Feature Contributions to Multimodal Interpretation of Common Ground

Ibrahim Khebour[1 [0009−0009−4374−7263]], Changsoo Jung[1 [0000−0002−2232−4300]], Jack Fitzgerald[1 [0009−0008−5604−7920]], Huma Jamil[1 [0000−0003−1097−4821]], and Nikhil Krishnaswamy[1 [0000−0001−7878−7227]]

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

Abstract. Large Language Models are excellent at processing and extracting semantic information from text. However, to understand the meaning of a real-world interaction, we often need to integrate additional modalities, like gestures, body language, and other non-verbal cues. Here we explore the difficulties that arise with the integration of real-time multimodal processing in AI systems, we also emphasize the disparity between human communication, which seamlessly incorporates multiple modalities, and the current limitations of AI. In this paper, we examine existing works, identify their weaknesses, and propose novel methods that aim to enhance the real-time integration of multimodal data. The results we present indicate that improving AI systems' ability to process multimodal information can lead to apparent advancements in their comprehension capabilities with dynamic and situated environments.

Keywords: Non-verbal features · real-time processing · Multimodal.

1 Introduction

In everyday communication, humans employ various modalities to convey information effectively. While language is often the primary and the most commonly used channel, in situated face-to-face interactions, non-verbal cues such as gestures, facial expressions, body language, sign language and even intonation play a significant role in complementing language and contributing to the interpretation of meaning. These non-verbal cues usually provide extra context, emphasize certain arguments, express emotional valence, or make communication more comprehensive. For instance, a simple gesture or a change in tone can completely change the meaning of an utterance. However, non-verbal cues alone are not always unambiguous or fully-specified, thus making it a challenge for interlocutors to follow.

The field of NLP has made significant strides recently due to the rise of large language models (LLMs), which learn to process and extract semantic information using large-scale neural network training. It has become difficult today to distinguish a conversation generated by an AI, or between an AI and a human from a real conversation that took place between two humans. These AIs have

the ability to generate coherent and linguistically appropriate responses making them valuable tools in a large variety of applications. But despite their outstanding capabilities, LLMs still have a strong bias toward concepts and processes that can be fully expressed in text. Even with the addition of images as a common modality in most recent chatbots, linguistic (and specifically textual) representations still serve as the most common method of information exchange, which poses a serious challenge when the AI has to interpret a real-word interaction where multiple modalities come in to play. The human brain can effortlessly attend to different modalities during a conversation, processing them in real-time to understand the nuances of a dialogue. This skill allows a fluid exchange and the ability to adapt to the changing context. Even modern AI systems find it challenging to integrate real-time multimodal data, as they often require some pre-processing or segmented data, and must keep track of the information being exchanged.

We follow recent work in *common ground tracking* [12], which analyzes small group collaborative dialogues and determines how each utterance affects the shared knowledge of the group. This was demonstrated in the Weights Task Dataset [11], which consists of videos of groups of 3 participants working together to deduce the weights of differently-colored blocks (Fig. 1).

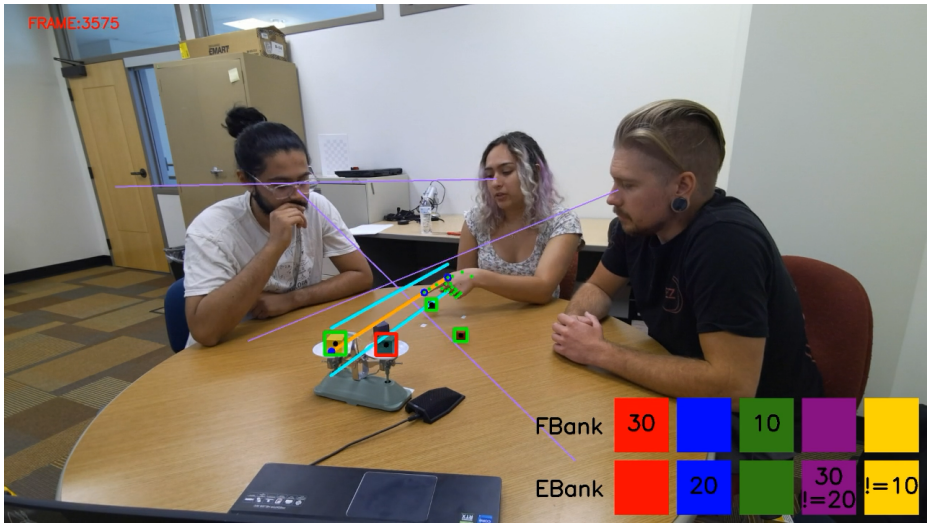

Fig. 1. Participant 2 is pointing at the block on the scale, with the task progression displayed at the bottom right of the frame.

