.: Facial gestures: taxonomy and application of non-verbal, non-emotional facial displays for embodied conversational agents. In: *Conversational Informatics: An Engineering Approach*, pp. 161–182 (2007)