Ex2Eg-MAE: A Framework for Adaptation of Exocentric Video Masked Autoencoders for Egocentric Social Role Understanding

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

Abstract. Self-supervised learning methods have demonstrated impressive performance across visual understanding tasks, including human behavior understanding. However, there has been limited work for self-supervised learning for egocentric social videos. Visual processing in such contexts faces several challenges, including noisy input, limited availability of egocentric social data, and the absence of pretrained models tailored to egocentric contexts. We propose Ex2Eg-MAE , a novel framework leveraging novel-view face synthesis for dynamic perspective data augmentation from abundant exocentric videos and enhance self-supervised learning process for VideoMAE [57] via: 1) reconstructing exocentric videos from masked dynamic perspective videos; and 2) predicting feature representations of a teacher model based on the corresponding exocentric frames. Experimental results demonstrate that Ex2Eg-MAE consistently excels across diverse social role understanding tasks. It achieves state-of-the-art results in Ego4D's *Talk-to-me* challenge (+0.7% mAP, +3.2% Accuracy). For the *Look-at-me* challenge, it achieves competitive performance with the state-of-the-art (-0.7% mAP, +1.5% Accuracy) without supervised training on external data. On the EasyCom dataset, our method surpasses both supervised Active Speaker Detection approaches and state-of-the-art video encoders (+1.2% mAP, +1.9% Accuracy compared to MARLIN [10]).

Keywords: Egocentric video understanding · Self-supervised learning · Social interactions

1 Introduction

The integration of visual perception systems in interactive technologies [3, 8, 15, 20, 22] has opened new frontiers for enhancing human-machine interactions, reshaping the way machines engage with users. These perceptual capabilities allow robots to perceive and interpret the surrounding environment through

* This work was partly done during at internship at Amazon Lab126.

visual cues, enabling them to better understand, interact with, and assist humans in various tasks. One of the most exciting applications of this technology is the enhancement of social interactions between humans in augmented reality settings [18, 46] or between humans and machines [11], yet the challenges associated with the complex egocentric vision impedes the realization of its full potential.

Visual processing of social egocentric videos, consisting of continuous streams of dynamic perspective frames, present distinctive challenges for systems designed to interact seamlessly with humans. These challenges arise from the inherent variability in visual appearance caused by motion, occlusions, and lighting conditions. Overcoming these obstacles requires the development of specialized learning techniques capable of extracting valuable insights from complex data. To advance egocentric vision applications, researchers have developed datasets of first-person perspective videos in social interactions [18,23,46,52]. However, these datasets remain limited in size, diversity and quality (Table 1) due to the considerable cost associated with collecting and annotating egocentric data in the wild.

Table 1: Overview of different datasets relevant for egocentric social interactions. Tasks include Talk-to-me (TM), Look-at-me (LM), Turn-taking prediction (TP), Active Speaker Detection (AS), Auditory Attention Localization (AL) and Speaker identification (ID). Face Quality (Qua.) scores are the average confidence scores of a facial landmark detection model [9] across all frames. ‡ : approximate numbers. ? : cannot be determined either due to missing face bounding boxes. !: synthesized/not natural.

Datasets	Ego.	Task	#vids	#spk	Qua.
Ego4D [23]	✓	TM/LM	194	914‡	0.31
EgoCom [46]	✓	TP	176	34	?
EasyCom [18]	✓	TM/AS	12	72‡	0.63
SAAL [52]	✓	AL	120	50	?
Vox2 [44]	✗	ID	145K	6K	0.76
EgoVox	!	-	63K	2.4K	0.78

To this end, self-supervised learning has emerged as a promising strategy for mitigating the supervised learning problems requiring large labeled datasets. By enabling the extraction of generic representations from unlabeled data, self-supervised learning enables transferring knowledge from pre-text tasks with minimal constraint on the input data to downstream supervised tasks. Such methods have demonstrated impressive performance in understanding exocentric videos, consistently outperforming supervised learning methods across a wide spectrum of applications, such as action recognition [57, 59], object detection [2], tracking [62], video captioning [35], and video editing [5]. Despite significant advancements in self-supervised video representation learning, developing models tailored to egocentric videos, especially within social interaction tasks, remains an under-explored area. We attribute part of the challenge to the substantial viewpoint gap between exocentric and egocentric videos and the absence of large and diverse egocentric datasets for social interactions.

In this study, we approach self-supervised learning for egocentric social interaction videos from two key points. To alleviate the scarcity of egocentered data, we propose leveraging an advanced 3D face synthesis method to transform exocentric videos (captured from a single camera view) into videos featuring dynamic camera views, resembling egocentric images. Such an approach capital-

izes on the wealth of available exocentric conversational videos, facilitating the creation of diverse synthetic data to be used for augmentations with simulated egomotions. In this work, we utilize VoxCeleb2 [14], a diverse and unlabeled corpus featuring social interactions through thousands of celebrity interviews, as the foundation dataset for perspective augmentations, resulting in *Ego Vox*. Notably, the augmented videos exhibit superior face quality compared to the existing datasets tailored for egocentric social interactions that better suit self-supervised learning purposes. Regarding the model architecture, we propose *Ex2Eg-MAE* , a self-supervised learning framework that enhances VideoMAE [57] for egocentric video understanding in two fundamental ways: first, by reconstructing the original exocentric videos using masked videos from *Ego Vox*, and second, by learning to estimate feature representations through a teacher-student distillation process in which the teacher model VideoMAE [57] is provided with the original exocentric inputs. We additionally introduce Perspective Shift Estimation Modules (PSEM), an add-on module attached to each encoder layer of Ex2Eg-MAE during the self-supervised learning process, to help the model learn perspective-invariant features via an adversarial process. In particular, PSEMs estimate camera movements relative to a reference frame, while Ex2Eg-MAE aims to adversarially produce features with zero estimated shift. Ultimately, our goal is to produce a robust facial encoder generalized for egocentric videos that are useful for a wide range of egocentric social interaction tasks.

