Tracking Individual Beliefs in Co-Situated Groups Using Multimodal Input

Mariah Bradford[0009−0009−2162−3307], Ibrahim Khebour[0009−0009−4374−7263], Hannah VanderHoeven[0000−0003−3234−6797], Videep Venkatesha[0009−0000−4635−3010], Nathaniel Blanchard[0000−0002−2653−0873], and Nikhil Krishnaswamy[0000−0001−7878−7227]

Colorado State University, Fort Collins CO 80523, USA
mbrad@rams.colostate.edu

Abstract. In co-situated collaborative groups, a challenge for automated interpretation of group dynamics is parsing and attributing input from individual group members to process their respective perspectives and contributions. In this work, we describe the necessary components for such a system to handle multimodal, multi-party input. We apply these methods over an audiovisual dataset of a co-situated collaborative task called the Weights Task Dataset (WTD) to track individual beliefs regarding the task. We find that combining audiovisual speaker detection (ASD) with utterance transcripts enables us to track individuals' beliefs during a task. We show that our system succeeds in individual belief tracking, achieving scores similar to those seen in dense-paraphrased common ground tracking. Further, we demonstrate that a combination of ASD and point target detection can be applied to transcripts for automated dense paraphrasing. We additionally identify where individual components need to be improved, including ASD and task-belief identification.

Keywords: Automatic Belief Tracking · Co-Situated Collaboration · Multimodal Alignment

1 Introduction

Group collaboration, which is commonplace in educational environments and the workforce, can be pleasantly productive or frustratingly inefficient for the group members. This may be due in large part to the group dynamics: individuals may feel disengaged or that they are unable to contribute and have their perspectives heard, or the group dynamics may lead to confusion and a lack of mutual understanding. In recent years, there has been a growing interest in using artificial intelligence (AI) systems to support groups in achieving productive collaboration [17, 25, 36, 35]. An assistive agent can help moderate these experiences to ensure meaningful collaboration between group members. For example, it may modulate the dialogue to create opportunities for everyone to speak, help the group work through disagreement, or detect when the group is behaving in a

way counter to available evidence. However, introducing an assistive agent into a multi-party setting, particularly a co-situated one, presents the inherent challenge of parsing the input from channels such as audio and visual signals and correctly attributing it to the appropriate group member.

An example of this challenge can be seen in Figure 1. Here, three participants are collaborating to detemine the weights of differently-colored blocks. Each of the participants express distinct views and information. Because the setup is captured by a single microphone, all utterances end up in the same input stream. Thus, the single audio segment, which by default may be considered a single "utterance" due to the continuous audio signal, needs to be parsed and each individual utterance must be attributed to the contributing group member. In this work we explore this challenge with a multimodal approach, using visual cues to identify the active speaker as well as ground gesture with speech. We then use that output along with a dialogue move classifier to determine the task-relevant beliefs of each participant. Our goal is to model the beliefs that are individually held by members of co-situated groups.

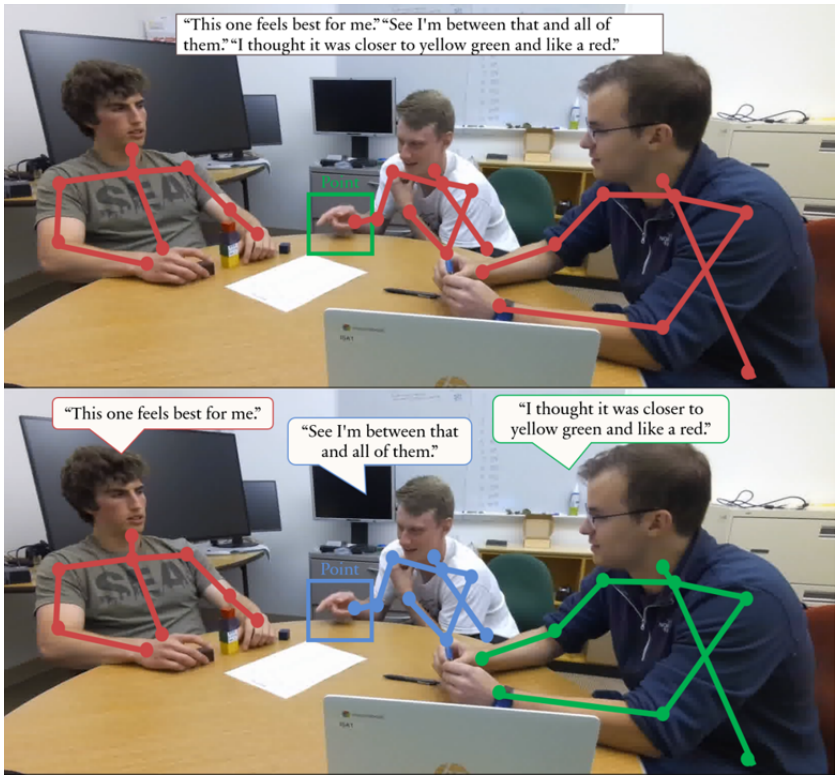

Fig. 1. Mapping multi-party input channels to individuals.

In this work, we describe the necessary components for a system to handle multimodal, multi-party input. We apply these methods over an audiovisual dataset of a co-situated collaborative task called the Weights Task Dataset (WTD) [15] to extend work presented in [17], which modeled only the group common ground. Here, we model beliefs at the individual level, the intersection of which, at a high level, constitutes the common ground.

Our approach creates a temporary representation of each group member which can be discarded after processing. This can help an agent track contributions and conflicts within the group without infringing on the privacy of individual members. Take, for example, an instance where one group member is not looking at the task area, and later has a disagreement rooted in something said or done while they were not paying attention. Modeling at the individual level enables pinpointing the cause(s) of diverging beliefs (cf. [23]). This paves the way for a single agent to assist both groups and individual members.

2 Related Work

In previous work, researchers modeled online group collaboration [35, 36] and decision making [13]. These are important indicators of group success, and an AI can use them to infer opportune moments for intervention. However, in an in-person scenario like that shown in Figure 1, which replicates a realistic environment like a table in a classroom or office where people are not outfitted with individual mics or specialized equipment, there are no longer dedicated audio and visual inputs that can be used to individuate each person. Previous work has explored interpreting inputs from group members - for example, Palmer et al. [25] used body tracking to model engagement of face-to-face groups, but have yet to incorporate speech.