The problem of *common ground tracking* is one of inferring the common task knowledge of a collaborative group, including the pieces of evidence shared in common and the agreed-upon facts. These are denoted in Figure 1 as EBank and FBank, respectively. For instance the contents of the purple square in EBank

is interpreted as "there is evidence that the purple block weighs 30g, and there is evidence that it doesn not weigh 20g." The green square in FBANK is interpreted as "the group currently agrees that the green block weighs 10g" (according to their speech and actions as interpreted by the automated system). The common ground tracking system from Khebour et al. [12] was able to process a multitude of other modalities such as gestures performed by the participants, the actions undertaken with the different task-relevant objects, objects motion throughout the task, the collaborative status of the participants following a Collaborative Problem Solving framework [16], and prosodic features extracted using openSMILE [7] describing the voices of the participants and giving information on their vocal tone. However, it was limited to running offline and postprocessing recorded data given extensive transcription, annotation, etc. We recently developed an online (synchronous) version of the same system, but one that uses speech features only. In this work, we integrate multiple non-verbal features into sychronous common ground tracking and assess the comparative effect of each on modeling the shared understanding of the group.

Through this exploration, we seek to contribute to the growing body of research on multimodal AI systems, providing insights and solutions that could pave the way for more advanced and natural Human-AI interactions in the future.

2 Related Work

Multimodal AI is a prominent and wide-ranging subdiscipline that has captured the interest of a wide segment of the research community. Initial well-known approaches focused on the integration of text and images for tasks like visual question answering and caption generation [9, 23]. More recent expansions into other modalities have included speech, gesture, and physiological signals [21]. For instance, [14] shows how audio and video data could be combined using deep learning to improve speech recognition accuracy.

Multimodal processing frequently entails challenges like human preprocessing of heterogeneous data and asynchronous data streams. In [22], the authors try to understand human intentions from videos by integrating natural language, facial expressions and auditory cues. They highlight the difficulties in multimodal sequence fusion when dealing with temporal asynchrony and modality heterogeneity. Meanwhile, [1] presents a two-stage emotion recognition model that relies heavily on a preprocessing phase that encodes the original dataset with different modalities. However, real-life communication is (almost) instantaneous and dynamic, all while integrating multiple modalities without the need of preprocessing.

When it comes to tracking realistic human-human dialogue, many additional factors must be taken into consideration. An AI model must keep in memory prior dialogues as they can hide additional and important context [5, 10, 13]. The AI system we use to evaluate the contribution of non-verbal features addressed this particular challenge. The purpose behind the task of Common

Ground Tracking (CGT) [12] is to understand the flow of conversation between a group of multiple people, while keeping track of information that has been previously asserted [20], and determining what the group's epistemic consensus toward it is (that is, do they accept the proposition or not?)

Overall, this work addresses consequential challenges of working with multimodal data such as with preprocessing, temporal asynchrony, dialogue state and history tracking, and real-time performance. This study aims to address the limitations identified in the aforementioned prior work, in order to develop methods that empower AI and enables it to narrow the gap in Human-Computer interactions.

3 Approach

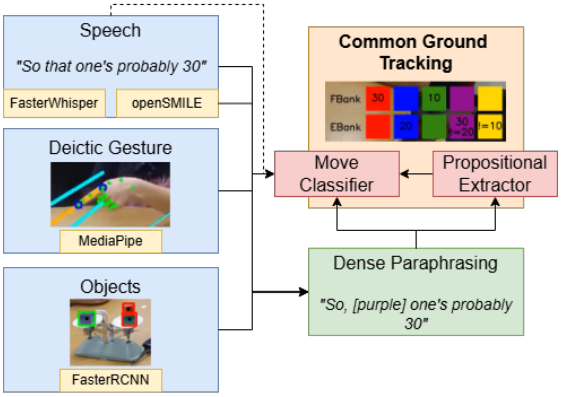

Fig. 2. Diagram of the TRACE system as used in this study.