To evaluate the efficacy of the proposed framework, we conduct extensive experiments with the Ego4D [23] and EasyCom [18] datasets. Our experimental results show that Ex2Eg-MAE can capture generic facial features from egocentric visual inputs that transfer well on a wide range of tasks relating to social role understanding, including *Talk-to-me*, *Look-at-me* and *Active Speaker Detection*. These tasks, while differing significantly in nature, collectively demonstrate the broad applicability of our pre-trained encoder. They all aim to assign socio-conversational roles to individuals appearing in the camera frame (LAM: a person looking-at the camera wearer, TTM: a person talking to the camera wearer, and ASD: a person speaking), and are essential for the future domestic robots and AR technologies, where it is important to identify social references. In summary, our main contributions are three-fold.

- We introduce Ex2Eg-MAE , a novel framework for effective self-supervised learning to generate robust and generalizable facial features in egocentric videos.
- We propose a challenging auxiliary task for egocentric representation learning: reconstructing exocentric frontal facial videos from dynamically augmented inputs mimicking egocentric environments, and introduce a distillation process and an adversarial learning process with PSEM to further assist VideoMAE [57] with the task.
- We conduct a comprehensive evaluation across diverse datasets, encompassing a broad spectrum of tasks related to social role understanding, including Talk-to-me, Look-at-me, and Active Speaker Detection. Our results showcase the superiority of *Ex2Eg-MAE* compared to existing exocentric video encoders.

2 Related Work

Self-supervised Pre-training in Vision. Learning image representation from unlabeled data has gained significant interest within the field of computer vision. Broadly, current approaches fall into two categories. The first category is contrastive learning, which centers on maximizing the agreement between diverse augmentations of an image through the application of a contrastive loss [13,25]. The second category is generative learning, where the methodology involves the random masking of image patches, and the network is trained to reconstruct the original image [6,24]. Following the success of Masked Autoencoder (MAE) [24], Tong *et al.* [57] proposes VideoMAE for video pre-training by extending MAE with cube embeddings and extremely high tube masking ratio. Most relevant to our work is MARLIN [10], which adapts VideoMAE for web-crawled facial data to create a generic, universal and task-agnostic facial encoder. By adding face-region-guided masking and adversarial adaptation to VideoMAE, MARLIN achieves state-of-the-art performance on a wide range of *exocentric* facial-related downstream tasks. Recently, Radsosavovic *et al.* [49] explores masked autoencoder pre-training on egocentric images and demonstrates consistent performance boost on robotics downstream tasks. Pramanick *et al.* [47] propose a video-language pre-training framework and achieve state-of-the-art performance on several non-social egocentric tasks. To the best of our knowledge, there has been no prior work exploring SSL for egocentric facial videos.

Social Role Understanding. Social role understanding involves the discernment of speakers' roles in a given social interaction, encompassing aspects such as active speaker detection (ASD), determining the participants' gaze directions, and establishing communication dynamics between individuals. Among various social roles, ASD from *exocentric* videos is an extensively researched problem. From early methods that relied upon lip movements and facial gestures to serve as identifying cues [19], the landscape has witnessed a transformative shift towards deep learning based approaches, such as the integration of 3D Convolutional Neural Networks [36,56], the implementation of relational context modules [63], and the application of graph neural networks [4,42]. However, given the current surge in assistive technology and augmented reality/virtual reality (AR/VR), there is a growing inclination to delve into additional dimensions of social roles in *egocentric* contexts, including *look-at-me* (identifying who is directing their gaze toward the camera-wearer) and *talk-to-me* (determining who is engaged in conversation with the camera-wearer). In light of this evolving landscape, existing datasets focused on egocentric social understanding [18,23,46] typically come with at least one facet of social roles, as demonstrated in Table 1. Along this line, we introduce the first self-supervised learning model for egocentric facial videos, and validate our method across multiple facets of social roles.

Egocentric Video Understanding. Over the past few years, the majority of video understanding algorithms have been developed with well-defined third-person video datasets. Yet, the unique characteristics of egocentric video data, such as viewpoint changes, large motions, and visual distortions, have been relatively under-explored. To bridge this gap, several egocentric datasets [16,23,37,

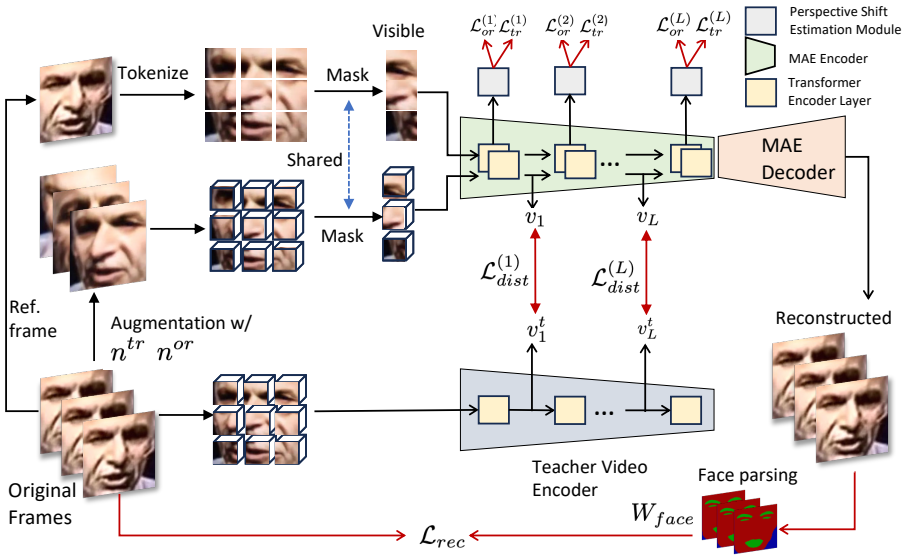

Fig. 1: An architectural overview of the Exo2Ego pre-training pipeline. Given a masked synthesized video (refer to Fig. 2 for our data augmentation pipeline) and reference frame, Ex2Eg-MAE learns to reconstruct the original exocentric video. To improve Ex2Eg-MAE 's capability to capture social signals, we add more weights to reconstruct the *mouth* and *eyes* regions via a face parser. We use a teacher encoder and an Perspective Shift Estimation Module (PSEM) to guide the generated features at each encoder layer. Specifically, at each layer, Ex2Eg-MAE learns to generate features that are (1) similar to the features generated by the teacher model, and (2) produce 0-estimated perspective shift via the layer's PSEM, while the PSEM learns to adversarially predict the correct rotation n_{or} and translation n_{tr} parameters.