A multitude of datasets have proliferated to facilitate the study of AI in group settings. These include groups working together online, as in [14], as well as in in-person meetings such as [6]. There also exist co-situated group datasets where groups work to solve a given task. One example is the emergent leader corpus, which explores roles in groups in a discussion-based task [33]. Another example, and the dataset used in the current study, is the Weights Task Dataset (WTD), which is a problem-solving task using blocks [15]. These datasets provide information for researchers to explore and compare results, furthering our understanding of group behavior.

To better understand these behaviors, previous work explored frameworks to identify collaborative behaviors [1, 10, 11, 38]. Some frameworks have split collaborative skills into *cognitive* and *social* elements [1, 11]. Others have described these elements as intertwined, such as Sun et al.'s collaborative problem solving coding framework [38], which identifies *constructing shared knowledge*, *negotiation and coordination*, and *maintaining team function* as the main pillars of collaboration. Another study took a broader look at collaborative skills and assessment, finding that many students and adults lack proficient collaborative problem-solving skills [10].

The aforementioned frameworks have been applied to behavior modeling. Some relevant studies rely on speech alone [28], while others rely on only video [7]. Several studies have shown that speech alone can offer significant insights into the group, but results can be improved with multimodal features and models [5, 35]. Another study explored implementing an agent to assist in remote groups using speech [36]. Further, group modeling has been used to predict performance of groups [2, 22, 37, 39].

Recently, theory of mind (ToM) approaches [27] have been applied to automated and AI analysis of collaborative groups. Of note, Zhu et al. [46] propose a *simulation theory of mind* (SToM), which posits that the underlying system of understanding others' mental states relies on our ability to maintain a simulation or "mind reading" of it [9], and use that to explain and predict behaviors, including beliefs, desires, and intentions (BDI). SToM finds its genesis in Shanton and Goldman's *simulation theory* and gave rise to AI research in multimodal simulations, including collaborative agents [18, 19, 29–31], but these were limited to peer-to-peer human-agent interactions, not group interactions. However, this work laid the foundation for automated modeling of *common ground* in dialogue, first in human-AI pairings [20], and then in groups of human collaborators [17].

Consensus-building is a critical component of collaboration because often a group must agree on the current step of a task before proceeding. In this paper and its immediate antecedents [17, 44], beliefs are a limited set of propositions regarding the task at hand. These include goals, constraints, and possible solutions. While Khebour et al. [17] made strides in tracking group agreement, similar techniques have yet to be applied to capture individual beliefs. The aforementioned studies demonstrate progress in the automated understanding and interpretation of group dynamics, but they have yet to address the challenge of attributing multimodal signals to individuals within co-situated groups for belief modeling. Thus, the import of this work becomes clear: an ability to model individual beliefs allows an agent to identify conflict and misunderstanding. A key requirement for this is aligning signals from multiple channels, such as speech and gesture, to the corresponding individual. To address this challenge we apply a combination of audiovisual speaker detection (ASD) and automatic gesture recognition to attribute task-relevant assertions to the correct group member, and evaluate on the tasks of *dense paraphrasing* and *individual belief tracking.*

3 Definitions

Here we present important defintions we rely on in the remainder of this paper. Subsequent uses of these terms should be assumed to follow these definitions.

- *Evidence* - Evidence allows an agent to make sensible decisions about its situation [12]. Under a dynamic epistemic logic [4, 24], evidence creates an accessibility relation between worlds $X \subset W$ such that for a given current world w accessible worlds X are those evidenced by w in their own unique way.

- *Belief* - Since it is possible to entertain evidence both for and against some proposition p, belief in p arises when the evidence for p is sufficiently strong that the set of worlds X accessible from w contain only propositions *non-contradictory* to the contents of w.
- *Individual Belief* - Because we treat each individual as having their own set of beliefs, it is possible for each actor to have their own unique mental model $\mathcal{M}$ that is supported under the current world w. Further, because we don't have access to what evidence an individual is working with, we allow for statements of p to immediately entail a (at least provisional) belief in p.
- *Dense Paraphrase* - To dense paraphrase an utterance is to decontextualize it; that is, provide all of the necessary context within the sentence to fill in missing information and reduce ambiguity [40, 41]. This is done by rewriting the original sentence. We do this by replacing pronouns in a sentence with the intended target.
- *Statement* - A statement occurs when a participant expresses a solution to the task. Here, that refers to expressing the weight of a block.
- *Accept* - An accept occurs when a participant agrees with another participant's statement.

4 Methodology

Our task is to take in audiovisual recordings of co-situated groups and output the belief states of each individual task participant. Our problem statement follows the format of *common ground tracking* (CGT) in multimodal dialogue [17]: for each utterance, we extract the epistemic positioning expressed by it (statement or acceptance) and the propositional content asserted, and use a set of closure rules derived from epistemic modal and public announcement logics to populate "banks" of shared evidence (*EBank*) and agreed-upon facts (*FBank*). The difference between CGT and *individual belief tracking* (IBT) is that the assumptions governing the closure rules that modulate the contents of EBank and FBank pertain not to correspondences between statements and acceptances from different people but to the epistemic state of individuals inherent in their own utterances. Thus rather than separate banks for shared evidence and facts that depend upon implicit or explicit agreement between multiple parties, we assume a single bank of individual *beliefs* that implicitly corresponds to an individually-held evidence bank.[1] Our pipeline begins with manually transcribed utterances, audiovisual speaker detection, and pointing detection. We combine these modalities for the textual enrichment technique knows as *dense paraphrasing* (Section 4.4). We then pass that output to a dialogue move classifier to detect *statements* and

[1] Though counterintuitive, under van Benthem et al.'s neighborhood semantics [4] where the belief and evidence operators $[B]$ and $[E]$ are the respective equivalents of the modal operators $\Box$ and $\Diamond$, for a single agent a, belief that there is evidence for φ ($[B]_a[E]_a\varphi$) is analogous to $\Box\Diamond\varphi$ ("necessarily possibly φ"). Since this entails that there is *some* sequence of accessibility relations between the current w and a world where φ holds, $[B]_a[E]_a\varphi \Rightarrow [E]_a\varphi$.

acceptances (Section 4.5), and a propositional extractor to identify claims being made (Section 4.6). These undergo simple closure rules (Section 4.7) to perform individual belief tracking. A visualization of our pipeline can be seen in Figure 2. We perform evaluations at several points in our pipeline, described in Section 5.

