Spherical World-Locking for Audio-Visual Localization in Egocentric Videos

Heeseung Yun[1,2]*, Ruohan Gao[2], Ishwarya Ananthabhotla[2], Anurag Kumar[2], Jacob Donley[2], Chao Li[2], Gunhee Kim[1], Vamsi Krishna Ithapu[2], and Calvin Murdock[2]

[1]Seoul National University, [2]Reality Labs Research at Meta
`https://hs-yn.github.io/SWL/`

Abstract. Egocentric videos provide comprehensive contexts for user and scene understanding, spanning multisensory perception to behavioral interaction. We propose Spherical World-Locking (SWL) as a general framework for egocentric scene representation, which implicitly transforms multisensory streams with respect to measurements of head orientation. Compared to conventional head-locked egocentric representations with a 2D planar field-of-view, SWL effectively offsets challenges posed by self-motion, allowing for improved spatial synchronization between input modalities. Using a set of multisensory embeddings on a world-locked sphere, we design a unified encoder-decoder transformer architecture that preserves the spherical structure of the scene representation, without requiring expensive projections between image and world coordinate systems. We evaluate the effectiveness of the proposed framework on multiple benchmark tasks for egocentric video understanding, including audio-visual active speaker localization, auditory spherical source localization, and behavior anticipation in everyday activities.

Keywords: Egocentric Vision · Audio-Visual Learning

1 Introduction

Egocentric videos provide comprehensive context from an individual's perspective, serving an essential role in user and scene understanding. Compared to conventional exocentric videos, egocentric videos capture in-the-wild context from a human-centric viewpoint covering daily routine activities and social interactions like conversation. Therefore, it is paramount to capture the interplay between visual, auditory, and behavioral modalities for comprehensive reasoning tasks in line with how humans would perceive their surroundings [6, 53]. Accordingly, a significant amount of research has been dedicated to exploring the integration of multiple modalities in egocentric videos [12,22,34,38,45,58,63]. In this work, we develop a general framework for multisensory egocentric perception and apply it to audio-visual and behavioral localization problems in egocentric videos.

* Work done during an internship at Meta.

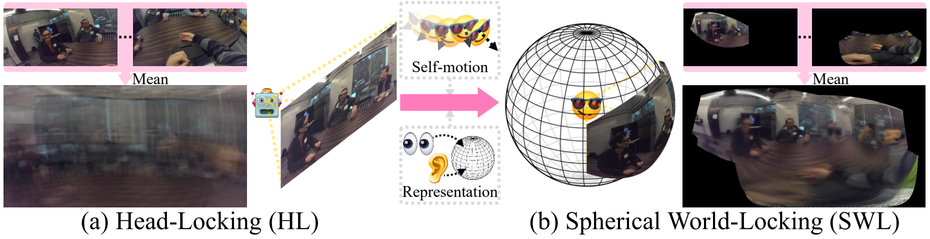


Fig. 1: The key idea of our framework. (a) In conventional *Head-Locked* (HL) frameworks, multisensory observations captured from head-mounted devices are used as-is, where self-motion introduces variability in otherwise static scenes. (b) Our Spherical World-Locking (SWL) framework compensates for self-motion with negligible overhead, leading to lower variability and better learnable scene representation.

One of the most distinctive characteristics of egocentric videos is self-motion. This is best illustrated in a group conversation where people frequently move their heads to engage in various actions such as nodding, making eye contact, or visually exploring their surroundings. That is, self-motion poses an important challenge for egocentric video understanding, as the relative location of stimuli in a head-locked reference frame would also move accordingly. Other factors stemming from self-motion, like motion drift and a limited field-of-view, also contribute to the increased complexity. Therefore, self-motion is often treated as a challenge in egocentric video understanding [5, 33, 42, 46, 58].

Nevertheless, self-motion is one of the core components of egocentric scene understanding, acting as a strong proxy of behavior and its underlying intention. For instance, the internal representations of our perceived surroundings do not change significantly with respect to drastic self-motion and always remain gravity-aligned thanks to behavioral responses like the vestibulo-ocular [70] and proprioceptive reflexes [20]. We can also effortlessly coordinate head motion to proactively sharpen our perception of attended contexts [7, 65]. In other words, humans are efficient stabilizers as well as utilizers of self-motion, and we claim that these traits can be beneficially adapted for egocentric video understanding.

To this end, we introduce **S**pherical **W**orld-**L**ocking (SWL) as a novel framework for integrating self-motion into egocentric videos. As depicted in Fig. 1, in contrast to conventional *Head-Locked* frameworks that learn to offset variability from self-motion in raw input streams, SWL forms a virtual sphere around a person and efficiently transforms audio-visual streams based on measurements of their relative head orientation. This is generally applicable in egocentric videos by leveraging sensors like inertial measurement units in commercial head-mounted devices [12,13,22,45]. SWL inherently offsets challenges posed by self-motion such as drift and visibility, while maintaining the strengths of a head-locked representation like compatibility with conventional frameworks.

