Modeling Multimodal Social Interactions: New Challenges and Baselines with Densely Aligned Representations
- Supplementary Material -

Sangmin Lee[1] Bolin Lai[2] Fiona Ryan[2] Bikram Boote[1] James M. Rehg[1]
[1]University of Illinois Urbana-Champaign [2]Georgia Institute of Technology
{sangminl,boote,jrehg}@illinois.edu {bolin.lai,fkryan}@gatech.edu

1. Network Structure Details

Table 1 shows the network structure details of the proposed baseline. The point, kinesics, position, and visual interaction encoders are included in the visual interaction modeling part. The multimodal transformer is used for aligned multimodal fusion. The point, kinesics, and position encoders are implemented with Multilayer Perceptron (MLP) structures, comprising fully connected (FC) layers. The visual interaction encoder and the multimodal transformer are based on typical transformer architectures [5]. For the MLP structures, "MLP Size" denotes the output dimension of each FC layer. ReLU activation is applied between FC layers in MLP structures. For transformers, "Hidden Size" refers to the feature size after passing through the feed-forward network, while "MLP Size" represents the intermediate feature size in the Multi-Head Attention (MHA) mechanism. The channel dimensions d_{point} and d are set to 64 and 512, respectively.

Network Structures				
Module	**Layers**	**Hidden Size**	**MLP Size**	**Multi-Heads**
Point Encoder	3	–	64	–
Kinesics Encoder	4	–	512	–
Position Encoder	4	–	512	–
Visual Interaction Encoder	3	512	1024	8
Aligned Multimodal Transformer	2	512	1024	8

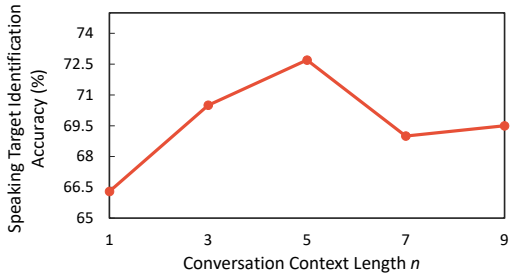
Table 1. Network structure details of the proposed baseline model including MLP and transformer structures.

2. Implementation Details

Language Models. We employ "bert-base-uncased", "roberta-base", and "electra-base-discriminator" as pre-trained BERT, RoBERTa, and ELECTRA language models, respectively. We leverage models and weights of them from Hugging Face [6].
Comparison Methods. For comparative analysis, we utilize (Language Model + MViT) and (Language Model + DINOv2). In the case of (Language Model + MViT), we leverage the visual features from the 24-layer multiscale vision transformer (MViT) [1], pre-trained on the Kinetics-400 video dataset, following the approach in [2]. For (Language Model + DINOv2), we use the visual features pooled from a 3-second window of DINOv2 features [3] (interval 0.5s) along the time axis. Both visual features are integrated with the language feature (*i.e.*, conversation context feature) through FC layers for task-specific predictions according to [2].

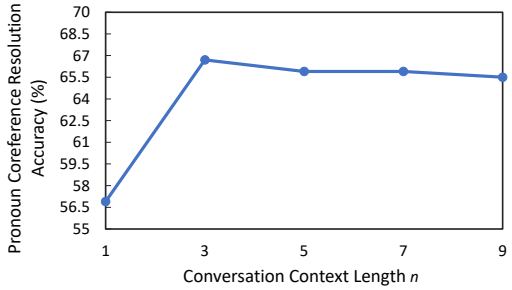
3. Effects of Conversation Context Length

We conduct experiments to investigate the effects of conversation context length n on the performance of each task. Figures 1, 2, and 3 show the validation results for speaking target identification, pronoun coreference resolution, and mentioned player prediction, respectively. When a context length of n is employed, the target utterance is concatenated with n preceding and n following utterances. As shown in the figures, we consistently obtain low performances with the shortest context length of $n = 1$, while achieving fairly good performances with a context length of $n = 5$ for all tasks. Note that we adopt $n = 5$ as our default setting for the baselines. These evaluations were conducted on YouTube dataset using the BERT-based model.



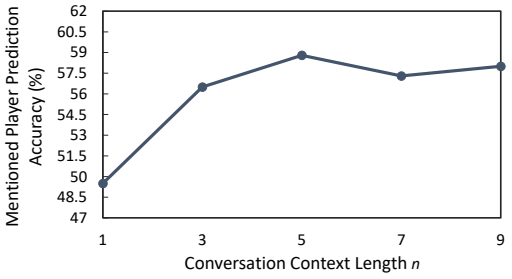
Target Task	Conversation Context Length	Accuracy (%)
Speaking Target Identification	$n = 1$	66.3
	$n = 3$	70.5
	$\boldsymbol{n = 5}$	**72.7**
	$n = 7$	69.0
	$n = 9$	69.5

Figure 1. Effects of conversation context length on the performance for speaking target identification.