53,55] have emerged, significantly advancing research in egocentric video understanding in areas such as human body pose estimation [29,38,45], activity recognition [32,64], anticipation [1,21] and human-object interaction [17,40,43]. Huang *et al.* [28] introduce an innovative approach for audio-visual object localization. At the core of their method is a geometry-aware temporal modeling module that transforms frames from various camera angles into a unified angle through homography. For egocentric social understanding, Jiang *et al.* [30] propose an end-to-end multi-channel audio-visual approach for active speaker localization. Ryan *et al.* [52] introduce an audio-visual network that utilizes both multi-channel audio and appearance-based features from egocentric videos for auditory attention localization. The mentioned methods are supervised approaches tailored to particular applications. In this work, we focus on self-supervised learning for egocentric facial videos with more universal applications.

3 Ex2Eg-MAE Framework

Figure 1 provides an overview of *Ex2Eg-MAE* pretraining process. Our framework begins with an input video featuring exocentric frames from both single

and frontal camera views. It then generates synthetic video sequences with dynamic perspectives with an auxiliary task involving reconstructing the original exocentric videos from these augmented sequences. We first discuss details of our data augmentation pipeline in Section 3.1. Then, we discuss the components of Ex2Eg-MAE and its self-supervised learning process in Section 3.2. In particular, our approach involves the utilization of a Perspective Shift Estimation Module (PSEM) at each layer of the encoder, which learns to estimate the translation and rotation parameters pertaining to the synthesized frames relative to a chosen reference frame (from the target frames). The PSEM-derived estimations are subsequently integrated into the transformer encoders to facilitate the generation of perspective-invariant representations. The representations generated at each layer of the encoder are also guided by the representations produced by a teacher model (VideoMAE [57]), which is derived from the original exocentric frames. Lastly, in Section 3.3, we discuss the optimization of *Ex2Eg-MAE* with its three loss terms: a reconstruction loss $\mathcal{L}_{rec}$ (Section 3.3a), a distillation loss $\mathcal{L}_{dist}$ (Section 3.3b), and a perspective estimation loss (Section 3.3c).

3.1 EgoVox for Data Augmentations

A good dataset for self-supervised egocentric representation learning requires four key attributes: diversity in speakers and video quantity, high-quality visible faces despite prevalent of occlusions and movements in egocentric videos, diverse points of view as egocentric visual inputs often contend with rapid perspective changes, and sufficient scale to support data-hungry learning with large models. Unfortunately, existing egocentric datasets for social tasks lack one or more of the desired properties. On the other hand, exocentric social data are abundant but lack egomotions (see Table 1). To fill this gap, we leverage large-scale exocentric conversational datasets [44] with advanced 3D face synthesis methods to create a dataset that can meet all of these criteria. We specifically use *Triplanenet* [7] as our multi-view face synthesis module based on the VoxCeleb2 [14] dataset. *Triplanenet* is a real-time inversion framework for EG3D [12] that provides high-quality face reconstruction with multi-view consistency by directly using tri-plane representations. The method produces high-quality images compared to recent GAN inversion methods [50, 51] while running an order of magnitude faster, making it suitable for large-scale data synthesis.

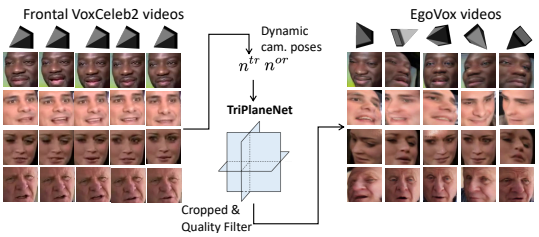

Fig. 2: An overview of the data augmentation pipeline. We select a diverse set of frontal videos from the VoxCeleb2 dataset [14], and generate dynamic view videos with *TriPlaneNet* [7].

Our data augmentation process is illustrated in Figure 2. For a given exocentric facial video, we start by defining a target sequence camera positions, gov-

erned by translation parameters $n_{tr} \in \mathcal{R}^{T\times 3}$ and rotation matrices $n_{ro} \in \mathcal{R}^{T\times 6}$ with respect to the original exocentric frames. The sequence of new camera positions are then fed into *Triplanenet* [7] to generate the desired synthetic facial frames with dynamic perspective. To verify the quality of the generated videos, we use a facial landmark detection model's confidence score to filter out low-quality videos [9]. We also crop the face in each frame to mitigate the impact of artifacts in the background during pre-training. The resulting EgoVox synthetic data contain around 63K videos with around 2.4M frames (around 135 hours) of 2.4K unique speakers. We provide samples of EgoVox in the supplementary materials.

3.2 Ex2Eg-MAE Training

Notations. We denote the input synthesized video as $v^{(in)} \in \mathcal{R}^{T\times h\times w\times 3}$, the target exocentric video as $v^{(tgt)} \in \mathcal{R}^{T\times h\times w\times 3}$, the reference frame as $r^{(in)} = v_i^{(tgt)} \in \mathcal{R}^{h\times w\times 3}$, the ground-truth egomotion translations as $n^{tr} \in \mathcal{R}^{T\times 3}$, and the ground-truth rotation matrices as $n^{or} \in \mathcal{R}^{T\times 3\times 3}$, where T is the sequence length and $i \in [T]$. The ground-truth translations and rotation matrices are the collected noises n_T and $R(n_R)$ during the creation of *EgoVox*. We denote pt, ph, pw, and ps as the number of patches along the temporal, height, width and spatial dimensions of the input. Hence, $ps = ph \times pw$ and $pt = \frac{T}{tub}$, where tub is the tubelet size.