We propose the **Mu**ltisensory **S**pherical World-Locked **T**ransformer (MuST), a novel architecture that incorporates self-motion-aware multisensory inputs for representation learning. MuST leverages self-motion as a useful cue for learn-

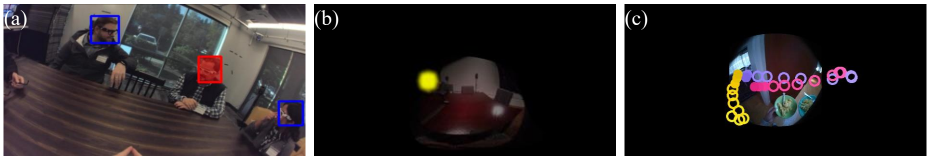


Fig. 2: Three multisensory localization tasks in egocentric videos that we tackle in this work: (a) audio-visual active speaker localization (§5.1), (b) auditory spherical source localization (§5.2), and (c) egocentric behavior anticipation (§5.3).

ing with negligible computational overhead via spherical position embeddings. To further facilitate the interaction of multisensory inputs on the world-locked sphere, we introduce modality-wise self-attention and quaternion-based spatial similarity. In addition, multiple classification tokens improve localized predictions and allow for flexible decoding on multiple target domains, *e.g.*, field-of-view, spherical, and pointwise prediction.

We validate the effectiveness of our framework with three representative benchmarks for multisensory egocentric perception and understanding. First, we outperform prior arts by a significant margin in the audio-visual active speaker localization benchmark on the EasyCom dataset [13]. Second, we obtain superior auditory spherical source localization performance on the RLR-CHAT dataset [50, 76]. Finally, to demonstrate the generalizability of our framework, we extend it to egocentric behavior anticipation in everyday activities on the Aria Everyday Activities (AEA) dataset [45], where we establish a new benchmark of jointly predicting a set of cohesive behaviors from multisensory contexts.

2 Related Works

Egocentric Video Understanding. A variety of reasoning tasks have been studied from the wearer-centric perspective based on the wearer's spatiotemporal or multimodal context [12, 22, 23]. Some prior works focus on improving egocentric human-object interaction by predicting the wearer's hand motion [43] or hand segmentation [30] and discerning distracting objects [71]. In addition, information from third-person videos can be transferred to the egocentric domain by means of knowledge distillation [40] or temporal alignment without paired data [75]. Other works learn topological maps with longer temporal dependencies [51] or pre-training with embodied agents [52]. Also, fine-grained temporal relationships among multiple modalities are modeled [34, 35].

Egocentric videos with additional modalities measuring the wearer's behavior offer unique challenges in understanding the context. Pose estimation is one of the primary tasks of egocentric user understanding, making use of dynamic motion signatures [31], body segmentation with motion history [32], or an intersection between kinematics and dynamics [44]. Gaze has been estimated by learning the correlation between global context and local information of visual tokens [38]. A more recent line of research utilizes the wearer's pose information, *i.e.*, IMU sensors, for efficient action recognition [63] or translation to textual description via contrastive pre-training [48]. While previous works make use of *unimodal* models for IMU and train them with cross-modal learning objectives, we directly incorporate the wearer's pose with audio-visual embeddings for *multimodal* models preserving spherical world-locked structure.

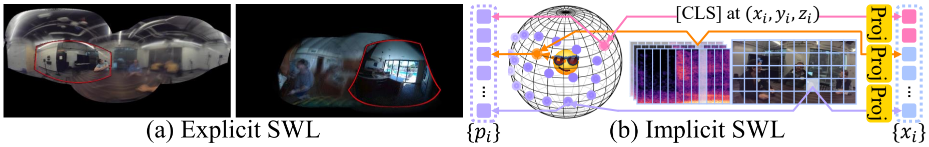
Audio-Visual Localization. Extensive research has been conducted to exploit audio-visual correspondence for localization in videos in-the-wild with self-supervision [3,18,54,61], cross-modal clustering [25,47], parameter-efficient adaptation [41], and pixel-level correspondence learning [72, 80, 81], to list a few. Audio-visual speaker detection is another major task in identifying the coherence between the two modalities under multi-speaker scenarios [1, 36, 57, 64, 68]. Other recent works on audio-visual localization exploit a richer set of modalities to enhance localization capabilities, like question-answer grounding [79], language-guide separation [62], cross-view consistency of source directions [9], homography [26], and audio-visual saliency [73].

Conversations in egocentric videos are often more complex than in conventional videos due to noisy environments and unconstrained multi-speaker interactions. Jiang et al. [33] combine audio-only and audio-visual networks to perform spherical and inner field-of-view active speaker localization. Ryan et al. [58] refine this localization capability by only detecting the attended conversation partner, while Jia et al. [29] further propose to predict the complete ego-exo conversational graph from egocentric video. Closest to our work is [50], where self-motion behaviors are used as a self-supervised learning objective. Unlike SWL, the self-supervisory signal cannot solely offset the challenges of self-motion.