The platform used to evaluate the contribution of non-verbal features in CGT is known as TRACE (TRAnsparency in Collaborative Exchanges). This system ingests data from various audiovisual channels in order to keep track of the common belief among participants as they go through a collaborative task. Khebour et al.'s original common ground tracking paper [12] is equivalent to an offline version of TRACE, and has a rich input feature set, including speech transcriptions, gesture annotations, object annotations, prosodic features, and facets from a collaborative problem solving coding framework. In this work, we incorporate the full set of features used in the offline CGT task into the real-time version. As some of these modalities were manually annotated, we extract each one automatically while minimizing latency due to inference time, we fine-tune different processing modules to improve the performance, and add additional features.

TRACE considers each modality to be a separate input, and uses an underlying dependency graph to ensure that one modality is processed through

the system once all the features it may need have already been automatically processed. This addresses the asynchrony issue, and is quick and optimized for computation time because we do not process a feature unless a change in its state has been observed. For example, if a model needs textual features, these are only looked up if there is new speech coming in, whereas if the participants remain silent, the speech transcription module will not be called. This method allows TRACE to track the conversation relatively quickly, enabling it to keep up with the ongoing dialogue in real-time. As we add more (specifically non-verbal) features, we remain faithful to this method. We evaluate the contribution of non-verbal features by using the Weights Task Dataset [11], a shared collaborative task where a triad must deduce the weights of differently-colored blocks with the aid of a balance scale.

3.1 CPS Facets

Collaborative Problem Solving (CPS) facets are a way of representing different dimensions of a group's interaction as they contribute to successful problem-solving in a team setting. We use the framework by Sun et al. [16]. Specifically, we used as features the CPS *facets*, or the highest level of this framework's hierarchy. These facets include *constructing shared knowledge*, *negotiation/coordination*, and *maintaining team function*. Successful exhibition of these facets in the course of an interaction are assumed to facilitate the exchange of information, align team efforts, and ensure that all members' opinions are considered, thus enhancing team function, encouraging collective understanding, and enabling teams to tackle complex tasks which would have been challenging for individuals to solve alone. A crucial aspect of CPS is that certain facets may be expressed non-verbally. Intonation, facial expressions and body language all play significant roles in conveying intent, emotion, agreement, confusion or other significant indicators of the current state of the collaboration. These non-verbal cues support verbal communication, making interactions richer, which improves group performance. In this work, we use speech and prosody as input for a random forest model, as first reported in [2], to infer the presence or absence of CPS facets from an utterance.

3.2 Propositional Extractor

Propositions include the semantic content of an utterance that is relevant to defining the state of the task. For instance, one of the participants may have a relatively lengthy utterance that intends to propose a solution such as "I think the blue block weighs the same as the red block is it's also equal to 10 grams", in this case the proposition expressed is $blue = 10$. A key problem in propositional extraction from natural dialogues is that people may have radically different ways of expressing the same underlying semantic content—they use filler words, disfluencies, and have different idiosyncracies and preferences for expressing certain content. This problem was previously addressed by Venkatesha et al. [20],

and we use their system as our propositional extractor. We also use the information of presence or absence of a proposition as a feature for one of TRACE's components (the move classifier—see below) as we observed a high correlation between the feature and the type of move an utterance contains (STATEMENT, ACCEPT, DOUBT).

Propositions are primarily extracted from speech transcriptions, as the semantic content mostly resides in the words spoken. However, like CPS facets, non-verbal features play a role. Due to the situated nature of the Weights Task, a lot of information is expressed using aligned speech and gestures—specifically demonstrative pronouns and deictic gestures (pointing). For instance, $green = 20$ might be expressed by the utterance "I think *that one's* 20 grams" while pointing to the green block. In this case, the transcribed speech alone will not enable the model to recognize which block the participant is referring to. Thus we use a *dense paraphrasing* procedure [17] which decontextualizes the reference by rewriting it with explicit information from other modal channels. Under this transformation, "I think *that one's* 20 grams" plus pointing at the green block gets rewritten to "I think *[green block]'s* 20 grams."