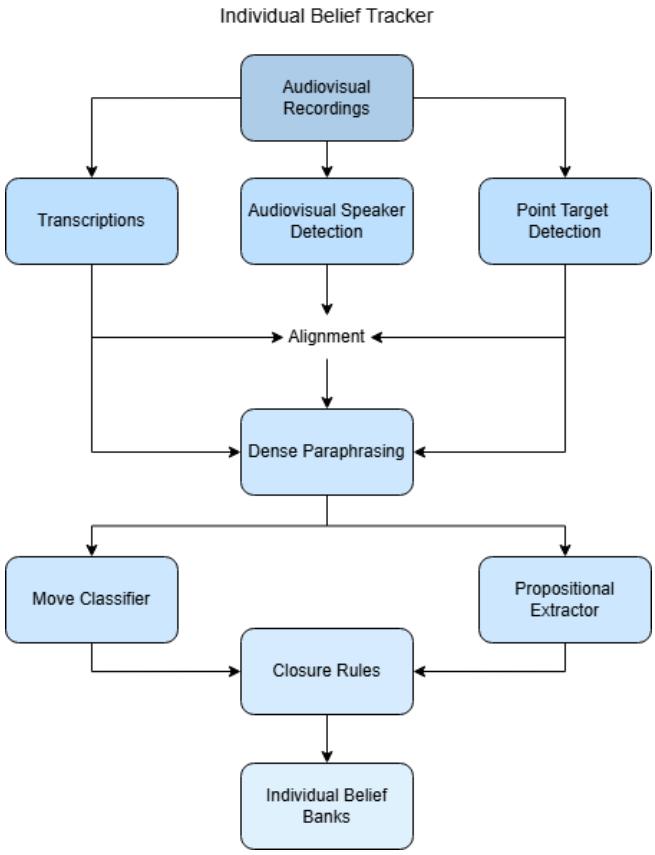

Fig. 2. Individual Belief Tracking pipeline.

4.1 Dataset

In this work we use the publicly released and IRB-approved Weights Task Dataset (WTD) [15], which can be seen in Figure 1. This dataset contains 170 minutes of video, comprising 10 groups working together with a set of blocks and a balance scale. The group's first task is to identify the weights of the blocks using the scale. For the following tasks, the scale is removed and they must

identify a pattern to solve the remaining block weights. In this work we only use data from the first task. The dataset also contains manual transcriptions of the participants' utterances, and annotations for *common ground*, when participants make statements regarding the answers for the task, based on [46]. For example, when Participant 2 says "The purple one is thirty", it is annotated as "S0358: STATEMENT(P2, purple = 30)". When Participant 3 replies with, "Purple is thirty, yeah", it is annotated as "AC0401: ACCEPT(P3, S0358)". We use these existing annotations, focusing on the statements individuals made, to create *individual belief banks*. Groups 1, 2, 4, and 5 were annotated for block locations in the video, for use in the point target detection model. The transcribed utterances were also enriched with explicit mentions of the blocks in place of demonstrative pronouns. For instance, if a participant said "So that one's 20" while pointing to the green block, the enriched utterance would be "So green one's 20" (see Figure 5). This constitutes a manual dense paraphrase (see Section 4.4). These were dual annotated (Cohen's $\kappa = 0.88$) and adjudicated by an expert.

4.2 Audiovisual Speaker Detection

Audiovisual speaker detection (ASD) is the process of attributing speech to individuals in the frame using audio and visual features. For audiovisual speaker detection, we used the Light-ASD model presented in [21]. We selected Light-ASD based on its small size and complexity paired with high performance [21]. This model utilizes a visual encoder, an audio encoder, and a gated recurrent unit to process audiovisual recordings and output frame-by-frame bounding boxes of the active speaker. This model also provides scores for each speaker. Any score of zero or higher is classified as speaking. An example of Light-ASD used on the WTD can be seen in Figure 3. We use several methods to aggregate these scores to retrieve utterance-level predictions. These experiments are further described in Section 5.

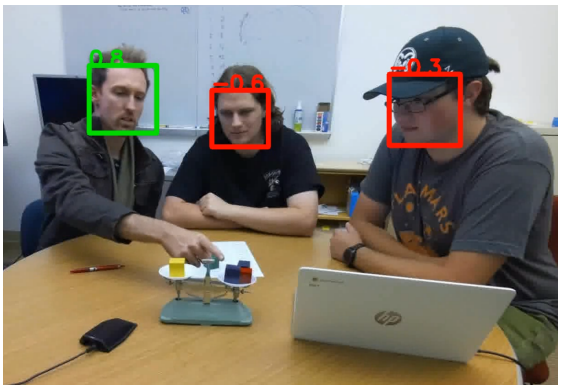

Fig. 3. Example of Light-ASD over the WTD, showing Participant 1 (left) detected as the active speaker while the other two participant are silent.

4.3 Gesture and Body Detection

Gesture is an important form of nonverbal communication that can be used to help add additional context group work tasks. By detecting and identifying gestures, we can understand how individuals interact with each other and the physical space around them. One form of gesture that occurs commonly in group work are deictic gestures, such as pointing. To perform meaningful gesture detection, we use the robust gesture detection pipeline target detection model described in [42]. The pipeline uses MediaPipe [45] to detect 21 joint positions across individual hands. These landmarks can then be fed into a classifier trained to identify gestures of interest and used to find the average change in motion of the hand, to then identify the key phases or frames of a gesture of interest. These key frames can then be used in varying detail to help identify a target of interest at any given timestamp, thus aiding in automatic detection of the intended targets of an individual's gesture. For pointing, this is achieved by extending a vector through the index finger of the hand, and creating a detection region in 3 dimensional space using a conical frustum shape. Objects that intersect with this region would be flagged as a potential target of interest [43] (see Figure 4).