Spherical Scene Representation. There has been a surge of interest in representing spherical data, from modeling 360° videos [39, 60] to the Earth's climate [49]. Omnidirectional videos are generally projected into multiple normal field-of-view images to mitigate distortion and discontinuity [39, 60]. On the other hand, some prior works propose invariant or equivariant architectures on a sphere [11, 16, 27], discretization with polyhedral approximation [15], or data structures like the spherical binoctree [28] and balanced spherical grid [10] for more faithful representations of the spherical scene. Other works focus on the transferability of convolutional networks [59] or transformers [77] from the normal image domain to the 360° domain. However, egocentric videos are planar and not spherical by nature, making it less practical to adopt their spherical architectures for learning. Instead, our method interprets egocentric videos on a world-locked sphere while preserving their original format, without requiring additional expensive mechanisms like the Spatial Transformer [27] to incorporate the spherical nature of egocentric observations.

3 Spherical World-Locking

In conventional head-locked frameworks, egocentric videos are provided as-is, and the model must learn the complex nature of self-motion from end to end. In contrast, we propose Spherical World-Locking (SWL) to represent videos on



(a) Explicit SWL (b) Implicit SWL

Fig. 3: Comparison of explicit and implicit spherical world-locking. While explicit SWL maps the original inputs to the spherical reference frame, implicit SWL retains the original inputs to process position ($\{p_i\}$) and semantic information ($\{x_i\}$) separately.

a world-locked sphere around the wearer's head, serving as an effective means to model self-motion in egocentric videos. Since these audio-visual data are not spherical in nature, we first need to establish the connection between the world-locked sphere and audio-visual egocentric streams. We use multisensory egocentric inputs comprising a video frame $\mathcal{V}$ $(3 \times H_v \times W_v)$, the corresponding multichannel audio spectrogram $\mathcal{A}$ $(C_a \times H_a \times W_a)$, and behaviors $\mathcal{B}$ as 3D unit vectors $(N_b \times 3)$ if available. We consider three different directional behaviors of the wearer, *i.e.*, eye gaze, head orientation, and motion trajectory.

As visualized in Fig. 3, we formulate two different SWL methods that can equivalently represent the world-locked sphere, where each has its distinct advantages. *Explicit* spherical world-locking (§3.1) maps the original video to a 360° panorama, *i.e.*, $f_{EX} : \mathcal{V} \mapsto \mathcal{V}'$, where $\mathcal{V}'$ $(3 \times H_p \times W_p)$ is a panoramic image. Whereas in *implicit* spherical world-locking (§3.2), we keep the original inputs and construct the mapping from patches (for audio-visual inputs) or vectors (for other inputs) to tuples of semantic and position embeddings that encode the corresponding self-motion.

3.1 Explicit Spherical World-Locking

We first outline the procedure of placing egocentric videos on a world-locked sphere. Among the three multisensory inputs, behaviors with direction $\mathcal{B}$ can be trivially placed on a sphere with scalar multiplication. However, assigning a precise direction to audio inputs $\mathcal{A}$ is difficult. In fact, this is often the ground truth the model aims to predict. Instead, we pair each multichannel audio segment with the readily available head pose information, *e.g.*, IMUs. Since the wearer's pose determines the microphone array's orientation, the model can directly consider self-motion instead of capturing subtle signals about self-motion in audio during training. These pairs can further be utilized to explicitly synthesize the spatial audio locked to a specific direction if necessary [2, 55].

Spherical Projection. Egocentric field-of-view videos can be visually projected onto a sphere. This is analogous to a partially observable 360° panorama where the observable region changes with respect to self-motion. An equiangular mapping from the ij-th pixel in $\mathcal{V}$ to the XY-th pixel in $\mathcal{V}'$ can be computed as follows, given horizontal and vertical angular fields-of-view of θ_{HF} and θ_{VF}, where $q \in \mathbb{R}^4$ is the current head pose provided in quaternions, $g_{ij} \in \mathbb{R}^4$ is the spherical world-locked position of ij-th pixel in pure quaternions, and R is the radius of a world-locked sphere such that $(H_P, W_p) = (\pi R, 2\pi R)$:

$$g'_{ij} = (0, \tan(\theta_{\mathrm{HF}} \times (j/W_v - 0.5)), \tan(\theta_{\mathrm{VF}} \times (i/H_v - 0.5)), 1), \tag{1}$$

$$g_{ij} = (0, g^x_{ij}, g^y_{ij}, g^z_{ij}) = q(g'_{ij}/\left\|g'_{ij}\right\|_2)q^{-1}, \tag{2}$$

$$(X, Y) = (R \times \mathrm{atan2}(g^z_{ij}, \sqrt{(g^x_{ij})^2 + (g^y_{ij})^2}), R \times (\mathrm{atan2}(g^y_{ij}, g^x_{ij}) + \pi)). \tag{3}$$

Although complete pose information with rotation and translation could be combined to further improve the spatial consistency in Fig. 3-(a), we only consider rotation in light of three observations: (i) it is more efficient and straightforward to integrate in our model due to the unit quaternion assumption (Eq. (7)), (ii) rotation suffices for common seated activities like conversations, and (iii) the influence of translation becomes negligible within a certain length of time, where we use less than one second in all experiments.

