Supplementary materials for Ex2Eg-MAE

Minh Tran[1], Yelin Kim[2], Che-Chun Su[2], Cheng-Hao Kuo[2], Min Sun[2,3], and Mohammad Soleymani[1]

[1] University of Southern California, Playa Vista, CA, USA
[2] Amazon Lab 126, USA
[3] National Tsing Hua University, Taiwan

A Datasets

Ego4D. Ego4D [3] is an extensive dataset of egocentric videos, boasting over 3600 hours of content capturing various daily life activities across diverse settings — households, outdoor environments, workplaces, and more. The dataset includes recordings from 931 distinct camera wearers across 74 locations. A subset of around 100 hours (with *unblurred* faces) is dedicated to social interactions, showcasing natural scenarios like dining, shopping, and board games involving multiple individuals. Given its real-world setting, the videos exhibit substantial egomotions and limited face quality. The dataset accommodates 14 challenges spanning 5 benchmarks. Among these, *Look-at-me* and *Talk-to-me* constitute the *social understanding* benchmark, designed with the aim of fostering the development of more capable virtual assistants and social robots. The dataset provides tracked face bounding boxes for all visible faces in each frame for the social interactions subset.

Talk-to-me: The goal of the *talk-to-me* task is to identify if a visible face is talking to the camera wearer, given a corresponding video and audio segment. The *talk-to-me*'s annotations come from the vocal activity annotation for the Audio-Visual Diarization (AVD) benchmark, followed by the classification of whether the speech segment is directed towards the camera wearer. Detailed statistics of TTM labels are available in the original Ego4D paper [3]. In summary, the training set of TTM contains around 1.2M frames, with around 75% of them labeled as TTM.

Look-at-me: The goal of the *look-at-me* task is to identify if a visible face is looking at the camera wearer, given a corresponding video (and no audio). The *look-at-me*'s annotations come from the segment-level annotation of when an individual is looking at the camera wearer. Detailed statistics of LAM labels are available in the original Ego4D paper [3]. In summary, the training set of LAM contains around 7.1M frames (with 7.5% labeled as positive).

For both challenges, a fixed set of 389 clips (32.4 hours) are selected for training, 50 clips (4.2 hours) are selected for validation, and 133 clips (11.1 hours) are selected for testing. The test set results must be submitted to an independent server for evaluation purposes[4][5].

[4] https://eval.ai/web/challenges/challenge-page/1624/overview
[5] https://eval.ai/web/challenges/challenge-page/1625/overview

EasyCom. The EasyCom dataset offers a distinct collection of over 5 hours of synchronized egocentric multi-channel audio and video data, specifically designed for AR glasses applications in noisy environments like restaurants or social gatherings. Each session involves a host (camera wearer) interacting with 3 to 5 participants around a table for 30 minutes. The activities are improvised, starting with self-introductions and progressing through tasks like food ordering, puzzle solving, and game playing. The dataset provides annotations of Voice Activity, Speech Transcriptions, Target of Speech, and Face Bounding Boxes.

Using the provided speech transcriptions, we derive both *talk-to-me* and *Active Speaker* labels. Specifically, we leverage the timestamps of each speaker's utterance in the transcript to define their positive ASD segment, designating the rest as negative. To refine *talk-to-me* labels, we assess the speech's target, identifying whether it is directed towards the camera wearer or not. As a high-level overview, the dataset contains more than 250K frames, with around 30% labeled as positive-ASD and 18% labeled as positive-TTM.