We build *Ex2Eg-MAE* based on the model architecture and pre-training procedure of *VideoMAE* [57]. Given $v^{(in)}$ and $r^{(in)}$, our model first uses a cube embedding Φ_{emb} to generate two sequences of tokens, then supplement them with sinusoidal positional embeddings E_v^P and E_r^P along with learnable data type embeddings E_v and E_r. $p\%$ of the tokens are masked to create inputs for the transformer encoders $v_{mask}^{(in)}$ and $r_{mask}^{(in)}$. We pad $r^{(in)}$ with a zero frame to match the minimal temporal requirement of Φ_{emb}. Similar to VideoMAE [57], we employ tube masking and share the masks between the video frames with the reference frame to minimize information leakage.

$$\begin{aligned} v_{masked}^{(in)} &= Mask_p(\Phi_{emb}(v^{(in)}) + E_v + E_v^P) \\ r_{masked}^{(in)} &= Mask_p(\Phi_{emb}(r^{(in)}) + E_r + E_r^P) \end{aligned} \tag{1}$$

where $v_{masked}^{(in)} \in \mathcal{R}^{(pt\times ps)\times d}$ and $r_{masked}^{(in)} \in \mathcal{R}^{ps\times d}$ with d being the embedding dimension. The video and reference frame representations are updated for each encoder layer according to

$$v_i = Blk(v_{i-1})\ ;\ r_i = Blk(r_{i-1}) \tag{2}$$

where v_i is the video representation and r_i is the reference frame representation at layer i. Blk is a standard transformer encoder layer consisting of a self-attention module, a multi-layer linear perceptron (MLP) and a normalization layer. For the first layer, $v_0 = v_{masked}^{(in)}$ and $r_0 = r_{masked}^{(in)}$. For each layer, an PSEM receives v_i and r_i to estimate camera displacements between the video

and reference frame. We provide further details of PSEM in the following section. Similar to *VideoMAE*, we forward the output of the last layer of the encoder v_L to the decoder to reconstruct the original video $v^{(tgt)}$.

Perspective Shift Estimation Module . We add the Perspective Shift Estimation Module (PSEM) to each layer of the encoder. The goal of PSEM is to estimate the difference in camera orientation and rotation between $v^{(in)}$ and the reference frame $r^{(in)}$. PSEM takes as inputs the video and reference frame representations at each layer v_i and r_i, and outputs the translation and rotation estimation for each frame in the video $\tilde{n}^{tr}$ and $\tilde{n}^{or}$. Each PSEM is a light-weight cross-modal transformer encoder layer that consists of a Linear Layer, a cross-modal attention module, an MLP and a normalization layer, followed by two MLPs to predict $\tilde{n}^{tr}$ and $\tilde{n}^{ro}$ (as demonstrated in Fig 3). Given $v_i \in \mathcal{R}^{pt \times ps \times d}$ and $r_i \in \mathcal{R}^{ps \times d}$, the module first reduces the dimensionality of the inputs from d to d_r via a linear projection. A cross-attention layer translates information from the video to the reference frame information.

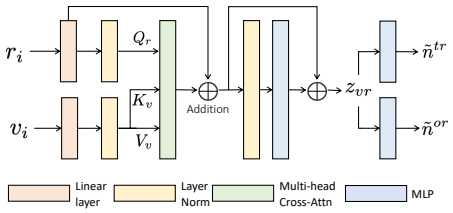

Fig. 3: An overview of the Perspective Shift Estimation Module.

$$z_{vr} = softmax(\frac{Q_r K_v^T}{\sqrt{d_k}})V_v \tag{3}$$

where $Q_r = r_i W_{Q_r}$, $K_v = v_i W_{K_v}$, and $V_v = v_i W_{V_v}$ with $W_{Q_r}, W_{K_v}, W_{V_v} \in \mathcal{R}^{d_r \times d_k}$. r_i is repeated along the temporal dimension to match the dimensionality of v_i. z_{vr} is then forwarded to a MLP and normalization layer (as in the standard transformer layer) to produce the output $z_{vr} \in \mathcal{R}^{pt \times ps \times d_r}$. We use two independent MLPs to estimate camera translation and rotation matrices between the video and reference frames.

$$\begin{aligned} \tilde{n}^{tr} &= MLP_{tr}(MeanPool(z_{vr}, dim = 1)) \\ \tilde{n}^{or} &= MLP_{or}(MeanPool(z_{vr}, dim = 1)) \end{aligned} \tag{4}$$

For camera translation, we estimate three independent parameters represent the x, y, z coordinates while for the rotation matrices, we estimate six independent parameters as in [27]. Therefore, MLP_{tr} and MLP_{or} map inputs of size $\mathcal{R}^{pt \times d_r}$ to outputs of size $\mathcal{R}^{pt \times (tub \times 3)}$ and $\mathcal{R}^{pt \times (tub \times 6)}$, respectively. We resize these outputs to make frame-level predictions of size $\mathcal{R}^{T \times 3}$ and $\mathcal{R}^{T \times 6}$. The 6D rotatation estimation output can then be mapped to rotation matrices as in [27].

3.3 Ex2Eg-MAE Training Losses

Figure 1 summarizes the training process for *Ex2Eg-MAE* . Our model is optimized with respect to (a) Reconstruction Loss, (b) Distillation Loss, and (c)

Perspective Shift Estimation Loss (adversarial).