3.3 Move Classifier

The *move classifier* from Khebour et al. [12] is designed to capture multimodal contextual information for the detection of STATEMENT, ACCEPT, and DOUBT classes in dialogues. These classes indicate the epistemic positioning of the speaker toward the utterance, such that $STATEMENT(p)$ is taken to indicate the assertion of evidence for proposition p, $ACCEPT(p)$ signals belief in p, etc. The combination of propositions and the associated epistemic positions expressed by the utterances and other features are used to populate the common ground of shared beliefs and evidence. The original architecture integrates features from multiple modalities, including language, audio, actions, and gesture-based inputs, processing them through modality-specific linear layers, LSTMs, and finally a shared classification head. To address class imbalance during training, the original paper employed SMOTE for oversampling, and used triplet loss for pretraining followed by cross-entropy loss for fine-tuning.

In this work, we adopt the same base architecture with a few key modifications. First, we normalize the audio features derived from OpenSmile to ensure consistent scaling, enabling their effective integration into the model. Additionally, we enhance the LSTM layers by introducing ReLU activations after each linear layer, improving non-linear transformations and providing greater flexibility for learning temporal dependencies across utterances.

Gesture Processing We made substantial alterations to the handling of gestures compared to Khebour et al.'s common ground tracking. The original version uses Gesture Abstract Meaning Representation (GAMR) [4] annotations to capture the interpretation of gestures in context as a meaningful input feature. Previously, annotations from the Weights Task Dataset were used to train models to classify GAMR, however, since this method requires human annotators

and preprocessing, this makes it unsuitable for use in a real-time system. Therefore we avoid the need for human annotators by using RGBD to detect hand movements [18]. We focus on pointing gesture due to their overwhelming prevalence in the data and the amount of salient context they carry relative to the task. When one is detected, we look for the target of that gesture [19] among the detected task-relevant objects (see Section 3.4). We then deterministically follow the GAMR specification using the detected gesturer and target to construct the GAMR annotation.

In the original CGT task, gesture features were encoded using k-hot representations. We replace these with embeddings generated by an attention-based graph encoder-decoder architecture (see Fig. 4). Gesture AMRs are naturally represented as rooted, acyclic directed graphs, making this architecture a more suitable choice compared to k-hot encodings for learning richer GAMR features. In these graphs, nodes represent argument values (e.g., `ARG0`, `ARG1`, `ARG2`), and edges define the type of relationships between these arguments. Specifically, `ARG0` represents the gesturer, `ARG1` represents the content of the gesture, and `ARG2` represents the recipient of the gesture. The gesture type itself serves as the root node, while the argument values act as leaf nodes (see Figure 3). Each leaf node is connected to the root node through bidirectional edges, allowing the leaf nodes to learn not only from the root but also from their neighboring nodes. This bidirectional connectivity ensures that the embeddings effectively capture both local and global dependencies within the graph, enhancing the representation of gesture semantics.

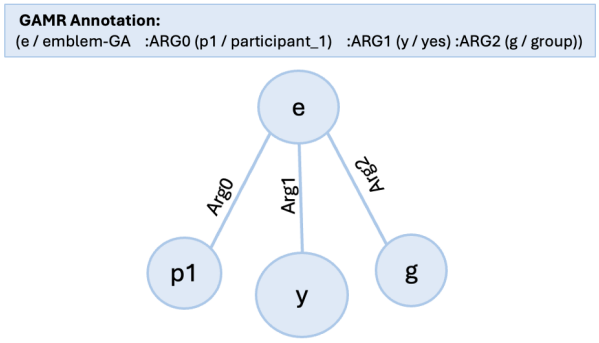

Fig. 3. A GAMR annotation represented as a structured semantic graph.

Graph Autoencoder Architecture We adopt the attention-based message passing mechanism, EdgeGAT, from [24] to construct the encoder. For each node in the graph, attention scores are computed for all neighboring nodes by

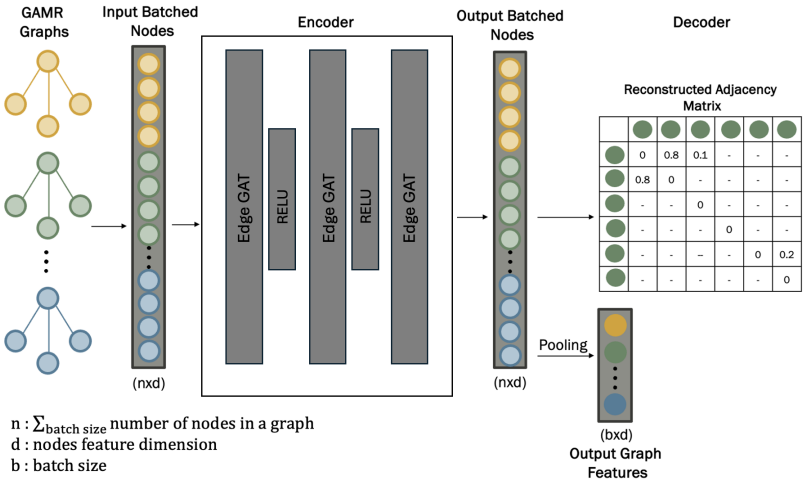

Fig. 4. Attention based graph encoder-decoder architecture.

