LONG-TERM SOCIAL INTERACTION CONTEXT: THE KEY TO EGOCENTRIC ADDRESSEE DETECTION

Deqian Kong[†*] *Furqan Khan*[‡] *Xu Zhang*[‡] *Prateek Singhal*[‡] *Ying Nian Wu*[†]

[†] University of California, Los Angeles [‡] Amazon AGI

ABSTRACT

As embodied agents learn to interact, it is crucial for them to understand when, what, and to whom they should respond. While advances in natrual language processing and speech technologies have enabled conversational agents to focus on what to respond, they still struggle to determine when and to whom they should respond. In this paper, we address the addressee detection (Talking-To-Me, TTM) problem under the egocentric view. Instead of relying solely on short-term audio and video data, we propose a simple architecture SICNet with self/cross-modality attention that leverages long-term social interaction context. By leveraging long-term information, our approach has achieved a mean Average Precision (mAP) of 68.98% on the Ego4D TTM task, surpassing the previous state-of-the-art single-task model by 10.07%. We also conducted a detailed ablation study to demonstrate the effectiveness of each component in the long-term social interaction context.

***Index Terms*—** talking-to-me detection, social interaction detection, multimodal analysis, human-centric analysis

1. INTRODUCTION

Addressee detection is a task of identifying the intended target or recipient of an utterance given audio (and sometimes video) information of a conversation [1, 2, 3, 4, 5, 6]. With the rapid evolution of the Metaverse and the development of social robots, there is a growing demand to solve the addressee detection problem from an egocentric view. Our work focuses on the Talking-To-Me (TTM) problem [7] which involves detecting if the target of one utterance is the camera-wearer.

Previous TTM research often take limited audio and video context (generally 1-2s from the start of the speech) as the input to a multi-modal network [7, 8]. However, relying solely on short-term information proves insufficient for addressing the problem. Consider a multi-agent setting where an utterance like "What happens next?" could be directed at any participant. The ambiguity in such cases is often resolvable by incorporating longer-term social interaction context, which includes conversational context and the visual information of all participants involved in the conversation [2, 9].

We handle the long-term social interaction context from both conversational and visual perspectives. By extracting information from long-term audio and video, we utilize a self-attention module to fuse conversational interaction context (including the context of the conversation and the diarization information) and visual social interaction context (including the face and body information for each participant). We name the proposed network the Social Interaction Context (SIC) network. The proposed network outperforms state-of-the-art (SOTA) single task TTM model by 10.07% on mAP and the SOTA multi-task TTM model (which utilizes annotation from other egocentric tasks) by 2.44% [8].

As far as we know, this is the first work showing the importance of long-term social interaction context in addressing the egocentric TTM problem. Our contributions include:

1. Establishing a simple yet effective baseline SICNet, incorporating long-term social interaction context to solve the ego-centric TTM problem.
2. On the Ego4D dataset [7], the proposed SICNet achieves mAP of 68.98% to establish a new state-of-the-art, outperforming the previous SOTA single task TTM model by 10.07%.
3. A detailed ablation study is conducted to evaluate the effectiveness of each component in the long-term social interaction context.

2. METHOD

Given an utterance U_t ending at time t, with corresponding video V and audio A segments from $t-T$ to t, denoted as $V_{[t-T,t]}$ and $A_{[t-T,t]}$, where T is a fixed time interval, the TTM problem seeks to determine if U_t addresses the egocentric agent (the camera wearer). For simplicity, throughout this paper, U_t, $A_{[t-T,t]}$, and $V_{[t-T,t]}$ are referred to as U, A, and V, respectively.

To fully understand the context of the conversation, T should be sufficiently large (we set $T = 10$s instead of 1-2s in the experiment). The major challenge of TTM is then how to effectively extract features from long-term audio A and video V. We propose a conversational context (CC) branch and a visual social context branch (VSC) to encode audio and video

[*] Author performed the work while interned at Amazon.