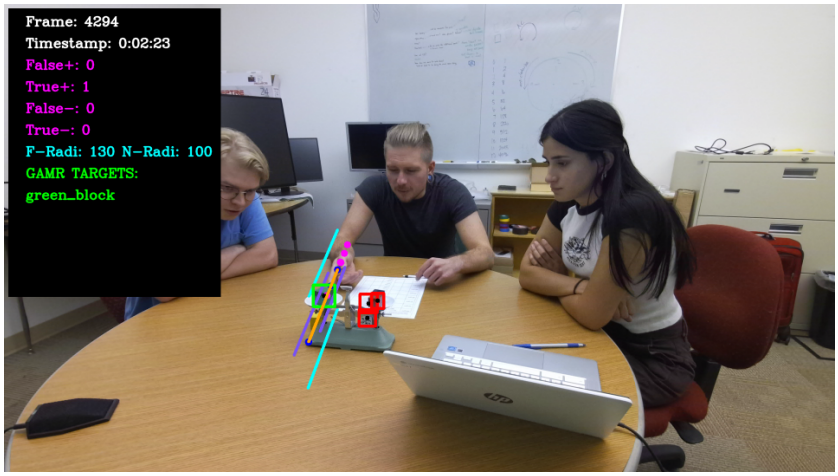

Fig. 4. Example of object selection using recognized pointing gestures (reproduced from [43]).

In order to associate deictic gestures and target objects with statements made by an individual in a group, we need to track the locations of participants throughout the group task. This is done by using the Azure SDK to return body landmarks on a frame by frame basis. We can then use the less detailed hand locations on the body to create a bounding box around each of the participant's hands. By tracking the general hand locations on each participant we are able to more consistently associate gestures and targets with individuals, in addition

leveraging these bounding boxes allows for more efficient hand tracking, by only requiring MediaPipe to be run in a subset of the frame.

4.4 Dense Paraphrasing

Dense paraphrasing is the process of decontextualizing speech using other modal channels such as gesture, action, or references to the environment and items within it [40, 41]. In our case, we supplement speech with visual information by replacing pronouns with their targeted objects. Manually dense paraphrased utterances are included in the WTD annotations (Section 4.1), providing ground truth, and we also experiment with automatic, gesture-based dense paraphrasing, replacing demonstrative pronouns with the objects selected by automatically detected pointing gestures, as described above. To align speech with gesture, we apply dense paraphrasing using speech, speaker detection, and gesture detection. Deictic gestures in particular allow individuals to identify targets or locations of interest in 3 dimensional space. When deictic gestures align with statements that use demonstratives such as "this" or "that" the ability to identify the intended target of the gesture and associate it with the individual communicating is vital to add context to speech. A schematic of this is shown in Figure 5.

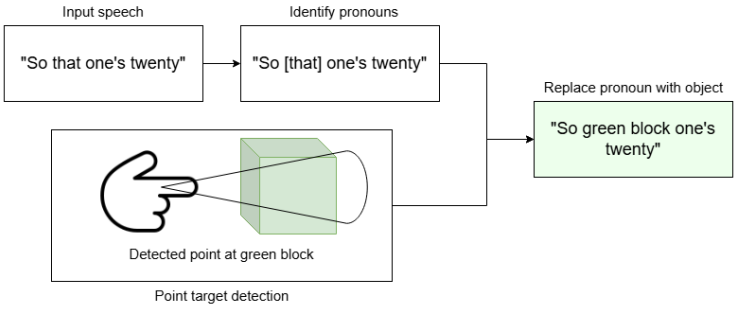

Fig. 5. Example of dense paraphrasing.

4.5 Move Classifier

We use a classifier model to determine when an utterance is a specific dialogue move. Here, we focus on *statement* and *accept*, since these are the two moves considered in the closure rules (see Section 4.7). We use a slightly modified version of the dialogue move classifier presented in [17]. This model sends the target utterance—along with 3 prior utterances for context—through two linear layers, and then a ReLU activation layer. That output is passed through a 512-unit Long Short-Term Memory (LSTM) block. It is then passed through a linear layer, a *tanh* activation, another linear layer, and then SiLU before the output

layer, which classifies the utterace into *statement*, *accept*, or neither. All layers, excluding the classification layer, were trained using a triplet loss with a margin of 1. This was done for 200 epochs with a learning rate of 1e-4. Then, the entire classifier was trained further with cross-entropy loss for 100 epochs with a learning rate of 1e-3, and then 200 more epochs with a learning rate of 1e-4. Hyperparameters were tuned based on a search using one group as validation and another as test. We then used leave-one-group-out validation wherein we trained 10 instances of the model with 9 groups used for training and 1 group for testing. We then averaged the resulting model performance across all 10 instances in order to estimate the average performance of the model on an unseen group, as well as the expected standard deviation on an unseen group.

4.6 Propositional Extractor

To effectively track individual beliefs in co-situated group interactions, we employ a propositional extraction model adapted from prior work on identifying task-relevant assertions in collaborative discourse [44]. The propositional extractor follows a pairwise classification approach inspired by co-reference resolution methods, where utterances and candidate propositions are jointly encoded using a cross-encoder model. The cross-encoder assesses whether a given utterance expresses a specific proposition by computing a contextualized similarity score. We train the model using supervised learning, where ground-truth propositions are labeled within manually transcribed utterances. The model is optimized with binary cross-entropy loss, treating the task as a binary classification problem: given an utterance-proposition pair, the model predicts whether the proposition is entailed by the utterance. For our purposes, we only extract propositions if a block or weight is mentioned. The model is trained for 12 epochs with a learning rate of 1e-6 for the transformer backbone and 1e-4 for the classifier head, ensuring stable learning across data splits.

4.7 Closure Rules

We use logical closure rules to determine the behavior of belief banks when we detect a *statement* or *accept*. We use closure rules modified from the descriptions in [46], such that is an individual makes a statement, they can be assumed to believe that there is evidence for it.[2] When we detect a statement attributable to an individual, we move the associated proposition directly into their belief bank. If a participant agrees with a statement, that statement's propositions will also go directly into their belief bank. Take the aforementioned example of a participant stating "The purple one is thirty", and another participant replying, "Purple is thirty, yeah". The proposition "purple = 30" will be placed in *both* of

[2] Following [17], we adopt similar public announcement logics *a la* Plaza and Baltag [26, 3] and a public announcement operator !, such that given an agent (here, participant) a, proposition φ, belief operator $[B]$ and evidence operator $[E]$, $[!\varphi]_a[B]_a[E]_a\varphi \Rightarrow [!\varphi]_a[E]_a\varphi$.

their belief banks, due to one being a *statement*, and another an *accept* of that statement.