concatenating node and edge features, passing them through a fully connected layer, and applying Leaky ReLU followed by a softmax function. The neighborhood information is then aggregated using these attention scores, normalized, and combined with the original node feature, weighted by a parameter λ.

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

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

$$\hat{A} = \sigma(ZZ^T), \tag{1}$$

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

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

$$\mathcal{L} = -\frac{1}{|E^+|} \sum_{(i,j)\in E^+} \log \hat{A}_{ij} - \frac{1}{|E^-|} \sum_{(i,j)\in E^-} \log(1 - \hat{A}_{ij}). \tag{2}$$

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

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

$$g = \frac{1}{|V|} \sum_{i \in V} h_i, \tag{3}$$

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

3.4 Object Detector

Data Augmentation Challenges of object detection in our scenarios include the diversity of positions, light conditions, and occlusions of the colored blocks. In collaborative tasks such as the Weights Task, the task-relevant objects are usually located at the center of the task area (here, the table), roughly equidistant to all participants. This means that in most object annotations, they remain stationary or there are non-significant changes in their positions in the image frame. In addition, the colors of the blocks may appear different when they are shadowed, e.g., covered by the participants' forearms when they point or reach toward another block. Additionally, participants' hands frequently occlude blocks when they interact with them, or blocks are occluded by other blocks when placed close together. These factors induce annotation biases which are then transferred to the object detection model trained over those annotations, resulting in missed object detections.

Table 1. Number of Frames on Various Light Conditions

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

Table 2. Number of Frames on Various Gestures

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

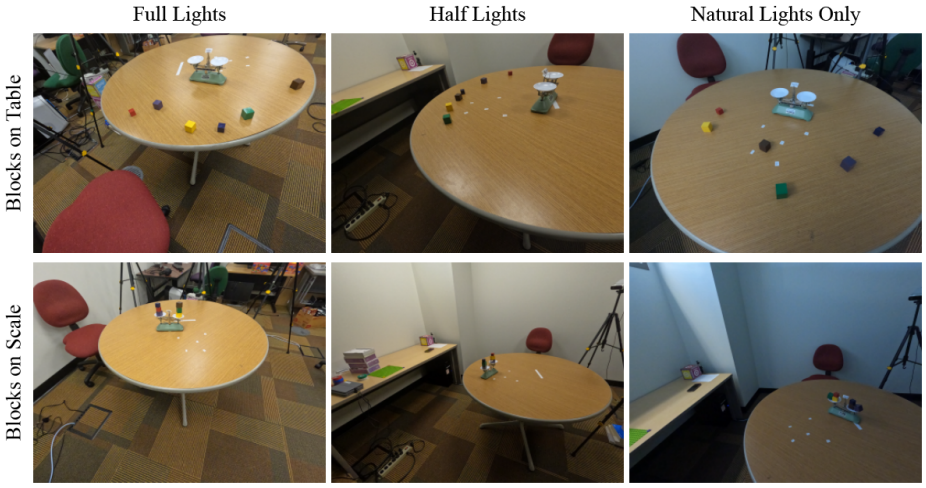

Fig. 5. Additional data collection in variant light conditions.

To overcome the challenges, we collected data on additional samples of corner cases to achieve a more robust model generalizable to various conditions. The additional data was collected twice for different purposes. The first additional set was for sampling light conditions, and the second set targeted occlusions from diverse interactions. Both sets contain frames of the blocks placed on the table and the scale, as occurs in the actual Weights Task. The blocks-on-scale scenario contained many cases of occlusion induced by stacking blocks horizontally and vertically. In addition, we changed the positions of the blocks and scale in every case to have various object positions in each frame.

The first set of augmented data was collected with varied light conditions: full lights, half lights, and natural lights (Figure 5). Table 1 shows how many frames were gathered in the different lighting conditions. In the second set, we sampled four different gestures to address occlusions during interactions. The four gestures included four mainly used in the collaborative task: pinching the center of block, pinching the top of block, covering the top of block, putting the blocks on palms. Table 2 shows the number of samples for various gestures.