(a) Reconstruction Loss. Given an input masked tokens of synthesized video $v^{(in)}_{masked}$, our masked auto-encoder learns to reconstruct the original real frames $x^{(tgt)}$. Unlike common masked auto-encoders that only learns to reconstruct the masked tokens [10,24,57], our model learns to reconstruct the whole frames because $v^{(in)} \neq v^{(tgt)}$. To achieve this goal, we use a mean squared error loss to minimize the distance in the 3D token space between $X_{rec} = Decoder(v_L)$ and x^{tgt}.

Furthermore, we leverage a face parsing model [58] to enhance our model's ability to capture social signals relating to social role understanding. Inspired by MARLIN [10], we use FaceXZoo [58] to classify face parts in $x^{(tgt)}$ at pixel-level, and double the weights for pixels identified as `mouth` or `eyes` while halved the weights for pixels identified as `background`. Therefore, the reconstruction loss is formulated as

$$\mathcal{L}_{rec} = W_{face} \odot ||X_{rec} - X^{tgt}||_2 \tag{5}$$

where W_{face} is the pixel-level face weights.

(b) Distillation Loss. We leverage a teacher model, VideoMAE [57], that learns from exocentric videos to guide our model's feature generation process. Specifically, we use the mean squared error loss to minimize the distance between the generated features by Ex2Eg-MAE and VideoMAE at each encoding layer.

$$\mathcal{L}_{dist} = \frac{1}{L}\sum_{i=1}^{L} ||v_i - v_i^t||_2 \tag{6}$$

where L is the number of layers in the video encoder and v_i^t is the features generated at the i-th layer of VideoMAE, using $x^{(tgt)}$ as the input.

(c) Perspective Shift Estimation Loss. The Perspective Estimation losses help PSEM learn to estimate translation and rotation differences between $v^{(in)}$ and $r^{(in)}$. In particular, we use the mean square error loss to estimate the target translations and use the Geodesic loss [27] to estimate the target rotation matrices for each frame of the input.

$$\mathcal{L}_{tr}(\tilde{y}^{tr}, y^{tr}) = \frac{1}{T}\sum_{t=1}^{T} ||\tilde{y}_t^{tr} - y_t^{tr}||_2 \tag{7}$$

$$\mathcal{L}_{or}(\tilde{y}^{or}, y^{or}) = \frac{1}{T}\sum_{t=1}^{T} cos^{-1}\left(\frac{tr(\tilde{y}_t^{or}(y_t^{or})^T) - 1}{2}\right) \tag{8}$$

where $\tilde{y}^{tr}, \tilde{y}^{or}$ are the predicted orientation parameters and y^{tr}, y^{or} are the target orientation parameters.

Since the goal of Ex2Eg-MAE is to produce egomotion-invariant representation, we adversarially train Ex2Eg-MAE with PSEM according to

$$\mathcal{L}_{PSEM} = \mathcal{L}_{tr}(\tilde{n}^{tr}, n^{tr}) + \mathcal{L}_{or}(\tilde{n}^{or}, n^{or}) \tag{9}$$

$$\mathcal{L}_{MAE} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{dist} + \lambda_3 (\mathcal{L}_{tr}(\tilde{n}^{tr}, \mathbf{0}^{tr}) + \mathcal{L}_{or}(\tilde{n}_t^{or}, \mathbf{0}^{or})) \quad (10)$$

where λ_1, λ_2, and λ_3 are hyper-parameters. During the pre-training phase, $\mathcal{L}_{MAE}$ updates the parameters in the masked autoencoder while $\mathcal{L}_{PSEM}$ updates the parameters of PSEMs.

4 Experiments

We have comprehensively compared our method across various downstream tasks relating to egocentric social understanding using existing datasets, including *Look-at-me*, *Talk-to-me*, and *Active Speaker Detection*. In this section, we begin by introducing the datasets used in our evaluations, followed by the baselines, and finally, we provide implementation details.

4.1 Datasets

Ego4D. The Ego4D dataset [23] comprises more than 3,000 hours of audio-visual data for egocentric video understanding. For social understanding tasks, it features the *Talking-to-me* (TTM) and *Looking-at-me* (LAM) challenges, which focus on detecting whether a person is talking or looking at the camera wearer from tracked faces in video segments. The dataset includes 389 training clips (32.4 hours), 50 validation clips (4.2 hours), and 133 test clips (11.1 hours, no ground-truth labels). Each clip is divided into segments with binary labels for tracked faces (talking or looking). We use the data filtering pipeline from Lin et al. [39] for the *Talk-to-me* challenge.
EasyCom. The EasyCom dataset [18] is a multimodal egocentric collection with 12 sessions spanning about 5.3 hours. In each session, 3 to 5 people sit around a table and engage in a natural conversation with a host for several tasks such as introductions, solving puzzles or playing games. The egocentric video recordings are obtained from the eyeglass cameras of all participants, excluding the host. The dataset offers annotated speech transcriptions for each conversation, including information about who said what and to whom. This enables us to extract Talk-to-me and Active Speaker Detection labels.

4.2 Baselines

We compare our model with state-of-the-art self-supervised video learners.
VideoMAE [57] is a state-of-the-art video encoder. The model was pre-trained on the Kinetics400 dataset [31] with more than 650 hours of video content of 400 human actions.
MARLIN [10] is a state-of-the-art facial video encoder. The model adversarially pre-trained VideoMAE on the YouTubeFace dataset [60] to achieve state-of-the-art performance on different facial video downstream tasks.
Eg4D-MARLIN/VideoMAE is the self-supervised adapted [34] version of MARLIN/VideoMAE on the social interaction subset of the egocentric Ego4D data. It is worth mentioning that we intentionally keep *EgoVox* to have a fairly

Table 2: Ego4D Talk-to-me results. *: state-of-the-art performance on leaderboard. ◇: challenge baseline.