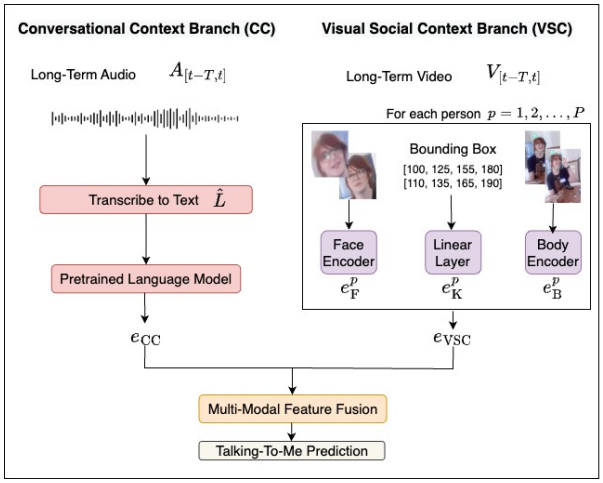

Fig. 1. SICNet: High-level diagram of the proposed Social Interaction Context Network. It consists of two seperate branches: conversational context branch and visual social context branch.

respectively. For CC branch, we leverage the pretrained language model (PLM) to obtain conversational context embedding. For VSC, we utilize pretrained visual feature extractors to aggregate the information of the face, body and position of all participants in the conversation. Those two types of embeddings are fused by a multi-modal self-attention layer to produce a joint embedding, followed by a classifier for the final TTM prediction (see Fig. 1).

2.1. Conversational Context Branch

We leverage a powerful pretrained language model (PLM), specifically RoBERTa [10], to encode long-term audio. The audio clip A is transcribed to text L, using a speech-to-text engine, such as Whisper [11]. Transcribed tokens are then converted into word embeddings using the PLM.

Directly sending the full plain conversation text into the PLM will lose the diarization information of the conversation. To overcome this issue, we introduce a few special tokens to help the PLM better understand the conversation. As multi-modal diarization is not the main focus of the work, we directly use the diarization information in the dataset. In practise, we can leverage some diarization tools [11] for this. We introduce 4 special tokens:

- [C]: the most recent utterance from history
- [H]: the start of the context history
- [EA]: an utterance from the egocentric agent
- [NEA]: an utterance from a non-egocentric agent

[C] and [H] are introduced to distinguish historical context from the current utterance. While [EA] and [NEA] are used to distinguish if one utterance is spoken by the egocentric agent. We find this distinction to be helpful. Specifically, if the previous utterance is spoken by the egocentric agent, it is likely that the current utterance is spoken by a non-egocentric agent. The [EA] and [NEA] information are also easy to derive since we can leverage audio magnitude for this purpose, as the egocentric agent's voice is louder due to proximity to the microphone. Below is an example of the enhanced language token sequence $\hat{L}$, augmented with special tokens to capture the full conversational context:

[H] [EA] How are you going to annotate this dataset? [NEA] Actually we don't need to do the annotation task. [EA] Who will do that? [C] [NEA] Some university students.

The enhanced language sequence $\hat{L}$ serves as the input to the PLM. We first fine-tune the PLM with classification head using cross-entropy loss with the TTM binary label. Once the model is fine-tuned, we freeze its parameters and use it to extract the language features, denoted as e_{CC}, from the second last layer, i.e., $e_{\mathrm{CC}} = \mathrm{PLM}(\hat{L})$. The dimension of e_{CC} is 768 in RoBERTa settings.

2.2. Visual Social Context Branch

For the long-term visual social context (VSC), the input is the video $V = \{I_\tau\}_{\tau=0}^{T}$, where τ is the offset timestamp from the start of the video and I is the image frame. We propose extracting the face and body information for each participant to represent the visual social context. Since face and body detection and tracking is not the major focus of the paper, we directly use the body and face bounding box information provided in the dataset. For each participant p, we extract the following information for all frames in V.