3.2 Implicit Spherical World-Locking

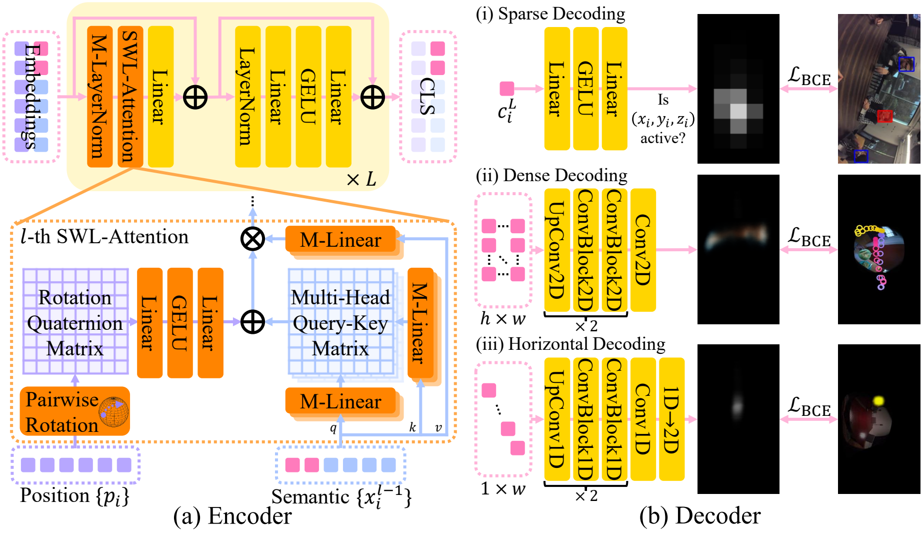
While explicit spherical world-locking can be compelling in terms of interaction and visualization, it is less practical to use this representation for model training. For example, an irregular array access in Eq. (3) incurs nontrivial overhead. Distorted images as in Fig. 3-(a) also introduce another challenge of distortion-aware methods discussed in §2. To circumvent these issues, we suggest an implicit way to construct a spherical world-locked representation of multisensory inputs. As depicted in Fig. 3-(b), we leave all inputs intact and pair them with coordinates on a world-locked sphere to maintain semantic and position embeddings of multisensory inputs separately.

Multi-CLS Embeddings. Classification tokens are commonly employed for capturing the global context from a set of input embeddings. Since our goal is to localize signals spatially, we exploit multiple classification tokens $\{c_i\}_{i=1}^{N_c}$ parametrized with a point $p_i = (x_i, y_i, z_i)$ on a world-locked sphere to capture semantic information around p_i, where $\mathbf{W}_c \in \mathbb{R}^{d\times 3}$, $\mathbf{b}_c \in \mathbb{R}^d$ are learnable weights:

$$c_i = \mathbf{W}_c p_i + \mathbf{b}_c. \tag{4}$$

Some recent works use multiple CLS tokens for capturing class-specific information in semantic segmentation [74] or ensembling in language understanding [8]. The key difference in our work is that we deploy multiple CLS tokens to predict spatially localized signals on a sphere for flexible decoding (§4.2).

Semantic Input Embeddings. Since implicit spherical world-locking naturally separates position embeddings on a world-locked sphere from semantic embeddings, we can use off-the-shelf feature encoders for processing unmodified multisensory inputs. To obtain visual embeddings $\{v_i\}_{i=1}^{N_v}$, we use ResNet-18 [24] or 3-layer ConvNet to process frames or facial images. For audio embeddings $\{a_i\}_{i=1}^{N_a}$, we apply linear projection per patch as in [14,21,78], where we use vertical patches of the spectrogram, as shown in Fig. 3-(b), for the audio semantic embeddings to accurately align with the wearer's head pose. Finally, similar to



(a) Encoder
(b) Decoder

Fig. 4: Our MuST model architecture. M- indicates modality-wise operations.

Eq. (4), we assign a learnable embedding for behavioral input $\{b_i\}_{i=1}^{N_b}$, which is also parametrized with a point on a unit sphere.

