# Technical Report on Ego4D Talking To Me Challenge @ CVPR 2023

Zhengqi Gao, Duane S. Boning
Department of EECS
Massachusetts Institute of Technology
{zhengqi, boning}@mit.edu

## Abstract

*This technical report presents our solution for the Ego4D Talking To Me (TTM) challenge held at CVPR 2023. In order to address this challenge, we have introduced a novel technique termed double model ensemble (DME). Motivated by the correlation between different tasks, such as the close relationship between the Looking at Me (LAM) and the TTM task, we ensembled two task-specific models (one designed for TTM and the other for LAM) to leverage this correlation, yielding one fused model for the TTM task. With this model available, we further performed model-level ensemble upon it by employing weighted summation or random mixing. Through the implementation of our proposed DME approach, we secured the* ***second*** *position in the TTM challenge, achieving 59.66% mAP and 57.87% accuracy on the unseen test dataset.*

## 1. Introduction

In the realm of video understanding, the primary focus has been on action recognition within the third-person perspective. However, the introduction of the extensive egocentric dataset, Ego4D [2], marks a significant shift towards capturing the intricate aspects of human-human and human-object interactions. The Ego4D challenge encompasses a wide range of tasks, classified into five main categories: (i) Episodic Memory, (ii) Hands and Objects, (iii) Audio-Visual Diarization, (iv) Social Understanding, and (v) Forecasting. The challenges within the Social Understanding category, i.e., the Talking to Me (TTM) and the Looking at Me (LAM) challenges, require us to determine whether someone in the scene is speaking to or looking at the camera wearer, given an egocentric video clip.

We observe that various tasks exhibit certain correlations [4, 5]. For example, individuals who direct their gaze towards the camera wearer are more likely to engage in conversation with them compared to those who do not. This observation serves as a motivation for us to ensemble models specifically designed for the TTM and the LAM tasks. The utilization of task synergy has also been explored in the existing literature such as [4, 5]. Besides task-level ensemble, we also adopt ensemble at the model level to further boost our model performance. In the next section, we will explain our double model ensemble (DME) approach.

## 2. Proposed Method

Fig. 1 shows a simplified schematic of the overall proposed DME approach. Our approach consists of two stages.

In the first task-level ensemble stage, we obtained two pre-trained baseline models designed for TTM and LAM tasks from the organizer. [1] The two models are frozen and only utilized to extract features during training. Specifically, given one input video clip, we employ the LAM and TTM model to extract features before the temporal pooling layer. The two task-specific temporal features are then passed through a linear projection layer, enriched with learnable positional embeddings, and fed as input to one transformer layer. We set the hidden dimension and dropout rate of the transformer layer as 128 and 0.5, respectively. Subsequently, we average the tokens output by the transformer layer and append a linear layer for outputting the final prediction. Note that we only optimize the projection, transformer and output layers with respect to the final classification loss while keeping the two task-specific models fixed, resulting in a small training computation cost. In all, in the first stage, we ensemble task-related information to augment TTM model performance.

In the second model-level ensemble stage, we construct two 'samples' of the aforementioned fused model, such as by training the neural networks twice with different initializations, or saving two models at different epochs during one training. We experimented with two different model ensemble approaches in our study. The first method involves computing a weighted sum of the soft predictions given by the two 'sample' models. The weight ratio, denoted as $\alpha$, is a hyper-parameter that varies within the range of $[0, 1]$.

[1] The models could be retrieved from https://github.com/EGO4D/social-interactions.

In the second method, we employ a probabilistic ensemble of the two 'sample' models. With a probability of $\beta$, the prediction from the first model is selected as the final prediction, while with a probability of $1-\beta$, the prediction from the second model is chosen. The hyper-parameter $\beta$ ranges in $[0, 1]$. It should be noted that the first method follows the common approach of model ensemble [1], typically applied to independent models (e.g., models with different architectures). However, in our case, due to the fixed nature of the backbone LAM and TTM models in the 'samples', independence is not possible. On the other hand, the second method, though seem less conventional, has been established in the statistics community as an effective technique for Neyman-Pearson hypothesis testing [3].

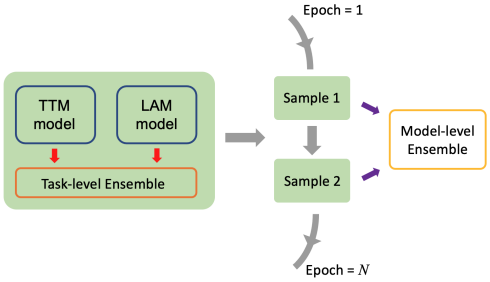


Figure 1. An illustration of our proposed DME approach.

## 3. Numerical Results

Table 1. Results when using weighted sum in model ensemble

| $\alpha$ | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|
| Test mAP (%) | 59.16 | 59.66 | 59.66 | 59.49 |
| Test Acc (%) | 58.13 | 57.87 | 57.80 | 57.11 |

Table 2. Results when using random mixing in model ensemble

| $\beta$ | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|
| Test mAP (%) | 57.50 | 59.59 | 56.70 | 56.51 |
| Test Acc (%) | 57.90 | 58.09 | 57.46 | 57.27 |

In the first task-level ensemble stage, we tried various hyper-parameter values when training the TTM and LAM models, and the transformer layer, and found that they only had minor effects on the performance. In the second model-level ensemble stage, we experimented with the aforementioned weighted sum and random mixing approaches. Tables 1 and 2 comprehensively list the results of the two model ensemble approaches. We mention that without using the DME approach, the best model we could get achieves about $57\%$ mAP. Thus, Tables 1 and 2 reveal that both weighted sum and random mixing could boost the model performance. Furthermore, we notice that in the weighted sum approach, the hyper-parameter $\alpha$ seems to have a relatively stable and positive effect on the model performance across the range $[0.1, 0.4]$. On the contrary, the overall ensembled model performance drops quickly when $\beta$ approaches to $0.4$.

## 4. Limitation and Future Work

Our current method is specifically designed for the TTM challenge. In the future, we plan to extend its applicability to a variety of different video tasks.

## 5. Conclusion

In this technical report, we present the DME approach for the Ego4D Talking to Me (TTM) challenge at CVPR 2023. Our proposed approach resulted in second place in the TTM challenge, achieving a 59.66% mAP and 57.87% accuracy on the unseen test dataset.

## Acknowledgement

This project was supported in part by Millennium Pharmaceuticals, Inc. (a subsidiary of Takeda Pharmaceuticals).

## References

[1] Mudasir A Ganaie, Minghui Hu, AK Malik, M Tanveer, and PN Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 2

[2] Kristen Grauman, Michael Wray, Adriano Fragomeni, Jonathan PN Munro, Will Price, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, et al. Around the world in 3,000 hours of egocentric video. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[3] Erich L Lehmann. The fisher, neyman-pearson theories of testing hypotheses: one theory or two? *Journal of the American statistical Association*, 88(424):1242–1249, 1993. 2

[4] Zihui Xue, Yale Song, Kristen Grauman, and Lorenzo Torresani. Egocentric video task translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2310–2320, 2023. 1

[5] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 1