Method	Modality	Acc-1	mAP
Majority Guess	-	53.6	54.2
Whisper* [48]	A	57.8	67.4
QuaVF [39]	V	54.8	56.2
VideoMAE [57]	V	56.8	57.8
MARLIN [10]	V	57.2	58.9
Eg4D-VideoMAE	V	56.1	56.4
Eg4D-MARLIN	V	57.4	58.3
Ex2Eg-MAE	V	**59.6**	**60.5**
ResNet-18 Bi-LSTM◇ [23]	A+V	54.3	53.9
EgoTask [61]	A+V	55.9	57.5
QuaVF [39]	A+V	57.1	65.8
VideoMAE w/ Whisper	A+V	58.7	67.2
MARLIN w/ Whisper	A+V	59.4	67.7
Ex2Eg-MAE w/ Whisper	A+V	**61.0**	**68.1**

similar size compared to the *Ego4D*-social interaction subset ($\sim$ 100hrs) to show that the benefit is mainly due tp the diversity within the data and not just the volume of the data.

Relevant state-of-the-art methods. For each task, we also include the performance of other recent supervised methods for reference. For the Ego4D challenges, we also include the best performance on the leaderboards. It is important to note that some of the SOTA methods use external annotated datasets for performance boost while our method do not use any. For example, SOTA in the LAM task uses AFLW2000 [65] for head-pose estimation.

4.3 Implementation Details

We follow most pre-training setup of MARLIN [10], including initializing our model with VideoMAE's weights [57]. 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 [41] 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 [23]. We provide more implementation details in the supplementary materials. 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.

5 Results and Discussion

Talk-to-me Results. In Tables 2 and 3, we compare the fine-tuning performance of Ex2Eg-MAE with popular transfer learning (i.e., VideoMAE [57],

Table 3: EasyCom Talk-to-me results.

Method	Pre-train	Acc-1	mAP
Majority Guess	-	80.4	52.4
VideoMAE [57]	K400 [31]	83.9	78.3
MARLIN [10]	YTF [60]	87.0	82.8
Eg4D-VideoMAE [57]	Ego4D [23]	86.3	80.1
Eg4D-MARLIN [10]	Ego4D [23]	87.6	84.2
Ex2Eg-MAE	EgoVox	**88.1**	**85.4**

MARLIN [10]) and other competitive supervised methods on the Ego4D TTM challenge and EasyCom dataset, respectively. From Tables 2 and 3, it is observed that Ex2Eg-MAE outperforms MARLIN by 1.6%/2.6% and VideoMAE by 2.7%/7.1% on the mAP metric, respectively. When combining with the Whisper model [48] for the audio modality via late fusion, we achieve state-of-the-art result on the Ego4D TTM challenge with an improvement of 0.7%. Notably, because the visual aspect of Ego4D TTM is prone to swift egomotions, previous work has shown diminishing performance when combining it with audio predictions [39]. Our model, however, demonstrates the capability of extracting meaningful information even in these challenging settings, leading to improved A-V performance (Ego4D: 67.4% → 68.1%).

Look-at-me Results. In Table 4, we report the performance of Ex2Eg-MAE for the Look-at-me task on the Ego4D dataset. In terms of the mAP metric, the results indicate that Ex2Eg-MAE outperforms VideoMAE (68.2% → 69.4%) and MARLIN (66.9% → 69.4). We further compare Ex2Eg-MAE with supervised methods on the dataset, namely ResNet18 Bi-LSTM, which uses a ResNet18 [26] model first supervised training on the large-scale gaze prediction Gaze360 dataset [33] then fine-tuning on Ego4D LAM, and PKU-WICT-MIPL, which is the state-of-the-art solution on the challenge's leaderboard. The results suggest that Ex2Eg-MAE performs competitively with other methods without any supervised pre-training, achieving nearly SOTA performance (mAP: 78.3% vs. 79.0%) while outperforming on the Acc-1 metric (Acc1: 93.5% vs. 92.0%). More importantly, we demonstrate the scalability of Ex2Eg-MAE with different encoder architectures. We observe that the larger model size results in better performance.

Active Speaker Detection Results. In Table 5, we compare Ex2Eg-MAE with the baselines for the ASD task on the EasyCom dataset. We also include performance for TalkNet [56] and SPELL [42], two supervised methods for ASD. The results show that Ex2Eg-MAE achieves substantial improvements over MARLIN (91.7% → 92.9%) and VideoMAE (90.0% → 92.9%). This further strengthens our claim that Ex2Eg-MAE is able to learn robust and transferable egocentric social signals.

Ablation Study. We progressively add different modules to Ex2Eg-MAE on the Ego4D-TTM and EasyCom-TTM task to investigate the contribution of individual modules (Table 6). We provide the ablation on the masking ratio in the supplementary materials, which indicates an optimal masking ratio of around

Table 4: Ego4D Look-at-me results. $^{\diamond}$: supervised pre-training. * : state-of-the-art performance on the leaderboard.

Method	Pre-train	Acc-1	mAP
Majority Guess	-	92.0	50.8
VideoMAE [57]	K400 [31]	91.8	68.2
MARLIN [10]	YTF [60]	90.8	66.9
Eg4D-VideoMAE	Ego4D [23]	88.6	61.2
Eg4D-MARLIN	Ego4D [23]	89.8	64.5
Resnet-18 Bi-LSTM$^{\diamond}$	Gaze360 [33]	86.0	71.9
PKU-WICT-MIPL*$^{\diamond}$	[33], [65]	92.0	**79.0**
Ex2Eg-MAE (tiny)	EgoVox	92.0	69.4
Ex2Eg-MAE (base)	EgoVox	92.7	73.9
Ex2Eg-MAE (large)	EgoVox	**93.5**	78.3

Table 5: EasyCom Active Speaker Detection results. ‡: supervised methods with supervised pre-trained encoders.