Position Input Embeddings. For each semantic input embedding, we assign a 3D point on a sphere so that all of the multisensory input embeddings are implicitly located on a world-locked sphere, as illustrated in Fig. 3-(b). For classification tokens and behavioral inputs, we use the coordinates identical to the ones used in semantic inputs. For audio and visual inputs, we assign the head orientation q in Eq. (2) and spherical world-locked location of each visual token, respectively. In short, our multisensory input embeddings are summarized as

$$\{x_i\}_{i=1}^{N} = \{x_i^0\}_{i=1}^{N} = \{c_1, ..., c_{N_c}, a_1, ..., a_{N_a}, v_1, ..., v_{N_v}, b_1, ..., b_{N_b}\}, \tag{5}$$

$$\{p_i\}_{i=1}^{N} = \{(x_1, y_1, z_1), ..., (x_N, y_N, z_N)\}. \tag{6}$$

4 Multisensory Spherical World-Locked Transformer

We propose the Multisensory Spherical World-Locked Transformer (MuST) to perform audio-visual localization tasks in egocentric videos, building upon the concept of implicit spherical world-locking and multimodal transformers. Using multisensory input embeddings in Eq. (5–6), we exploit MuST encoder blocks that focus on multisensory interactions on a world-locked sphere (§4.1), followed by a lightweight decoder for tackling various localization tasks flexibly (§4.2).

4.1 MuST Encoder

Transformers [69] can effectively integrate multisensory embeddings and their corresponding positions, forming an ideal combination of position and semantic embeddings from implicit spherical world-locking. Our MuST block is built upon the general Transformer block [14], with two key differences in self-attention to incorporate multiple senses on a world-locked sphere: spatial similarity and modality-wise operations.

Spatial Similarity on Sphere. Since each multisensory embedding retains a position on a world-locked sphere, we can model the pairwise interaction between different modalities in the form of rotation quaternions. This spatial similarity matrix at l-th layer $\mathbf{P}^l \in \mathbb{R}^{N \times N}$ is integrated into each (semantic) query-key matrix in the multi-head self-attention, promoting spatial relations among embeddings. The rotation from a 3D unit vector p_i to p_j is computed as

$$\mathbf{P}^l_{ij} = \text{Linear}(\text{GELU}(\text{Linear}([1 + p_i \cdot p_j,\, p_i \times p_j]))), \tag{7}$$

where $(1 + p_i \cdot p_j, p_i \times p_j)$ is a rotation quaternion and the output of MLP is scalar, *i.e.*, $\mathbb{R}^4 \rightarrow \mathbb{R}$. Since the pairwise rotation remains identical for all layers, rotation quaternions are computed once for each input and used for all layers.

Modality-wise Operations. Since our goal is to encode a heterogeneous set of modalities in a single encoder, it is paramount to harmonize them during training. We promote cross-modal interactions in each encoder block by applying layer normalization [4] and q, k, v projection in multi-head attention in a modality-specific manner, *i.e.*, M-LN and M-Attn in Table 1-(b). By normalizing the embeddings with modality-wise means and variances while retaining the modality-specific mapping before dot-product attention in each layer, *i.e.* using different linear projection per modality, the model notably promotes an interplay among different modalities than its unimodal counterparts. Note that not all modality-specific modules positively influence the model training, which is further discussed in §5.1. In short, multi-head self-attention for each head is

$$\bar{x}^l_i = \text{Linear}(0.5 \times (\sigma(Q^l K^{lT}/\sqrt{d}) + \sigma(\mathbf{P}^l))V^l), \tag{8}$$

$$x^{l+1}_i = x^l_i + \bar{x}^l_i + \text{Linear}(\text{GELU}(\text{Linear}(\bar{x}^l_i))), \tag{9}$$

where σ denotes softmax and Q^l, K^l, V^l are queries, keys, and values projected with modality-wise linear layers (*i.e.*, M-Linear in Fig. 4), respectively.

4.2 MuST Decoder

Using multiple CLS tokens obtained from the last encoder layer, *i.e.*, $\{c^L_i\}^{N_c}_{i=1}$, we can employ different decoding strategies depending on the target task.

Sparse Decoding. For each token c^L_i, we apply pointwise decoding with an MLP to obtain score y_i that corresponds to our model's prediction on point p_i:

$$y_i = \text{Linear}(\text{GELU}(\text{Linear}(c^L_i))). \tag{10}$$

Unless mentioned otherwise, we use a sparse grid of 5×10 CLS tokens for training the model, *i.e.*, each token covers around 2% of the output region. It is possible to make the training more efficient by selecting a subset of classification tokens for the model like a set of pre-detected regions of interest or faces (*e.g.*, MuST with Sparse Point in Table 1-(d)). This reduces the number of CLS tokens by an order of magnitude smaller, which is far more efficient than the full grid.