The compact representation of the face: The face crop for participant p at time τ denoted as F^p_τ is passed through a face encoder (e.g., MagFace [12]), and the highest quality embedding is selected as face embedding e^p_{F}. For example, in the case of MagFace, it corresponds to the one with the largest magnitude (Euclidean norm), i.e. $e^p_{\mathrm{F}} = \arg\max_\tau ||\mathrm{MagFace}(F^p_\tau)||_2$.

The position of the body: We concatenate the 4 coordinates of each body box K^p_τ (padding zero if one bbox is missing) and pass the coordinate sequence through a linear layer to obtain a location embedding, i.e. $e^p_{\mathrm{K}} = f_{\mathrm{Linear}}(\mathrm{Concat}\{K^p_\tau\}_{\tau=0}^{T})$.

The compact features of the body: For each body image crop B^p_τ, it is processed through a shared body encoder to get a feature vector $\mathbf{b}^p_\tau$. The feature vectors are concatenated and projected with a linear layer to obtain body embedding $e^p_{\mathrm{B}} = f_{\mathrm{Linear}}(\mathrm{Concat}\{\mathbf{b}^p_\tau\}_{\tau=0}^{T})$.

For each participant p, the face embedding, the location embedding, and the body embedding are concatenated to form a joint embedding $e^p_{\mathrm{joint}} = \mathrm{Concat}(e^p_{\mathrm{F}}, e^p_{\mathrm{K}}, e^p_{\mathrm{B}})$. Those embeddings are then averaged across all participants to obtain the final visual context embedding, i.e. $e_{\mathrm{VSC}} = \frac{1}{P}\sum_p (e^p_{\mathrm{joint}})$ (see Fig. 1), where P is total num-

ber of people in the scene. In our experiments, e^p_{K} and e^p_{B} have the same shape of 128 and e^p_{F} has the shape of 512. Hence, e^p_{joint} has the shape of 768.

2.3. Multi-modal Feature Fusion

We derive two separate embedding vectors from distinct branches: the conversational context branch e_{CC}, and the visual social context branch e_{VSC}. These vectors are then processed through a multi-modal self-attention layer and subsequently, a classification head, leading to the TTM prediction. The network is trained using cross-entropy loss.

3. EXPERIMENTS

3.1. Datasets and Implementation Details

3.1.1. Datasets

We conduct experiments on the Ego4D social interaction benchmark [13], a subset of a large-scale egocentric dataset [7]. The social interaction subset is based on audio-video diarization and comprises 389 training clips (around 32.4 hours), 50 validation clips (around 4.2 hours), and 133 test clips (11.1 hours). Since the test subset is not publicly available and hence we cannot get the audio transcriptions, we report results on the validation set. Each utterance U in the dataset not belonging to the egocentric agent is assigned a binary TTM label. Following [7], it yields a total of $26,791$ training and $2,469$ validation utterances. We allocate 90% of the training utterances for model training and the remaining 10% for hyperparameter tuning.

3.1.2. Implementation Details

Model settings Our model combines conversational and visual context to determine the TTM label. We set the temporal receptive field to be 10s. For the conversational context branch, we use RoBERTa [10] to extract the conversational context embedding. In the visual context branch, we employ MagFace [12] as the face encoder, and a ResNet-18 [14] as the body encoder.

Training details The pre-trained RoBERTa is fine-tuned for 5 epochs with a newly-initiated classification head. We then fix RoBERTa and use the output features before the classification head as the conversational embeddings e_{CC}. The rest of the model, including the body encoder, the linear projection layers for body position and body embedding, the self-attention fusion layer and the final classification layer, is trained using the Adam optimizer [15] with a learning rate of 5×10^{-5} for a total of 50 epochs on 8 Nvidia V100 GPUs.

3.2. Model Performance

We adhere to the Ego4D benchmark [13] for evaluation, which uses mean average precision (mAP) and treats TTM as a two-label classification problem.