Target Task	Conversation Context Length	Accuracy (%)
Pronoun Coreference Resolution	$n = 1$	56.9
	$n = 3$	**66.7**
	$\boldsymbol{n = 5}$	65.9
	$n = 7$	65.9
	$n = 9$	65.5

Figure 2. Effects of conversation context length on the performance for pronoun coreference resolution.

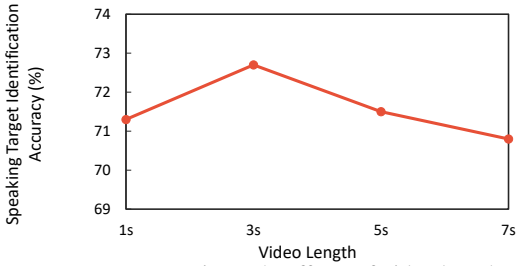


Target Task	Conversation Context Length	Accuracy (%)
Mentioned Player Prediction	$n = 1$	49.5
	$n = 3$	56.5
	$\boldsymbol{n = 5}$	**58.8**
	$n = 7$	57.3
	$n = 9$	58.0

Figure 3. Effects of conversation context length on the performance for mentioned player prediction.

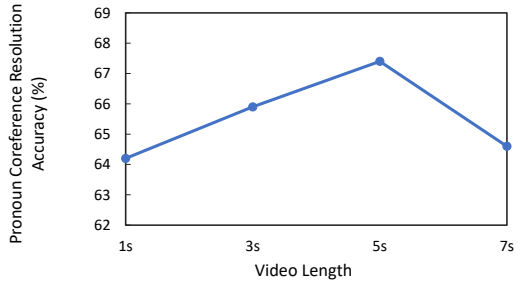
4. Effects of Video Length

We conduct experiments to investigate the effects of video length on the performance of each task. Figures 4, 5, and 6 show the validation results for speaking target identification, pronoun coreference resolution, and mentioned player prediction, respectively. We achieve fairly good performance with a video length of 3 seconds for all tasks. Note that we adopt 3 seconds as our default setting for the baselines. These evaluations were conducted on YouTube dataset using the BERT-based model.



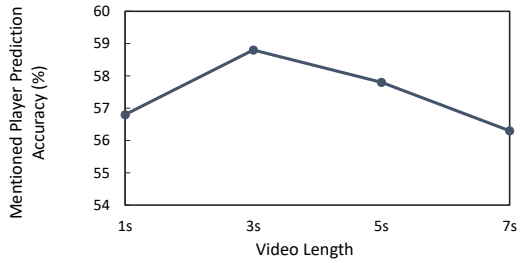
Target Task	Video Length	Accuracy (%)
Speaking Target Identification	1 sec	71.3
	3 sec	**72.7**
	5 sec	71.5
	7 sec	70.8

Figure 4. Effects of video length on the performance for speaking target identification.



Target Task	Video Length	Accuracy (%)
Pronoun Coreference Resolution	1 sec	64.2
	3 sec	65.9
	5 sec	**67.4**
	7 sec	64.6

Figure 5. Effects of video length on the performance for pronoun coreference resolution.



Target Task	Video Length	Accuracy (%)
Mentioned Player Prediction	1 sec	56.8
	3 sec	**58.8**
	5 sec	57.8
	7 sec	56.3

Figure 6. Effects of video length on the performance for mentioned player prediction.

5. Effects of Player Position Correction

We address scenarios where a player is temporarily undetected, such as being offscreen for a short time. In such cases, we proceed with player position encoding by leveraging the corresponding player position stored in a buffer to correct the missing player position. Table 2 shows the experimental results demonstrating the impact of missing player correction on the performance of all three social tasks. As shown in the table, the position correction contributes to improved performance. These experiments were conducted using the BERT-based model on YouTube dataset.

Target Task	Player Position Correction	Accuracy (%)
Speaking Target Identification	✗	70.1
	✓	**72.7**
Pronoun Coreference Resolution	✗	65.3
	✓	**65.9**
Mentioned Player Prediction	✗	58.1
	✓	**58.8**

Table 2. Effects of player position correction on the performances for three social tasks.