The additional data was annotated using SAM2 in the CVAT annotation tool.

Object Detection Model The object detection model used was the `torchvision` implementation of the Faster R-CNN [15] model (`fasterrcnn_resnet50_fpn`). Faster R-CNN uses a backbone network to generate feature maps, in our case this is a ResNet-50 feature pyramid network. Then a smaller convolutional network, referred to as the region proposal network (RPN), slides over the feature maps from the backbone network and generates bounding-box proposals. The RPN region proposals are then fed into a Fast R-CNN [8] detection network that

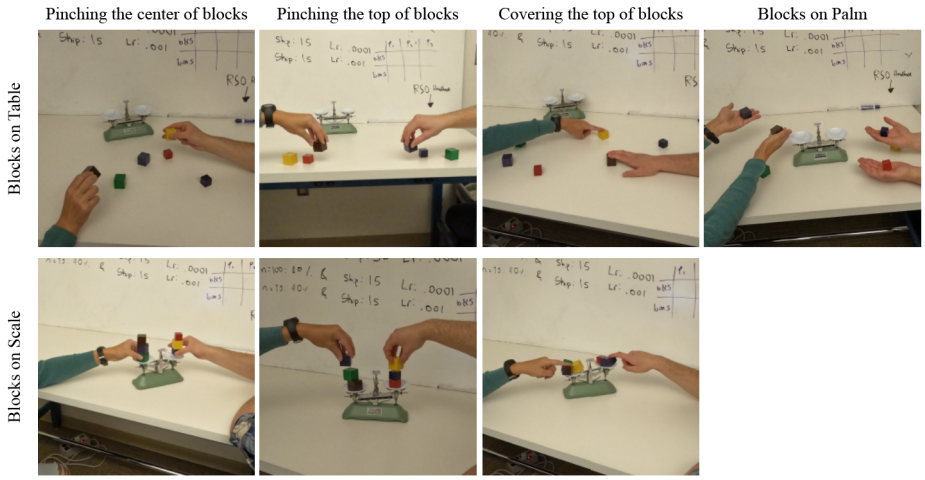

Fig. 6. Additional data collection of blocks under different gesture interactions. The images in this figure have been cropped for better visibility.

ultimately outputs the predicted bounding boxes. In addition to the bounding boxes, the model outputs confidence scores during inference for each predicted bounding box, and we take the bounding boxes that have the highest confidence scores for each class. Due to the ability of the RPN to generate region proposals quickly, Faster R-CNN is a suitable model for detecting objects in real-time.

The `fasterrcnn_resnet50_fpn` model is pretrained on ImageNet-1k [6], and is then further fine-tuned on 300,000 frames from various groups in the Weights Task Dataset (WTD), 5,428 frames from a private demo of the Weights Task, and on the extended light conditions and gesture data described in Section 3.4. The model is fine-tuned for 10 epochs on each set of extra training data using two RTX 3090 GPUs. A test set consisting of 1,786 frames from a separate private demo of the Weights Task is used to measure the performance of the `fasterrcnn_resnet50_fpn`. For each fine-tuning stage, Table 3 shows the global mean average precision (mAP), mAP at an intersection over union (IoU) threshold of 0.5 (mAP_{50}), mAP at an IoU threshold of 0.75 (mAP_{75}), and mean average recall for the model's top 10 predictions based on the confidence scores (mAR_{10}). The mAP metric takes the mean of the average precision for each class at IoU thresholds $\in \{0.5, 0.55, \ldots, 0.95\}$. The base fine-tuning stage represents the model weights after being fine-tuned on the WTD and the private demo data, and the following stages (light conditions, gestures, and light conditions + gestures) are further fine-tuned using the base stage as a starting point.