Method	mAP(%)
Single-task training	
Random Guess	50.77*
Ego4D-TTM [13]	52.85
TalkNet [16]	57.88*
EgoT2-TS [8]	58.91†
SICNet (Ours)	**68.98**
Multi-task training	
Multi-task [17]	61.91†
Late Fusion [18]	64.29†
EgoT2-g [8]	64.49†
EgoT2-s [8]	66.54†

Table 1. Comparison of SICNet with other approaches. * denotes that the numbers are obtained by reproducing the original code. † denotes the numbers are obtained from [8].

3.2.1. Baseline

We compare the proposed SICNet with the following baseline methods. **Random Guess**: It directly outputs the two-label classification as a Bernoulli trial. **Ego4D-TTM** [13]: It uses ResNet-18 for face features and ResNet-SE to obtain audio embeddings. Audio and video embeddings are then concatenated and passed through a fully-connected layer to predict TTM. **TalkNet** [16]: Different from Ego4D-TTM, it utilizes two cross-attention layers followed by a self-attention layer to combine the video feature and the audio feature. **EgoT2** [8]: It refines the outputs of various models optimized on separate tasks. EgoT2-TS is a TTM task specific model. EgoT2-s and EgoT2-g denote task-specific and task-general translations respectively, both requiring multi-task annotations (such as Looking-At-Me, Audio-Video Diarization). **Multi-task** [17]: It uses hard parameter sharing multi-task learning. **Late Fusion** [17]: It concatenates auxiliary task feature with the primary task feature and finetunes a few layers for the final prediction.

Branches		mAP(%)
CC	VSC	
✓	✗	67.17
✗	✓	63.72
✓	✓	68.98

Table 2. Ablation studies on different components. CC: Conversational Context, VSC: Visual Social Context.

3.2.2. Performance and Discussion

The performance of the proposed SICNet and all baseline methods are presented in Tab. 1. SICNet surpasses all baseline models especially for its single-task training competitors,

Location	Body	Face	mAP(%)
✓	✓	✗	59.35
✓	✗	✓	60.56
✗	✓	✓	61.29
✓	✓	✓	63.72

Table 3. Ablation studies on different components of VSC.

demonstrating its superior performance. The discrepancy arises because the baseline models only leverage short-term audio-visual information struggling to glean higher-order information directly from audio-visual signals due to their limited temporal context.

In contrast, the proposed SICNet employs PLMs for the Conversational Context branch and summarizes long-term video features based on participant behaviors. It helps derive semantic information from long-term audio and video.

3.3. Ablation Studies

3.3.1. Different Context Information

We train various configurations of the SICNet, each excluding one or more branches, to ascertain each branch's contribution to the final performance. The results are summarized in Tab. 2. The full model, encompassing the long-term CC and VSC branches, delivers the optimal performance, indicating the significance of all branches. The superior performance of the CC branch in isolation compared to the models without it highlights the crucial role of conversational context in TTM. Tab. 2 demonstrates that VSC is able to complement CC and improves mAP by 1.8%.

3.3.2. Different Components in VSC

We further examine the efficacy of each component in the VSC branch, as shown in Tab. 3. In this analysis, we don't use CC branch and systematically evaluate the contribution of each VSC component by removing one at a time. Results show that the face embedding from the MagFace encoder is the most valuable among all VSC features, as its exclusion leads to the most significant performance decrease (4.37%). On the other hand, the location embedding derived from bounding box coordinates appears to have the least impact.

Length of CC (s)	mAP(%)
cur (current utterance only)	59.46
5	63.18
10	67.17
15	66.65

Table 4. The impact of different length of conversational context.