6. Data Domain Generalization

We conduct experiments to validate the generalization capability of our proposed approach across two data domains. To this end, we train our baseline on YouTube dataset and evaluate its performance on Ego4D dataset. Table 3 shows the performance results according to the training data. As shown in the table, the model trained on YouTube data performs well on the Ego4D domain and even achieves better results compared to the model trained on Ego4D for all three social tasks. This improvement can be attributed to the larger amount of training data available in YouTube dataset. The experimental results demonstrate the generalization capability of our approach between different data domains and its potential to work in generalized environments. We adopt the BERT-based model for this experiment.

Target Task	Test Data	Training Data	Accuracy (%)
Speaking Target Identification	Ego4D	Ego4D	61.9
		YouTube	**70.5**
Pronoun Coreference Resolution	Ego4D	Ego4D	49.1
		YouTube	**58.0**
Mentioned Player Prediction	Ego4D	Ego4D	50.0
		YouTube	**57.3**

Table 3. Performance results according to the training data types for three social tasks.

7. Additional Quantitative Results

Utilization of Cropped Visual Features. We conduct experiments using cropped visual image features for visual interaction modeling. This approach is based on our dense alignment framework but utilizes cropped CLIP [4] features instead of keypoint features for the speaker kinesics part (green) in Figure 2 of the main paper. The performances achieved with the cropped CLIP features are 71.0% for Speaking Target Identification, 63.6% for Pronoun Coreference Resolution, and 57.7% for Mentioned Player Prediction. These results are lower compared to our proposed baseline with keypoint features, which achieves 72.7%, 65.9%, and 58.8% for the respective tasks. These evaluations are conducted on YouTube dataset using the BERT-based model.

Measurement of Recall and Precision. In addition to accuracy, we further measure the macro-precision and macro-recall performance for our proposed approach. It is worth noting that in the multi-class setting, accuracy represents both micro-precision and micro-recall. The results show that our approach achieves (macro-precision / macro-recall / accuracy) performances of (74.8% / 74.7% / 72.7%) for Speaking Target Identification, (64.9% / 63.3% / 65.9%) for Pronoun Coreference Resolution, and (61.9% / 60.6% / 58.8%) for Mentioned Player Prediction. These results are obtained using the BERT-based model on YouTube dataset. We could achieve the balanced performances across precision, recall, and accuracy metrics in our environment, considering precision vs recall and macro vs micro aspects.

8. Additional Qualitative Results

Figure 7 illustrates examples where our multimodal baseline model utilizing aligned language and visual cues outperforms the language-only model across the three social tasks. Our baseline with densely aligned multimodal representations enables corrected inferences compared to relying solely on language input. The BERT model is employed for these experiments. When we make inferences for these samples with RoBERTa language-only model, the inference results are #3 for STI, #1 for PCR, and #2 for MPP, which means it fails on 2nd and 3rd samples. These results demonstrate cases where the multimodal model (*i.e.*, BERT) with visual reasoning surpasses the more powerful language-only model (*i.e.*, RoBERTa).