B Implementation Details

We follow most pre-training setup of MARLIN [2], including initializing our model with VideoMAE's weights [8]. However, we train our model for 800 epochs with *tube masking*. Our empirical analysis suggests that Ex2Eg-MAE achieves the best performance with a masking rate of 75%. We empirically set $\lambda_1 = 1.0, \lambda_2 = 0.1, \lambda_3 = 0.01$. During fine-tuning and inference, the decoder and PSEMs are discarded. For experiments with the EasyCom dataset, we use AdamW optimizer [6] with a learning rate of $1e^{-4}$ for 10 epochs. For experiments with the Ego4D dataset, we follow the experimental setups of the challenges' baselines [3]. Unless otherwise noted, we report results for the `tiny` architecture for the self-supervised learning methods (around $30M$ parameters). Following the Ego4D Challenges, we use the mean Average Precision (mAP: primary metric) and Accuracy (Acc-1: secondary metric) as the evaluation metrics. **Ego4D-TTM.** We follow closely the pre-processing pipeline proposed by Lin *et al.* [5] (SOTA solution at CVPR23's Ego4D-TTM Challenge). In particular, the method uses a facial landmark detection model [1] to assess the quality of faces within the frames of Ego4D. We filter out low-quality frames when face quality falls below a certain threshold. We do not apply any augmentation during the training process. For each frame, we use the previous and following 8 frames (around 1s) as the context video for prediction. For model architecture, we simply add a linear head on top of the video encoder for classification. We follow QuAVF's optimization settings with the Adam optimizer [4], learning rate of $1e^{-5}$, batch size of 32 for 100 epochs.

Our training for the audio modality is similar to QuAVF [5]. In our audio-visual experiments, we separately train the audio and visual branches. Subsequently, we merge the predictions from both branches using a SVM trained on the predictions of the validation set. Since the final prediction is made frame-by-frame, we further apply a Gaussian filter to smooth the TTM predictions.

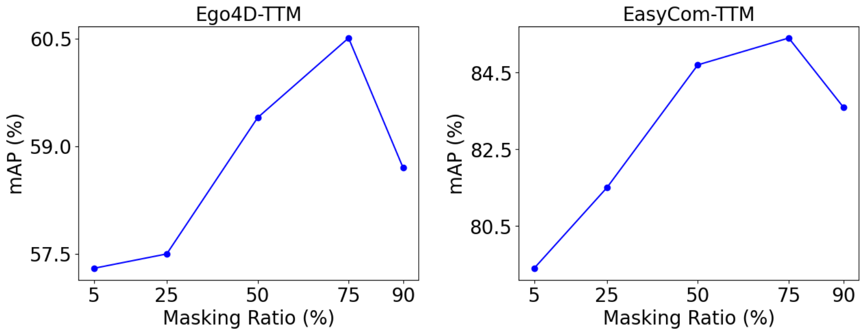

Fig. 1: Impact of mask ratio on downstream performance.

Ego4D-LAM. We follow the experimental setting of the official baseline [3]. Similar to TTM, for each frame, we use the previous and following 8 frames (around 1s) as the context video for prediction. For model architecture, we simply add a linear head on top of the video encoder for classification. For optimization, we uses the Adam optimizer [4], learning rate of $5e^{-6}$ for 40 epochs. For post-processing, we also apply a Gaussian filter to produce smooth predictions, as in TTM. We derive the face quality metric (similar to TTM) for LAM's test set. In cases where faces are undetected or possess low confidence scores, we substitute the lost frame with the nearest one and generate predictions for the frame accordingly.

EasyCom-TTM & EasyCom-ASD. Since the EasyCom dataset lacks a predefined task for its visual modality, we generate the frame-level labels for ASD and TTM as described above. We proceed by creating non-overlapping segments of 16 frames in length, along with their respective frame-level labels, serving as inputs and targets for our models. We use the same model architecture as Ego4D-TTM/LAM, excluding the mean-pooling process applied to the encoder's feature representations to generate frame-level predictions. The dataset is partitioned based on sessions, employing a split ratio of around 60%/20%/20% for train/validation/test set. For optimization, we uses the AdamW optimizer [6], learning rate of $1e^{-4}$ for 10 epochs. Given the relatively high quality of face data in EasyCom (compared to Ego4D), we do not any post-processing methods to the final predictions.