Dense Decoding. From a grid of CLS tokens, we can utilize a light deconvolutional network illustrated in Fig. 4-(b)-(ii) to obtain a dense output map $\mathbf{y}$ of the desired resolution in either a field of view $(H_v \times W_v)$ or a spherical panorama $(H_p \times W_p)$. One special case of dense decoding is horizontal decoding (Fig. 4-(b)-(iii)), which is applicable to spherical source localization. Since the spherical world-locking compensates self-motion in the data stream, our scene representation is gravity-aligned. So, for some tasks, it is possible to discard tokens except for the ones around the equator line without notable performance degradation. For example, there is only a marginal performance gap in our model reported in Table 2, *i.e.*, around 1°, in spite of reducing the number of CLS tokens by a factor of five. In this case, 1D operations instead of 2D can be used for decoding and converted to 2D by permuting channel dimension to vertical dimension, making both encoding and decoding more efficient.

Learning Objective and Details. We adapt the binary cross entropy loss for training the model, *i.e.*, $\frac{1}{N_c}\sum \mathcal{H}(y_i, \hat{y}_i)$ for sparse decoding and $\frac{1}{H \times W}\sum \mathcal{H}(\mathbf{y}, \hat{\mathbf{y}})$ for dense decoding where $\mathcal{H}$ denotes cross entropy. We use the Adam optimizer [37] with a learning rate of 1e-4 without scheduling. The model is trained end-to-end for 10 epochs until convergence, where we closely follow the hyperparameters used in a smaller variant of Vision Transformer (DeIT-S [67]), which has slightly fewer parameters than ResNet-50 [24].

5 Experiments

For a comprehensive demonstration of the effectiveness of our framework, we evaluate MuST with multiple benchmarks covering diverse egocentric videos. We first focus on egocentric audio-visual active speaker localization (§5.1). In addition, we report the performance of auditory spherical source localization (§5.2) to evaluate the model's capability of localizing directional signals on a sphere without visual shortcuts. Finally, we generalize our framework to more diverse everyday activities by developing a new suite of tasks on egocentric behavior anticipation (§5.3), which jointly predicts the direction of wearer's future behaviors, *i.e.*, gaze, head orientation, and trajectory, from multisensory inputs.

5.1 Audio-Visual Active Speaker Localization

Dataset. EasyCom [13] is a public dataset of egocentric conversations for augmented reality applications. It consists of 0.38M video frames and their corresponding sensory inputs like pose and multichannel audio. Due to the highly noisy nature of audio streams from the microphone array and frequent self-motion, the dataset covers various challenges in egocentric multi-speaker conversations such as speech enhancement. We focus on active speaker localization as one of the most common egocentric audio-visual localization tasks.

Table 1: Performance of active speaker localization on the EasyCom dataset [13].

(a) Methods	mAP$_\uparrow$
DOA [66]	52.62
DPT [56]	61.66
MRC [33]	64.24
BAVNet [72]	60.75
TalkNet [64]	69.13
BPN$_{\text{face}}$ [50]	75.22
AVLN [50]	85.11
MAVASL$_{\text{Spec}}$ [33]	85.49
MAVASL$_{\text{C+E}}$ [33]	86.32
MuST	**89.88**
Oracle	91.03

(b) Encoder	mAP$_\uparrow$
MuST$_{\text{w/o pose}}$	87.76
MuST$_{\text{w/o rotation}}$	88.83
MuST$_{\text{w/o M-ops}}$	88.53
MuST$_{\text{M-LN}}$	89.67
MuST$_{\text{M-LN,M-MLP}}$	89.16
MuST$_{\text{M-LN,M-Attn}}$	**89.88**

(c) Modality	mAP$_\uparrow$
$\mathcal{B}_{\text{pose}}$	47.95
$+\mathcal{A}_{\text{mono}}$	68.57
$+\mathcal{V}$	68.78
$+\mathcal{A}_{\text{mono}}+\mathcal{V}$	73.47
$+\mathcal{A}_{\text{multi}}$	89.50
$+\mathcal{A}_{\text{multi}}+\mathcal{V}$	**89.88**

(d) Decoder	mAP$_\uparrow$
Sparse Point	88.95
Dense	89.58
Sparse Grid	**89.88**

(e) Temporal	mAP$_\uparrow$
100ms	82.96
200ms	87.78
300ms	**89.88**

Experiment Settings. We closely follow prior works' experiment settings [33] like splits and metrics for a fair comparison. We mainly report the mean average precision (mAP), which captures both spatial and temporal localization of speech activity inside the camera's field-of-view. The mAP scores of all models are computed by pooling the maximum logit value within the corresponding head bounding boxes. We also devise an Oracle comparison as a potential upper bound on the egocentric models' performance using close microphone recordings from the other participants, which are unavailable from the wearer's perspective, to detect speech activity from cleaner near-field audio.

We compare our full model and its variants with a number of competitive baselines. We report the performance of a state-of-the-art audio signal processing method [66], visually oriented methods like mouth region classifier [33] and Dense Prediction Transformer [56], and competitive localization frameworks based on LSTM [72] or Transformer [64]. We also report more recent frameworks on egocentric audio-visual localization [33, 50] for thorough comparison. Finally, ablation studies on encoder-decoder design, input modalities, and temporal window are provided for a comprehensive analysis of MuST.