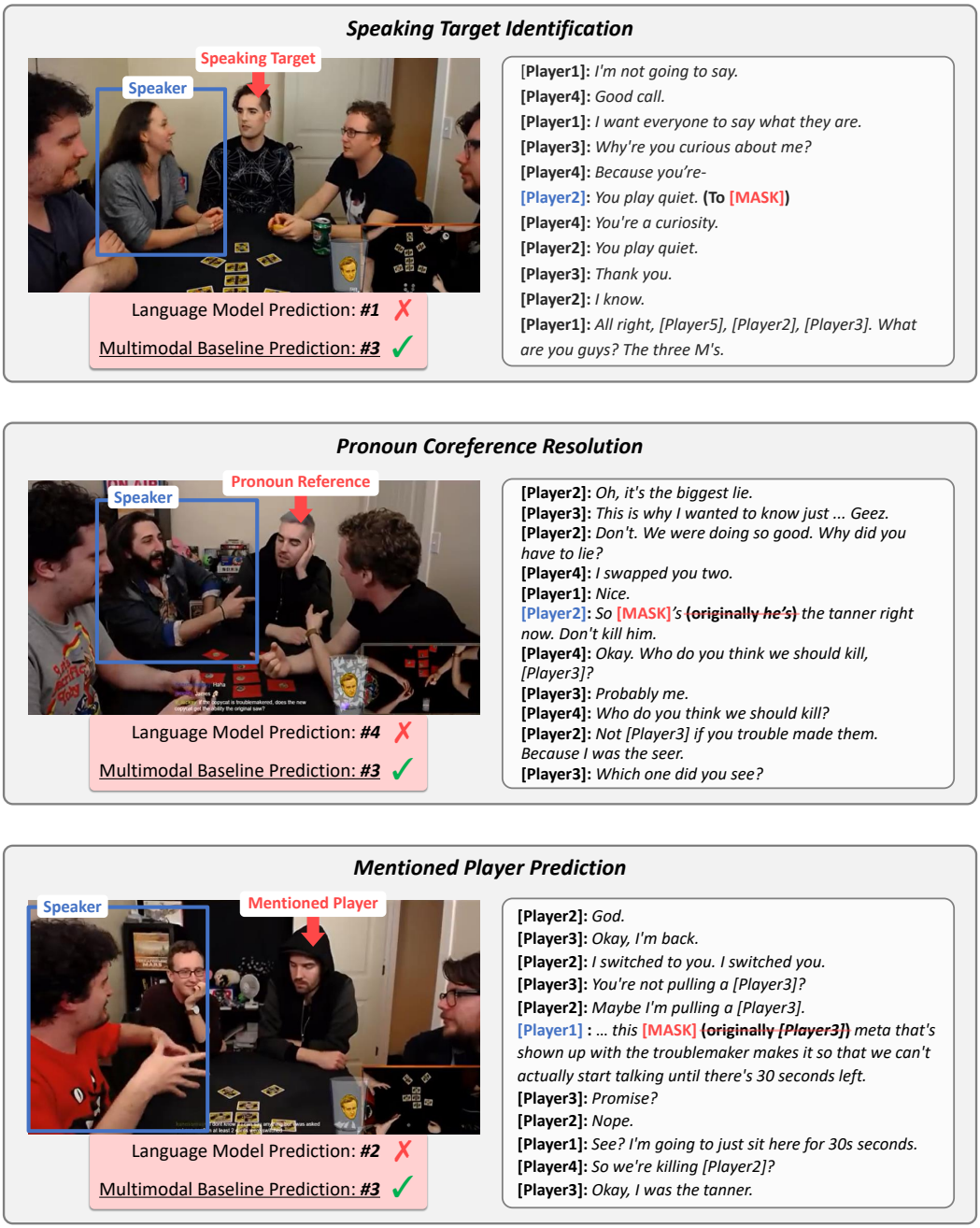


Figure 7. Qualitative results demonstrating the benefit of visual cues in multimodal analysis for three social tasks. Note that Player# are assigned in ascending order from left to right in the visual scenes of this figure.

9. Social Deduction Game Details

We leverage two social deduction game datasets: YouTube and Ego4D. Note that the data collections and annotations have been approved by the Institutional Review Board (IRB). YouTube dataset includes the games of *One Night Ultimate Werewolf* while Ego4D dataset contains the games of *One Night Ultimate Werewolf* and *The Resistance: Avalon*. Below are the details for each social deduction game: Werewolf and Avalon.

One Night Ultimate Werewolf. One Night Werewolf is a social deduction game in which players are secretly assigned to one of two primary factions - the villager team or the werewolf team. During the night phase, players close their eyes and characters with special abilities perform actions like swapping cards before opening their eyes again. The night phase may alter players' roles, though most remain unaware of these changes. Subsequently, players engage in discussion and negotiation to deduce the werewolf's identity. Werewolves, on their part, strive to conceal their identity and mislead others. At the end, everyone votes on who they believe is most suspicious. If at least one werewolf is eliminated, the village team wins, but if no werewolves are eliminated, the werewolf team wins. We refer One Night Ultimate Werewolf game's rules on Wikipedia https://en.wikipedia.org/wiki/Ultimate_Werewolf.

The Resistance: Avalon. This game splits players into two groups: the Minions and the Loyal Servants of Arthur. After roles are assigned secretly via card distribution, players commence with a round where each assumes the Leader role in turns. The Leader's role involves proposing a team for a Quest and all players discuss and vote on approving or rejecting the team assignment. Post the Team Building phase, the designated team decides the Quest's outcome. The Good Team is restricted to using only the Success card in the Quest phase, whereas the Evil Team has the option to use either the Success or Fail card. The Good Team claims victory upon completing three successful Quests, while the Evil Team wins either by causing three Quests to fail or by correctly identifying the character Merlin among the Good Team. We refer The Resistance: Avalon game's rules on Wikipedia https://en.wikipedia.org/wiki/The_Resistance_(game).

References

[1] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6824–6835, 2021. 2

[2] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of the Association for Computational Linguistics (Findings of ACL)*, pages 6570–6588, 2023. 2

[3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 2

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 4

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 1

[6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-Demos)*, pages 38–45, 2020. 2