5 Experiments

5.1 Audiovisual Speaker Detection

The Light-ASD model [21] for audiovisual speaker detection (ASD) makes predictions at the frame level. We conduct several experiments to apply these frame-level predictions at the utterance level. For each segment, we use one of the following methods to aggregate the frame scores:

1. **Count**: selecting the candidate with the most non-negative frames;
2. **Mean**: calculating the mean score;
3. **Mean (Positive Only)**: calculating the mean while excluding negative scores;
4. **Sum**: summing all frame scores; or
5. **Sum (Positive Only)**: summing the scores while excluding negatives.

We also include a random guess baseline ("Guess"), where we generated a random participant as our prediction. We evaluate our utterance-level predictions using F1 score and accuracy compared to the ground-truth speaker labels.

5.2 Dense Paraphrasing

We compare ground-truth annotations of dense paraphrasing with the original (non-dense paraphrased) transcript, and our automatic dense paraphrase approach. This allows us to verify whether the output of the automated process is closer, equal to, or further away from the goal than the original text. We compare only the utterances that contained a pronoun and at least one point target object during the time of speaking. Thus these conditions only rendered Groups 1, 2, 4, and 5 fully evaluable (as only those groups have complete object annotations). We evaluate the performance using cosine similarity of embeddings from a pretrained Sentence Transformer model [32] for semantic comparison and Levenshtein distance for literal comparison. This approach allows us to assess both the semantic and token-level differences between the goal state and automatic paraphrasing, as well as between the goal state and original transcripts.

5.3 Individual Belief Tracking

We conduct Individual Belief Tracking (IBT) experiments using the original transcripts, manually dense paraphrased transcripts, and automatically dense paraphrased transcripts described in Section 4.4. This is the final output of our pipeline. Following previous work [17], we evaluate our system using the Dice-Sørensen coefficient (DSC) [8, 34]. This method allows us to compare our final set of propositions in each participant's belief bank with the ground-truth banks. At

this point we take the average score of each group. This shows how each method performs on each group. Next, we pad the remaining dialogue steps by holding the final score constant until all groups reach the max length of dialogue seen in the data (141 dialogue steps). We then take the average of all groups at each step to compare across conditions. This allows us to compare results of each method as the group progresses.

6 Results

6.1 Audiovisual Speaker Detection

Out of our experiments, the best performing approach was to take the candidate with the largest sum of positive numbers. A paired t-test showed that this approach yielded a significant improvement over a random guess strategy ($\alpha = 0.05$, $p = 0.04$). The results of our experiments can be seen in Table 1.

Table 1. F1 and Accuracy Results for Light-ASD Experiments

	F1	SD	Acc.	SD
Guess	.267	.023	.250	.021
Count	.306	.061	.335	.052
Sum	.252	.055	.283	.051
Mean	.252	.053	.282	.048
Sum (Positive Only)	**.320**	.059	**.347**	.058
Mean (Positive Only)	.318	.050	.345	.049

6.2 Dense Paraphrasing

The results of our dense paraphrasing experiments can be seen in Table 2. While the cosine similarity between the dense paraphrase annotations and the automatic dense paraphrasing is higher than that of the dense paraphrase annotations and the original transcript, the Levenshtein distance is also higher. This means that, while the automated approach results in sentences that were semantically closer to the ground truth than the original transcripts goal, these were further from the exact text. We should note, however, the very high standard deviation in Levenshtein distance, indicating the wide range of variations between automatically dense paraphrased sentences and their ground truth counterparts. An example from the data allows us to further explore the results we see here: take an instance of the original transcript being, "Ok so this one is probably twenty, ten ten twenty". The manual dense paraphrase (ground truth) of this text is "Ok so green block one is probably twenty, ten ten twenty". Using the automatic dense paraphrasing method, we get the output "Ok so blue block, green block, red block one is probably twenty, ten ten twenty". Here, the point target

detection selected the blue block and red block in addition to the green block, because they were close together in space. In this case, the automatic dense paraphrase is semantically closer to the ground truth with a cosine similarity of .865 compared to the original transcript at .646. This may be because it contains the correct "green block" target. However, because the participant points to two additional blocks (red and blue), there are extra objects added to the automatic dense paraphrase, resulting in a larger edit distance of 23 compared to that of the the original transcript (12).

Table 2. Comparison of transcripts and automatic paraphrasing against ground truth

	Cosine Sim.	SD	Levenshtein Dist.	SD
Transcript	.766	.184	21.344	21.633
Auto DP	.855	.146	21.500	24.496

6.3 Individual Belief Tracking

Our results show the best performance using the original transcripts with no dense paraphrasing. Interestingly, using the manually dense paraphrased utterances yielded the lowest scores across the board. These can be seen in Table 3. Figure 6 shows the performance of the original transcript and the dense paraphrase methods as participants progress through the task. We see, especially in the original transcript method, an increase during the beginning and middle of the task, and a slight decrease at the end of the task. As the size of the ground truth belief bank expands, the predictions may start to drift away. This may be due to the propositional extractor and move classifier continually admitting items, as we see in Section 6.4. As a result, more false positives are introduced, and the increasing number of true items creates more opportunities for false negatives, leading to greater discrepancies.

Our results of the subsample (Groups 1, 2, 4, and 5) shows the original transcript again outperforms the other methods, though the automatic dense paraphrased method does perform better in Group 2 (see Table 4). We also note that in Figure 7 the automatic dense paraphrasing shows a high increase in the beginning of the task, performing similarly to the transcripts method in the middle of the task, but faces a sharper decrease later in the task. Still, the original transcripts method has outperformed the manual dense paraphrasing and the automatic dense paraphrasing the majority of the time.