Comparison with Prior Arts. In Table 1-(a), our proposed framework outperforms previous methods by a large margin, increasing the accuracy by 3.6%p. This gap becomes wider if the same modalities (faces and spectrograms) are used as inputs, *i.e.*, 4.4%p. The performance of our model is comparable with the Oracle's despite the usage of noisy microphone arrays (-1.2%p). This tight upper bound is likely due to the coarse-grained nature of annotations in EasyCom, *i.e.*, active speech labels are based on phrases instead of phonemes, whereas the model's prediction is evaluated every 50ms. Such a gap makes it hard for the model to differentiate pauses of active speakers from non-speakers, which is in line with performance degradation for shorter temporal windows in Table 1-(e).

Ablation Studies. Table 1-(b) shows the influence of different encoder components on the performance. Components based on spherical world-locking, like position information and rotation, significantly contribute to the full model's per-

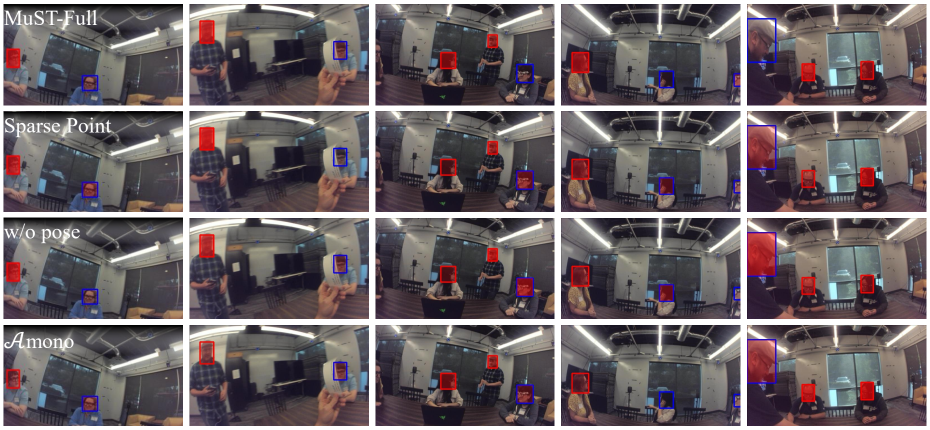


Fig. 5: Qualitative examples of egocentric active speaker localization on EasyCom [13]. The red/blue boxes indicate active/non-active speakers, and the red heatmap indicates model prediction. MuST can make correct predictions for scenes with gravity misalignment (col. 1), motion blur (col. 2, 4), and multi-speakers (col. 3, 5).

formance, (+2.1%p). Also, MuST without modality-specific operations (MuST$_{\text{w/o M-ops}}$) displays even worse performance than the model without visual information in Table 1-(c). Since not all modality-specific operations are beneficial to performance, it is essential to reconcile different modalities properly. Combining mono audio with visual input introduces a remarkable +4.9 mAP gain, even outperforming a large-scale pretrained prior art for single-channel localization [64]. Still, the largest performance boost comes from the multichannel microphone array, which can capture rich spatial signals with multiple microphones.

Qualitative Results. Fig. 5 compares our full model with selected variants of MuST, demonstrating correct active speaker localization in challenging scenarios with multiple active speakers and diverse self-motion. The sparse point decoder, which assigns a CLS token to each detected face, is also precise except for the last column with variable head size. Without pose information or multichannel audio, performance falls short due to insufficient spatial reasoning capability.

5.2 Auditory Spherical Source Localization

Dataset. RLR-CHAT [76] is a large-scale dataset of egocentric multisensory streams under a variety of configurations, covering an order of magnitude more recordings than in EasyCom [13]. RLR-CHAT encompasses more realistic scenarios like scene layouts, overlapping speech in free-form conversations, and varying degrees of background noise. Due to increased diversity and lower frame rates (*i.e.*, 200ms), it is paramount for the model to perform precise localization as well as disambiguate multiple speakers at the same time. We particularly focus on a more challenging setup of multichannel auditory source localization that lacks visual cues, which could prevent shortcuts like face regions.

Table 2: Comparison of auditory spherical source localization errors on the RLR-CHAT dataset [76].

	$\text{MAE}_{\text{g}\rightarrow\text{p}\downarrow}$	$\text{MAE}_{\text{p}\rightarrow\text{g}\downarrow}$
EchoNet [17]	66.99	65.29
MAVASL-A [33]	62.25	60.70
SAM [78]	46.95	44.90
$\text{MuST}_{\text{w/o pose}}$	29.33	28.55
MuST	**15.23**	**12.67**

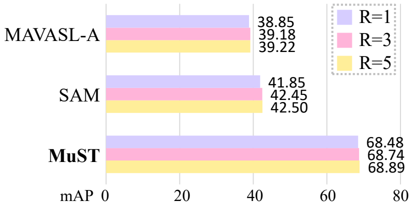


Fig. 6: Spherical mAP with varying angular precision on RLR-CHAT [76].