C Mask Ratio Ablation

We provide ablation study on the impact of mask ratio and downstream performance in Figure 1, which identifies 75% as the optimal mask ratio. We attribute the lower mask ratio compared to existing literature (95% for VideoMAE [8] and 90% for MARLIN [2]) to the task's increased difficulty in transforming from the altered perspective back to the original perspective. Furthermore, to ensure effective learning during the distillation process, a substantial amount of information must be presented. Nevertheless, a low mask ratio could render the

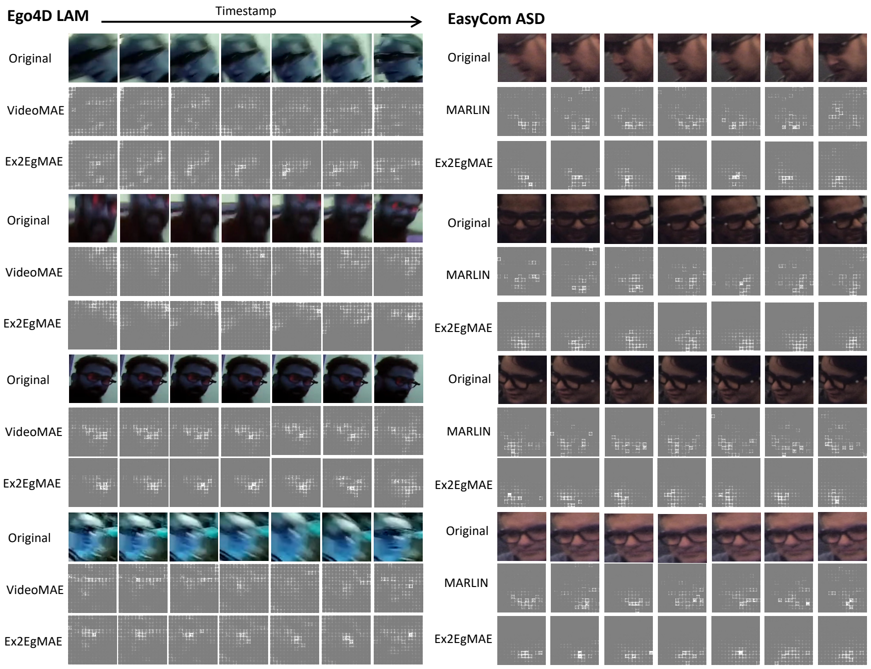

Fig. 2: Saliecy maps generated by *SmoothGrad* [7] of Ex2Eg-MAE and MARLIN/VideoMAE on EasyCom ASD and Ego4D LAM.

reconstruction task trivial, leading to poor representations and potentially introducing more artifacts from the face synthesis module into the pre-training process.

D More qualitative results

We provide more qualitative results in Figure 2. We select challenging sequences showcasing various egomotions or blur effects. We use *SmoothGrad* [7] to identify regions of the original images that the models focus on. Generally, in typical egocentric scenarios with noticeable egomotions, Ex2Eg-MAE consistently generates more dependable saliency maps, directing attention accurately towards relevant areas like the eyes/head pose for LAM or the mouth for ASD.

E Limitations

First, our method is restricted to a facial encoder, primarily due to its dependency on the novel-view synthesis module. In broader contexts, it is much more challenging to generate realistic images, impacting the overall effectiveness of our approach. Second, our study assumes reliable face bounding boxes, which are provided with the evaluation datasets. However, extracting precise face bounding boxes from egocentric videos can be challenging in real-world scenarios.

References

1. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE international conference on computer vision. pp. 1021–1030 (2017) 2
2. Cai, Z., Ghosh, S., Stefanov, K., Dhall, A., Cai, J., Rezatofighi, H., Haffari, R., Hayat, M.: Marlin: Masked autoencoder for facial video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1493–1504 (2023) 2, 3
3. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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. pp. 18995–19012 (2022) 1, 2, 3
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 2, 3
5. Lin, H.C., Wang, C.Y., Chen, M.H., Fu, S.W., Wang, Y.C.F.: Quavf: Quality-aware audio-visual fusion for ego4d talking to me challenge. arXiv preprint arXiv:2306.17404 (2023) 2
6. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018) 2, 3
7. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017) 4
8. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022) 2, 3