6.4 Error Analysis

To investigate the increased error seen with dense paraphrasing, we evaluated our move classifier and propositional extractor. We found that the addition of manual dense paraphrasing increased the total count of propositions by 12%. This

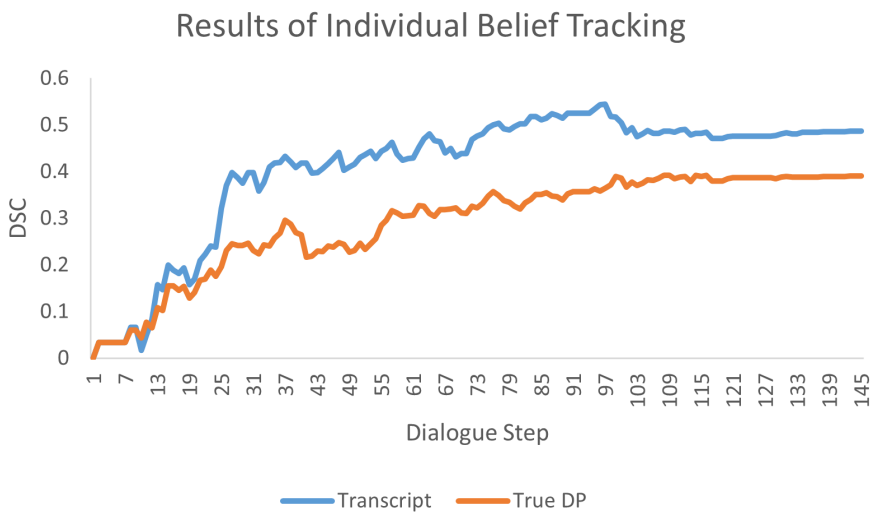

Fig. 6. Results of Individual Belief Tracking using original transcripts and manual dense paraphrasing (True DP).

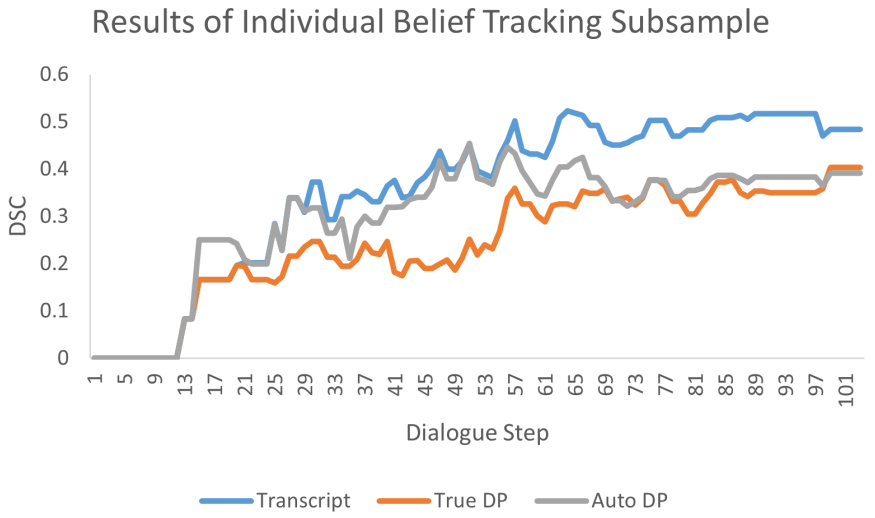

Fig. 7. Results of Individual Belief Tracking using original transcripts, manual dense paraphrasing (True DP), and automatic dense paraphrasing (Auto DP) over a subsample of the WTD.

Table 3. Average DSC of Individual Belief Banks by Group

	Group Number										
	1	2	3	4	5	6	7	8	9	10	Avg.
Transcript	**.484**	**.143**	**.283**	**.384**	**.298**	**.381**	**.349**	**.600**	**.287**	**.287**	**.350**
True DP	.228	.091	.235	.368	.213	.259	.300	.346	.202	.193	.316

Table 4. Average DSC of Individual Belief Banks by Group

	Group Number				
	1	2	4	5	Avg.
Transcript	**.484**	.143	**.384**	**.298**	**.355**
True DP	.228	.091	.368	.213	.245
Auto DP	.418	**.194**	.330	.219	.299

reflects the addition of more task objects due to replacing the pronouns with colored blocks. The propositional extractor will output a proposition whenever a color or weight is mentioned. For example, "Let's try this one" was manually dense paraphrased as "Let's try yellow block one". This resulted in the proposition "yellow > green". This is a hallucination, as the utterance did not make that claim. We additionally found that the move classifier detected statements 84% of the time, and accepts 16% of the time. This means the move classifier was permitting 100% of propositions into the belief banks. The aforementioned hallucination of "yellow > green" was subsequently added to the participant's belief bank. This type of outcome demonstrates the need for both the propositional extractor and the move classifier to be more resilient to negative cases, where a proposition isn't present and a statement was not made.

7 Discussion

Our results show potential for individual modeling in co-situated groups; however, they also underscore an opportunity for growth and improvement toward that end. The audiovisual speaker detection model should be greatly improved before deployment. Part of this may be working toward an utterance-based solution, such as an audiovisual voice activity detection model for segmentation. The automatic dense paraphrasing shows limited results, though it is closer semantically to the target dense paraphrasing than the original transcripts. This shows us that there is a meaningful signal there, but it's not fully developed. The original transcripts had the highest score on individual belief tracking. Surprisingly, the manually annotated dense paraphrasing yielded the lowest score on individual belief tracking. The added context overwhelmed the propositional extractor; that, paired with biased predictions toward statements/accepts in our move classifier, added extraneous beliefs which participants didn't actually express. This is supported by the findings of our error analysis, showing *more*

propositions detected than were made, and the move classifier permitting every dialogue step as a statement or accept. This highlights the need for the propositional extractor to be robust to non-propositional utterances and for the move classifier to reduce false positives. Still, our system shows promising results in individual belief modeling.

7.1 Design Implications

Our work shows feasibility in this design approach while highlighting specific areas in need of improvement for a successful system. We show that, while audiovisual speaker detection (ASD) is a route to speech attribution, the current available frame-by-frame approach does not yield very high performance. This component of the system may find more success in other approaches, such as using utterance-based ASD. This also has the potential to work alongside the automatic segmentation and speaker diarization. Further, this technique relies on visual input; a better approach would allow for speaker diarization even with the visual channel disabled. In addition to this, we find that the point-target detection brings us semantically closer to target dense-paraphrased text. This is notable as it allows for more grounding and context within the speech; audiovisual systems should look for more ways to contextualize the speech channel. However, we found that the dense paraphrased text hurt the performance of our system. After analyzing the output, we found that both our propositional extractor and our move classifier were sensitive to additional information. These components should be more discerning in order to identify when a claim is being made, and what that claim is. The move classifier in particular is meant to act as a selective gate for what should be passed forward, but in our case passed everything forward. This led to adding much more than the intended claims into the belief banks. Therefore, we suggest future work to improve these components of the system.