3.3.3. Conversational Context Branch

We further examine how different components in the CC branch affecting the final performance (Tab. 5). We first explore using different context lengths (T) cutting at [cur, 5, 10, 15] seconds from the end of the speech in the CC branch, where cur means only the current utterance is used. The result is presented in Tab. 4. Our findings indicate that the CC branch tends to generate the most effective embedding when the context length is approximately 10s. Contexts that are significantly longer tend to introduce excessive and irrelevant historical information.

We assess the impact of language transcription quality by comparing dataset-provided transcriptions with the generated ones using Whisper [11] in Tab. 5. The transcriptions provided by the dataset are significantly inferior to those produced by Whisper because $\sim 58\%$ of utterances lack transcriptions in the dataset. Final performance also reflects the difference, since the model trained with Whisper transcriptions outperforms the model trained with dataset-provided transcriptions by 3.25% on mAP.

Furthermore, we also probe the utilization of the four special tokens we introduced into the conversation context. As Tab. 5 illustrates, incorporating these special tokens into the dialogue history contributes to $\sim 5\%$ improvement in performance.

Special Tokens	Transcription Source	mAP(%)
✗	Dataset Provided	58.83
✓	Dataset Provided	63.92
✓	Whisper	67.17

Table 5. Performance of CC branch alone on the TTM task with different transcription quality, and w/wo the special tokens.

4. CONCLUSION

In this paper, we aim to tackle the Talking-to-Me identification problem, which involves identifying which conversations are directed towards the egocentric agent. Addressing the TTM problem necessitates understanding the long-term conversation as well as the interaction among individuals. To address this, we introduce the SICNet, which effectively models the long-term visual and conversational contexts and merges the both information through multi-modal feature fusion. Our multi-modal SICNet sets a robust state-of-the-art on the Ego4D benchmark. However, considering the overall TTM prediction performance, we are still some distance away from fully resolving the TTM problem. We believe that the development of an improved language model (i.e. GPT-4 [19]) integrated with a more potent social contextual model, can significantly enhance performance.

5. REFERENCES

[1] TJ Tsai, Andreas Stolcke, and Malcolm Slaney, "Multimodal addressee detection in multiparty dialogue systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2314–2318.

[2] TJ Tsai, Andreas Stolcke, and Malcolm Slaney, "A study of multimodal addressee detection in human-human-computer interaction," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1550–1561, 2015.

[3] Oleg Akhtiamov, Maxim Sidorov, Alexey A. Karpov, and Wolfgang Minker, "Speech and Text Analysis for Multimodal Addressee Detection in Human-Human-Computer Interaction," in *Interspeech*. 2017, pp. 2521–2525, ISCA.

[4] Oleg Akhtiamov, Dmitrii Ubskii, Evgeniia Feldina, Aleksei Pugachev, Alexey Karpov, and Wolfgang Minker, "Are you addressing me? Multimodal addressee detection in human-human-computer conversations," in *International Conference on Speech and Computer*. Springer, 2017, pp. 152–161.

[5] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva, "With whom do I interact? Detecting social interactions in egocentric photo-streams," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2959–2964.

[6] Alircza Fathi, Jessica K Hodgins, and James M Rehg, "Social interactions: A first-person perspective," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1226–1233.

[7] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al., "Ego4D: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18995–19012.

[8] Zihui Xue, Yale Song, Kristen Grauman, and Lorenzo Torresani, "Egocentric video task translation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),*, 2023, pp. 2310–2320.

[9] Jixu Chen, Ming-Ching Chang, Tai-Peng Tian, Ting Yu, and Peter Tu, "Bridging computer vision and social science: A multi-camera vision system for social interaction training analysis," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 823–826.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," Tech. Rep., Tech. Rep., Technical report, OpenAI, 2022.

[12] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14225–14234.

[13] "Ego4D Social Interaction Benchmark," `https://github.com/EGO4D/social-interactions`, 2022.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[16] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3927–3935.

[17] Sebastian Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[18] Minghuang Ma, Haoqi Fan, and Kris M Kitani, "Going deeper into first-person activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1894–1903.

[19] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.