4 Experiments and Results

Let us first recap and expand the explanation of the common ground tracking (CGT) task. CGT involves identifying the shared beliefs among participants in

Table 3. Faster R-CNN Fine-tuning Performance

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

a task-oriented multimodal dialogue. The input features (here including speech transcriptions, prosodic features, gestures and actions) contribute to the prediction of epistemic positions (or "moves"; Section 3.3) communicated by each participant. These are classified into STATEMENT, ACCEPT, and DOUBT. A set of logical closure rules help construct the final common ground structure, as they direct the expressed propositions into the right bank, or level of belief. These banks are 3 in total:

- Question Under Discussion Bank (QBank): Contains the current set of questions or topics that participants are actively seeking to solve. A DOUBT of a weakly evidenced proposition p may return p to the status of a question under discussion (QUD), though this does not occur in the WTD dialogues.
- Evidence Bank (EBank): Holds propositions for which there is support but not necessarily agreement. A STATEMENT of proposition p introduces evidence for it into EBank. A DOUBT of proposition p if p is already in FBank moves p back down to EBank.
- Fact Bank (FBank): Holds propositions that all participants believe in and have accepted. An ACCEPT of p if p is already in EBank, moves it to FBank. DOUBTs may remove propositions from FBank and send them back to EBank.

To investigate the contribution of each modality, we designed our experiments in a way that isolates them depending on the salient input channels. See Table 4 below. For certain experiments, we assume access to the "ground truth" (human annotations) for certain channels:

- Ground Truth Speech: Experiments using these use ground truth speech transcripts.
- Ground Truth Gesture: Experiments using these use the manually annotated GAMR representations of participants' gestures during the interaction.
- Ground Truth Object: Experiments using these use the manually annotated coordinates of the bounding boxes for the objects.

Table 4. Experiments with additional modalities and evaluation features (speech transcripts feature is always included and so is not shown here, e.g., Experiment 1 is an automatic speech transcription-only baseline).

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

We use Sørensen-Dice Coefficient (DSC) as our primary metric. DSC is an IoU-style metric that normalizes for the sizes of the sets being compared. This is also the primary metric used in Khebour et al.'s original CGT paper [12]. The test dataset is composed of 4 videos out of the 10 that exist in the Weights Task Dataset. These videos (Groups 1, 2, 4, and 5) contained ground truth annotations for all the relevant modalities, enabling a complete suite of experiments. We use a leave-one-group-out experimental format where models were trained over all but one group and evaluated on the remaining group. We calculate the average DSC for each one of the 4 test groups, then we compute the average across those 4 values.

Experiment 1 provides a baseline using only automatically-transcribed speech. Experiment 2 shows the maximum utility of speech alone, as we use the ground truth data. Experiments 1 and 2 show that with automatic transcriptions (Experiment 1), the models can get almost .40 DSC for F ∪ E, and that is over 83% of the potential of speech when using ground truth.

Table 5. Experimental results averaged across test groups. F ∪ E denotes the union of FBank and EBank [12] and this serves as a proxy for extraction of the correct propositional content even if the level of evidence assigned to it is incorrect. Bold shows which feature set performed best for each bank.

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

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

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

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

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

There's a noticeable decrease in performance when we compare Experiment 1 with Experiment 9, even though we added more modalities. This is very different from the results from Khebour et al. [12]. In that version, we see a mix of trends, but the decreases are not as big as with TRACE in real-time. In offline CGT the general trend across the test groups and all 4 values of DSC, is a drop of 1% when more modalities are used, whereas the live model shows an 8% drop. This is explained by the sparsity of the additional modalities in the move classifier. These features also go deeper into the model compared to the dense paraphrase, the outputs of which impact many other components of the TRACE, further indicating the data sparsity issue. The best performing experiment that did not use any ground truth data is Experiment 7, which used automatic speech transcriptions, dense paraphrases with ASR transcripts, automatically detected pointing gestures, and automatically-detected objects, as well as prosodic, CPS, and propositional features. This indicates a plausible best feature set for future work in real-time common ground extraction. GAMR features may also still have utility in a less-sparse data condition.

Table 6. Average DSC cver test groups (calculated from [12]).

Modalities	**Qbank**	**Ebank**	**Fbank**	**F ∪ E**
All modalities	0.714	0.535	0.313	0.851
Speech only	0.725	0.551	0.184	0.928

5 Conclusion

Many works from multiple subdisciplines have shown the importance of multimodality at the intersection of HCI and AI. In this paper we presented the effects of adding more modalities to a real-time multimodal common ground tracker, we have seen that the more modalities we add the more a model is able to approach its limit in performance. We have also proven the necessity for the additional modalities to be as continuous as they possibly can be. The more a modality is sparsely represented relative to another set of modalities, the imbalance creates too much noise for the model to extract the proper features.