7.2 Privacy

User privacy is an important consideration in the design of assistive agents, such as processing audiovisual input and tracking beliefs. It is essential that privacy is prioritized throughout the design process, ensuring users have informed control over their personal data. This includes what type of data is being processed and how it is used. To this end, the deployed audiovisual system must provide the ability to toggle each input modality on or off, allowing users to manage their privacy preferences. Additionally, audiovisual data is never stored, further protecting user privacy. Furthermore, the system is designed to respect users' privacy by strictly limiting belief tracking to pre-defined options relevant to the task and are not used for tracking participants outside the task context. Irrelevant or off-topic belief statements are neither tracked nor stored, ensuring that users' unrelated personal information remains private and protected.

8 Conclusion

In this paper, we described desiderata for an agent capable of modeling individuals in co-situated groups. We also presented experiments over our initial system, and uncovered the potential of such an approach, as well as much opportunity for future work. Overall, while we conclude this work is feasible, we also found that our current approach leaves much to be desired in performance; from our results we can see that much of this lies in the need for robust audiovisual speaker detection at the utterance level. While our system runs offline, our methods could be used to the benefit of real-time systems in similar tasks [16].

We should note some limitations to the scope and approach of the work presented herein. Regarding the approach, we explored a frame-level audiovisual speaker detection system for assigning utterances, and our methods for this were relatively simple. Future work should explore an utterance-level approach to audiovisual speaker detection. Given that, we did not apply audio-only speaker diarization methods, which would be a desirable alternate condition to this work. The automatic dense paraphrasing also inserted point target objects in place of single pronouns, but future work should consider sentences with multiple pronouns and objects. We also did not explore a combination of common ground and individual beliefs in this work; future work should consider incorporating both of these simultaneously to better understand the group dynamics and total progression of the task.

Acknowledgments. This research was supported in part by the NSF National 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. The opinions expressed are those of the authors and do not represent views of the Department of Defense, the National Science Foundation, or the U.S. Government.

Disclosure of Interests. The authors have no competing interests to disclose.

References

1. Andrews-Todd, J., Forsyth, C.M.: Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. Computers in Human Behavior **104**, 105759 (Mar 2020). https://doi.org/10.1016/j.chb.2018.10.025, https://www.sciencedirect.com/science/article/pii/S0747563218305156
2. Avci, U., Aran, O.: Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. IEEE Transactions on Multimedia **18**(4), 643–658 (Apr 2016). https://doi.org/10.1109/TMM.2016.2521348, conference Name: IEEE Transactions on Multimedia
3. Baltag, A., Moss, L.S., Solecki, S.: The logic of public announcements, common knowledge, and private suspicions. Springer (2016)
4. van Benthem, J., Fernández-Duque, D., Pacuit, E.: Evidence and plausibility in neighborhood structures. Annals of Pure and Applied Logic **165**(1), 106–133 (2014)

5. Bradford, M., Khebour, I., Blanchard, N., Krishnaswamy, N.: Automatic detection of collaborative states in small groups using multimodal features. In: Proceedings of the 24th International Conference on Artificial Intelligence in Education (2023)
6. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI Meeting Corpus: A Pre-announcement. In: Renals, S., Bengio, S. (eds.) Machine Learning for Multimodal Interaction. pp. 28–39. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2006). https://doi.org/10.1007/11677482_3
7. Cukurova, M., Zhou, Q., Spikol, D., Landolfi, L.: Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough? In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 270–275 (2020)
8. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
9. Goldman, A.I.: Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press (2006)
10. Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the Science of Collaborative Problem Solving. Psychological Science in the Public Interest **19**(2), 59–92 (Nov 2018). https://doi.org/10.1177/1529100618808244, https://doi.org/10.1177/1529100618808244, publisher: SAGE Publications Inc
11. 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: Methods and Approach, pp. 37–56. Educational Assessment in an Information Age, Springer Netherlands, Dordrecht (2015). https://doi.org/10.1007/978-94-017-9395-7_2, https://doi.org/10.1007/978-94-017-9395-7_2
12. Johnson-Laird, P.N.: The history of mental models. In: Psychology of reasoning, pp. 189–222. Psychology Press (2004)
13. Karadzhov, G., Stafford, T., Vlachos, A.: What makes you change your mind? An empirical investigation in online group decision-making conversations (Jul 2022). https://doi.org/10.48550/arXiv.2207.12035, http://arxiv.org/abs/2207.12035, arXiv:2207.12035 [cs]
14. Karadzhov, G., Stafford, T., Vlachos, A.: DeliData: A dataset for deliberation in multi-party problem solving. Proceedings of the ACM on Human-Computer Interaction **7**(CSCW2), 1–25 (2023), publisher: ACM New York, NY, USA
15. Khebour, I., Brutti, R., Dey, I., Dickler, R., Sikes, K., Lai, K., Bradford, M., Cates, B., Hansen, P., Jung, C., others: When Text and Speech are Not Enough: A Multimodal Dataset of Collaboration in a Situated Task. Journal of Open Humanities Data **10**(1) (2024)
16. Khebour, I., Jung, C., Fitzgerald, J., Jamil, H., Krishnaswamy, N.: Feature Contributions to Multimodal Interpretation of Meaning. In: International Conference on Human-Computer Interaction (HCII). Springer (2025)
17. Khebour, I.K., Lai, K., Bradford, M., Zhu, Y., Brutti, R.A., Tam, C., Tu, J., Ibarra, B.A., Blanchard, N., Krishnaswamy, N., others: Common Ground Tracking in Multimodal Dialogue. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 3587–3602 (2024)