Method	Modality	Acc-1	mAP
Majority Guess	-	68.5	51.8
Whisper [48]	A	66.0	66.6
VideoMAE [57]	V	83.9	90.0
MARLIN [10]	V	86.3	91.7
Eg4D-VideoMAE [57]	V	86.7	91.4
Eg4D-MARLIN [10]	V	87.2	92.1
Ex2Eg-MAE	V	**88.2**	**92.9**
TalkNet‡ [56]	A+V	82.6	86.9
SPELL‡ [42]	A+V	84.5	89.5
Ex2Eg-MAE w/ Whisper	A+V	**88.7**	**93.2**

75%. A high masking ratio is necessary for the model to learn meaningful features and reduce the impact of artifacts introduced by the face synthesis process. However, a too-high masking ratio makes the reconstruction task and distillation process more difficult to learn, resulting in a degraded performance. It is worth noting that the masking ratio of 75% is lower than most masking ratios for video learners (e.g., VideoMAE with 95%, MARLIN with 90%). We attribute this to the more challenging nature of the reconstruction task and distillation process.

Qualitative Analysis. To better understand the effectiveness of the learned representations for egocentric inputs, we visualize the important regions that Ex2Eg-MAE focused on via *SmoothGrad* [54]. We provide the qualitative results on two tasks, namely EasyCom ASD and Ego4D LAM, and compared the saliency maps produced by Ex2Eg-MAE with the second best performing method for the corresponding task (additional qualitative results are also available in the supplementary materials). To better illustrate the effectiveness of the method in egocentric settings, we intentionally choose inputs with egomotions, as shown in Figure 4. Ex2Eg-MAE tends to provide more accurate and focused

Table 6: Contribution of different modules to Ex2Eg-MAE .

Modules	Ego4D-TTM		EasyCom-TTM	
	Acc-1	mAP	Acc-1	mAP
Ex2Eg-MAE				
$+\mathcal{L}_{rec}$	57.1	58.4	86.4	82.4
$+\mathcal{L}_{dist}$	58.8	59.5	86.8	84.0
$+\mathcal{L}_{ego}$	59.2	60.1	87.6	84.9
w/ W_{face}	**59.6**	**60.5**	**88.1**	**85.4**

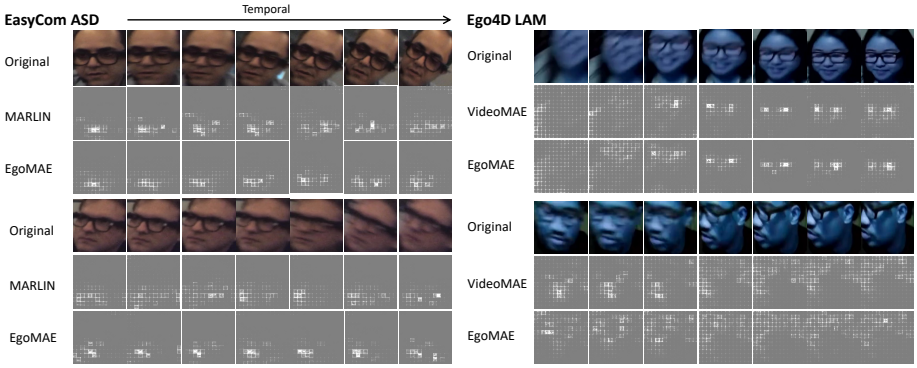

Fig. 4: Saliency map generated by *SmoothGrad* [54] of Ex2Eg-MAE and MARLIN/VideoMAE on EasyCom ASD and Ego4D LAM. We expect the saliency maps to focus on the mouth regions (for ASD) and eyes/head-pose regions. (for LAM).

attention to relevant regions (mouth for ASD and eyes/headposes for LAM) compared to MARLIN/VideoMAE, especially in the presence of egomotions (*e.g.,* [*ASD, sample 2, frame 6*] and [*LAM, sample 1, frame 2*]). The examples further demonstrate our method's ability to reduce the impact of egomotions.

6 Conclusion

In this paper, we present our novel *Exo2Ego* framework, adeptly adapting *exocentric* video encoders for *egocentric* contexts by leveraging novel-view face synthesis techniques for data augmentation, and enhance VideoMAE [57]'s representations for egocentric inputs with the task of reconstructing single-view videos from augmented ones with guidance from a teacher video encoder. Our experimental results on egocentric social role understanding tasks demonstrate the superior performance of our encoder compared to existing *exocentric* video encoders and supervised benchmarks. **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.

Acknowledgement The work was partially sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Abu Farha, Y., Richard, A., Gall, J.: When will you do what?-anticipating temporal occurrences of activities. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5343–5352 (2018) 5
2. Afouras, T., Asano, Y.M., Fagan, F., Vedaldi, A., Metze, F.: Self-supervised object detection from audio-visual correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10575–10586 (2022) 2
3. Alameda-Pineda, X., Sanchez-Riera, J., Wienke, J., Franc, V., Čech, J., Kulkarni, K., Deleforge, A., Horaud, R.: Ravel: An annotated corpus for training robots with audiovisual abilities. Journal on Multimodal User Interfaces **7**, 79–91 (2013) 1
4. Alcázar, J.L., Caba, F., Thabet, A.K., Ghanem, B.: Maas: Multi-modal assignation for active speaker detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 265–274 (2021) 4
5. Athar, S., Shu, Z., Samaras, D.: Self-supervised deformation modeling for facial expression editing. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 294–301. IEEE (2020) 2
6. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: International Conference on Learning Representations (2021) 4
7. Bhattarai, A.R., Nießner, M., Sevastopolsky, A.: Triplanenet: An encoder for eg3d inversion. arXiv preprint arXiv:2303.13497 (2023) 6, 7
8. Boissy, P., Corriveau, H., Michaud, F., Labonté, D., Royer, M.P.: A qualitative study of in-home robotic telepresence for home care of community-living elderly subjects. Journal of telemedicine and telecare **13**(2), 79–84 (2007) 1
9. 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, 7
10. 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) 1, 4, 9, 10, 11, 12, 13
11. Carros, F., Meurer, J., Löffler, D., Unbehaun, D., Matthies, S., Koch, I., Wieching, R., Randall, D., Hassenzahl, M., Wulf, V.: Exploring human-robot interaction with the elderly: results from a ten-week case study in a care home. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2020) 2
12. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022) 6

13. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 4
14. Chung, J., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. Interspeech 2018 (2018) 3, 6
15. Cruz, F., Parisi, G.I., Twiefel, J., Wermter, S.: Multi-modal integration of dynamic audiovisual patterns for an interactive reinforcement learning scenario. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 759–766. IEEE (2016) 1
16. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018) 4
17. Damen, D., Leelasawassuk, T., Mayol-Cuevas, W.: You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. Computer Vision and Image Understanding **149**, 98–112 (2016) 5
18. Donley, J., Tourbabin, V., Lee, J.S., Broyles, M., Jiang, H., Shen, J., Pantic, M., Ithapu, V.K., Mehra, R.: Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. arXiv preprint arXiv:2107.04174 (2021) 2, 3, 4, 10
19. Everingham, M., Sivic, J., Zisserman, A.: Hello! my name is... buffy"–automatic naming of characters in tv video. In: BMVC. vol. 2, p. 6 (2006) 4
20. Fritsch, J., Kleinehagenbrock, M., Lang, S., Fink, G.A., Sagerer, G.: Audiovisual person tracking with a mobile robot. In: Proc. Int. Conf. on Intelligent Autonomous Systems. pp. 898–906 (2004) 1
21. Furnari, A., Farinella, G.M.: Rolling-unrolling lstms for action anticipation from first-person video. IEEE transactions on pattern analysis and machine intelligence **43**(11), 4021–4036 (2020) 5
22. Gan, C., Zhang, Y., Wu, J., Gong, B., Tenenbaum, J.B.: Look, listen, and act: Towards audio-visual embodied navigation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9701–9707. IEEE (2020) 1
23. 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) 2, 3, 4, 10, 11, 12, 13
24. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 4, 9
25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 4
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 12
27. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: 6d rotation representation for unconstrained head pose estimation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2496–2500. IEEE (2022) 8, 9
28. Huang, C., Tian, Y., Kumar, A., Xu, C.: Egocentric audio-visual object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22910–22921 (2023) 5

29. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from egocentric video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3501–3509. IEEE (2017) 5
30. Jiang, H., Murdock, C., Ithapu, V.K.: Egocentric deep multi-channel audio-visual active speaker localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10544–10552 (2022) 5
31. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 10, 12, 13
32. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5492–5501 (2019) 5
33. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6912–6921 (2019) 12, 13
34. Kim, D., Wang, K., Sclaroff, S., Saenko, K.: A broad study of pre-training for domain generalization and adaptation. In: European Conference on Computer Vision. pp. 621–638. Springer (2022) 10
35. Kim, S., Jeong, S., Kim, E., Kang, I., Kwak, N.: Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 13171–13179 (2021) 2
36. Köpüklü, O., Taseska, M., Rigoll, G.: How to design a three-stage architecture for audio-visual active speaker detection in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1193–1203 (2021) 4
37. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 1346–1353. IEEE (2012) 4
38. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023) 5
39. 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) 10, 11, 12
40. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 704–721. Springer (2020) 5
41. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018) 11
42. Min, K., Roy, S., Tripathi, S., Guha, T., Majumdar, S.: Learning long-term spatial-temporal graphs for active speaker detection. In: European Conference on Computer Vision. pp. 371–387. Springer (2022) 4, 12, 13
43. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8688–8697 (2019) 5
44. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: Voxceleb: Large-scale speaker verification in the wild. Computer Speech & Language **60**, 101027 (2020) 2, 6
45. Ng, E., Xiang, D., Joo, H., Grauman, K.: You2me: Inferring body pose in egocentric video via first and second person interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9890–9900 (2020) 5

46. Northcutt, C., Zha, S., Lovegrove, S., Newcombe, R.: Egocom: A multi-person multi-modal egocentric communications dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 2, 4
47. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5285–5297 (2023) 4
48. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023) 11, 12, 13
49. Radosavovic, I., Xiao, T., James, S., Abbeel, P., Malik, J., Darrell, T.: Real-world robot learning with masked visual pre-training. In: Conference on Robot Learning. pp. 416–426. PMLR (2023) 4
50. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2287–2296 (2021) 6
51. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on graphics (TOG) **42**(1), 1–13 (2022) 6
52. Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., Ithapu, V.K.: Egocentric auditory attention localization in conversations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14663–14674 (2023) 2, 5
53. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626 (2018) 4
54. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017) 13, 14
55. Su, Y.C., Grauman, K.: Detecting engagement in egocentric video. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 454–471. Springer (2016) 4
56. Tao, R., Pan, Z., Das, R.K., Qian, X., Shou, M.Z., Li, H.: Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3927–3935 (2021) 4, 12, 13
57. 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) 1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14
58. Wang, J., Liu, Y., Hu, Y., Shi, H., Mei, T.: Facex-zoo: A pytorch toolbox for face recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3779–3782 (2021) 9
59. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14549–14560 (2023) 2
60. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: CVPR 2011. pp. 529–534. IEEE (2011) 10, 12, 13

61. Xue, Z., Song, Y., Grauman, K., Torresani, L.: Egocentric video task translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2310–2320 (2023) 11
62. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7177–7188 (2021) 2
63. Zhang, Y., Liang, S., Yang, S., Liu, X., Wu, Z., Shan, S., Chen, X.: Unicon: Unified context network for robust active speaker detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3964–3972 (2021) 4
64. Zhou, Y., Berg, T.L.: Temporal perception and prediction in ego-centric video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4498–4506 (2015) 5
65. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016) 11, 13