We've also seen how that cost an important requirement in data quality which is modality balance, as some modalities have become sparser. This motivates us to investigate other modalities that are present in as many individual frames as possible, such as eye gaze, body pose, and the addition of more task relevant objects such as the scale used in the Weights Task (e.g., whether it is tipping to one side or the other and how that correlates with participant utterances or informs their beliefs). While the GAMR embedding features in our experiments proved too sparse to be truly useful, future work could examine how more continuous gesture features may impact performance of offline CGT, or how they may be contributing other information to the interaction such as level of engagement. We also focus only the *common* ground, or set of shared beliefs, without attributing these to individuals [3].

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. Ai, W., Zhang, F., Meng, T., Shou, Y., Shao, H., Li, K.: A two-stage multimodal emotion recognition model based on graph contrastive learning. In: 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS). pp. 397–404. IEEE (2023)

2. Bradford, M., Khebour, I., Blanchard, N., Krishnaswamy, N.: Automatic detection of collaborative states in small groups using multimodal features. In: International Conference on Artificial Intelligence in Education. pp. 767–773. Springer (2023)
3. Bradford, M., Khebour, I., VanderHoeven, H., Blanchard, N., Krishnaswamy, N.: Modeling Individual Beliefs in Co-Situated Groups. In: International Conference on Human-Computer Interaction (HCII). Springer (2025)
4. Brutti, R., Donatelli, L., Lai, K., Pustejovsky, J.: Abstract meaning representation for gesture. In: Proceedings of the thirteenth language resources and evaluation conference (2022)
5. Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Ultes, S., Ramadan, O., Gašić, M.: Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. arXiv preprint arXiv:1810.00278 (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
7. 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 (2010)
8. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
9. Jabri, A., Joulin, A., Van Der Maaten, L.: Revisiting visual question answering baselines. In: European conference on computer vision. pp. 727–739. Springer (2016)
10. Jacqmin, L., Rojas-Barahona, L.M., Favre, B.: " do you follow me?": A survey of recent approaches in dialogue state tracking. arXiv preprint arXiv:2207.14627 (2022)
11. Khebour, I., Brutti, R., Dey, I., Dickler, R., Sikes, K., Lai, K., Bradford, M., Cates, B., Hansen, P., Jung, C., et al.: When text and speech are not enough: A multimodal dataset of collaboration in a situated task. Journal of Open Humanities Data **10**(1) (2024)
12. Khebour, I., Lai, K., Bradford, M., Zhu, Y., Brutti, R., Tam, C., Tu, J., Ibarra, B., Blanchard, N., Krishnaswamy, N., et al.: Common ground tracking in multimodal dialogue. arXiv preprint arXiv:2403.17284 (2024)
13. Liao, L., Long, L.H., Ma, Y., Lei, W., Chua, T.S.: Dialogue state tracking with incremental reasoning. Transactions of the Association for Computational Linguistics **9**, 557–569 (2021)
14. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., et al.: Multimodal deep learning. In: ICML. vol. 11, pp. 689–696 (2011)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
16. 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)
17. Tu, J., Rim, K., Ye, B., Lai, K., Pustejovsky, J.: Dense paraphrasing for multimodal dialogue interpretation. Frontiers in artificial intelligence **7**, 1479905 (2024)
18. VanderHoeven, H., Blanchard, N., Krishnaswamy, N.: Robust motion recognition using gesture phase annotation. In: International conference on human-computer interaction. pp. 592–608. Springer (2023)

19. VanderHoeven, H., Blanchard, N., Krishnaswamy, N.: Point target detection for multimodal communication. In: International Conference on Human-Computer Interaction. pp. 356–373. Springer (2024)
20. 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. In: Proceedings of the 17th International Conference on Educational Data Mining. pp. 169–180 (2024)
21. Verma, G.K., Tiwary, U.S.: Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. NeuroImage **102**, 162–172 (2014)
22. Yang, D., Li, M., Qu, L., Yang, K., Zhai, P., Wang, S., Zhang, L.: Asynchronous multimodal video sequence fusion via learning modality-exclusive and-agnostic representations. IEEE Transactions on Circuits and Systems for Video Technology (2024)
23. Yang, Z., Yuan, Y., Wu, Y., Cohen, W.W., Salakhutdinov, R.R.: Review networks for caption generation. Advances in neural information processing systems **29** (2016)
24. Zhang, Z., Ji, H.: Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 39–49. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.4, https://aclanthology.org/2021.naacl-main.4/