18. Krishnaswamy, N., Narayana, P., Bangar, R., Rim, K., Patil, D., McNeely-White, D., Ruiz, J., Draper, B., Beveridge, R., Pustejovsky, J.: Diana's world: A situated multimodal interactive agent. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13618–13619 (2020)
19. Krishnaswamy, N., Pickard, W., Cates, B., Blanchard, N., Pustejovsky, J.: The voxworld platform for multimodal embodied agents. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1529–1541 (2022)
20. Krishnaswamy, N., Pustejovsky, J.: A formal analysis of multimodal referring strategies under common ground. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 5919–5927 (2020)
21. Liao, J., Duan, H., Feng, K., Zhao, W., Yang, Y., Chen, L.: A Light Weight Model for Active Speaker Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22932–22941 (Jun 2023)
22. Murray, G., Oertel, C.: Predicting Group Performance in Task-Based Interaction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 14–20. ACM, Boulder CO USA (Oct 2018). https://doi.org/10.1145/3242969.3243027, https://dl.acm.org/doi/10.1145/3242969.3243027
23. Nath, A., Venkatesha, V., Bradford, M., Chelle, A., Youngren, A., Mabrey, C., Blanchard, N., Krishnaswamy, N.: "any other thoughts, hedgehog?" linking deliberation chains in collaborative dialogues. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 5297–5314 (2024)
24. Pacuit, E.: Neighborhood semantics for modal logic. Springer (2017)
25. Palmer, D., Zhu, Y., Lai, K., VanderHoeven, H., Bradford, M., Khebour, I., Mabrey, C., Fitzgerald, J., Krishnaswamy, N., Palmer, M., Pustejovsky, J.: Speech Is Not Enough: Interpreting Nonverbal Indicators of Common Knowledge and Engagement (2024), https://arxiv.org/abs/2412.05797, _eprint: 2412.05797
26. Plaza, J.: Logics of public communications. Synthese **158**, 165–179 (2007)
27. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and brain sciences **1**(4), 515–526 (1978)
28. Pugh, S., Subburaj, S.K., Rao, A.R., Stewart, A.E., Andrews-Todd, J., D'Mello, S.K.: Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. Proceedings of The 14th International Conference on Educational Data Mining (2021), https://par.nsf.gov/biblio/10494306, publisher: Educational Data Mining
29. Pustejovsky, J., Krishnaswamy, N.: Embodied human-computer interactions through situated grounding. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. pp. 1–3 (2020)
30. Pustejovsky, J., Krishnaswamy, N.: Embodied human computer interaction. KI-Künstliche Intelligenz **35**(3), 307–327 (2021)
31. Pustejovsky, J., Krishnaswamy, N., Draper, B., Narayana, P., Bangar, R.: Creating common ground through multimodal simulations. In: Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication (2017)
32. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), https://arxiv.org/abs/1908.10084
33. Sanchez-Cortes, D., Aran, O., Mast, M., Gatica-Perez, D.: A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. IEEE Transactions on Multimedia **14**, 816–832 (Jun 2012). https://doi.org/10.1109/TMM.2011.2181941

34. Sorensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biologiske skrifter **5**, 1–34 (1948)
35. Stewart, A.E.B., Keirn, Z., D'Mello, S.K.: Multimodal modeling of collaborative problem-solving facets in triads. User Modeling and User-Adapted Interaction **31**(4), 713–751 (Sep 2021). https://doi.org/10.1007/s11257-021-09290-y, https://doi.org/10.1007/s11257-021-09290-y
36. Stewart, A.E., Rao, A., Michaels, A., Sun, C., Duran, N.D., Shute, V.J., D'Mello, S.K.: CPSCoach: The design and implementation of intelligent collaborative problem solving feedback. In: International Conference on Artificial Intelligence in Education. pp. 695–700. Springer (2023)
37. Subburaj, S.K., Stewart, A.E., Ramesh Rao, A., D'Mello, S.K.: Multimodal, Multiparty Modeling of Collaborative Problem Solving Performance. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 423–432. Association for Computing Machinery, New York, NY, USA (Oct 2020), https://doi.org/10.1145/3382507.3418877
38. Sun, C., Shute, V.J., Stewart, A., Yonehiro, J., Duran, N., D'Mello, S.: Towards a generalized competency model of collaborative problem solving. Computers & Education **143**, 103672 (2020). https://doi.org/https://doi.org/10.1016/j.compedu.2019.103672, https://www.sciencedirect.com/science/article/pii/S0360131519302258
39. Sun, C., Shute, V.J., Stewart, A.E.B., Beck-White, Q., Reinhardt, C.R., Zhou, G., Duran, N., D'Mello, S.K.: The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. Computers in Human Behavior **128**, 107120 (Mar 2022). https://doi.org/10.1016/j.chb.2021.107120, https://www.sciencedirect.com/science/article/pii/S074756322100443X
40. Tu, J., Rim, K., Holderness, E., Pustejovsky, J.: Dense Paraphrasing for Textual Enrichment. arXiv preprint arXiv:2210.11563 (2022)
41. Tu, J., Rim, K., Ye, B., Lai, K., Pustejovsky, J.: Dense Paraphrasing for Multimodal Dialogue Interpretation. Frontiers in Artificial Intelligence **7**, 1479905 (2024), publisher: Frontiers
42. VanderHoeven, H., Blanchard, N., Krishnaswamy, N.: Robust Motion Recognition Using Gesture Phase Annotation. In: Duffy, V.G. (ed.) Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. pp. 592–608. Springer Nature Switzerland, Cham (2023)
43. VanderHoeven, H., Blanchard, N., Krishnaswamy, N.: Point Target Detection for Multimodal Communication. In: Duffy, V.G. (ed.) Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. pp. 356–373. Springer Nature Switzerland, Cham (2024)
44. Venkatesha, V., Nath, A., Khebour, I., Chelle, A., Bradford, M., Tu, J., Pustejovsky, J., Blanchard, N., Krishnaswamy, N.: Propositional Extraction from Natural Speech in Small Group Collaborative Tasks. Educational Data Mining Conference (2024)
45. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C., Grundmann, M.: Mediapipe hands: On-device real-time hand tracking. CoRR **abs/2006.10214** (2020), https://arxiv.org/abs/2006.10214
46. Zhu, Y., VanderHoeven, H., Lai, K., Bradford, M., Tam, C., Khebour, I., Brutti, R., Krishnaswamy, N., Pustejovsky, J.: Modeling Theory of Mind in Multimodal HCI. In: International Conference on Human-Computer Interaction. pp. 205–225. Springer (2024)