Experiment Settings. Following prior works [33], we report the mean angular error (MAE) from prediction to ground truth and vice versa, reflecting how far the model's prediction deviates from the ground truth source direction on a sphere. In order to consider the differences among models' output distributions, we select a fixed number of peaks in predictions with non-max suppression, *i.e.*, the number of active speakers in a 200ms timeframe.

We report the performance of several state-of-the-art models on spatial reasoning with audio using identical audio features (multichannel STFT) and ground truth for a fair comparison. We use the audio network of MAVASL [33], EchoNet [17] for spatial reasoning with echolocation, and the SAM audio network [78] for dense indoor prediction with sound. We also provide spherical mAP with varying angular resolutions to get a better grasp of spatial precision. Please refer to [45] for details regarding the microphone array configuration used in experiments.

Performance Analysis. Table 2 summarizes the performance of different auditory spherical source localization methods. Our framework achieves superior performance in both MAE metrics with or without pose, and the usage of pose information substantially improves the audio-based localization performance. Fig. 6 illustrates the mAP scores on a sphere with varying angular precision, where radius values of 1, 3, and 5 correspond to detection thresholds of 2.25°, 6.75°, and 11.25°, respectively. Despite generally lower performance due to higher complexity, angular precision variations do not have a notable influence on mAP scores, implying the precise localization ability of correct predictions.

Fig. 7 depicts qualitative examples of audio-only localization performance, where all models efficiently bypass a visually challenging scenario of a hair occlusion in the first column. In addition, our framework displays better source detection and localization capability than the competitive baseline [78].

5.3 Egocentric Behavior Anticipation

Dataset. We extend MuST to perform multisensory localization in more diverse egocentric daily activities to examine the generalizability of our proposed framework. The Aria Everyday Activities (AEA) Dataset [45] covers diverse egocentric videos of daily activities like cooking or chatting for scene comprehension, comprising 143 recordings from five different environments. Understanding and anticipating the wearer's behaviors, *i.e.*, eye gaze, head orientation, and motion

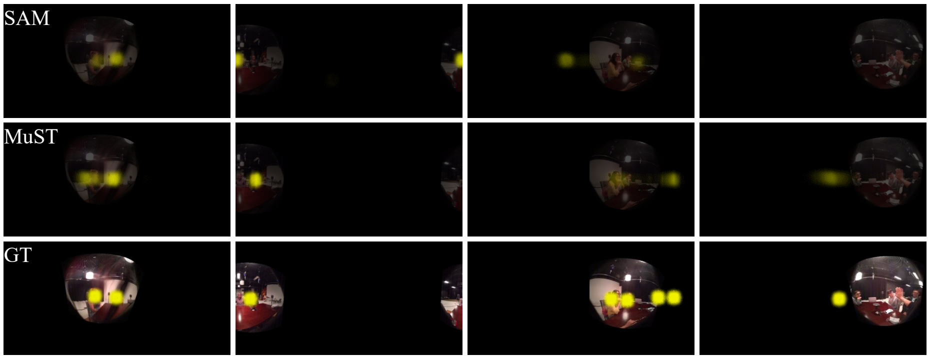


Fig. 7: Qualitative examples of auditory spherical source localization on RLR-CHAT [76]. Our model displays precise detection as well as localization capability over the prior art [78]. Note that visual frames are not used in all models.

Table 3: Comparison of behavior anticipation errors on the AEA Dataset [45].

	Gaze			Orientation			Trajectory		
MAE$_{\downarrow}$	$T_{300\text{ms}}$	$T_{500\text{ms}}$	$T_{700\text{ms}}$	$T_{300\text{ms}}$	$T_{500\text{ms}}$	$T_{700\text{ms}}$	$T_{300\text{ms}}$	$T_{500\text{ms}}$	$T_{700\text{ms}}$
MultitaskGP [19]	11.42	15.59	18.40	**4.70**	9.28	12.27	13.75	17.86	20.02
MuST$_{\mathcal{AV}}$	12.26	14.48	16.59	5.68	8.28	10.75	92.38	92.46	92.71
MuST$_{\mathcal{B}}$	**8.78**	11.98	14.65	5.02	7.65	10.18	**9.77**	**12.04**	**13.48**
MuST$_{\mathcal{VB}}$	8.92	**11.96**	**14.57**	4.82	7.40	9.91	10.05	12.36	13.90
MuST$_{\mathcal{AVB}}$	9.17	12.15	14.75	4.78	**7.36**	**9.90**	9.96	12.38	13.95
MuST$_{\mathcal{AVB}\text{-singletask}}$	9.19	12.35	15.02	5.02	7.74	10.28	10.08	12.53	14.03

trajectory, in daily activities can be crucial in egocentric user understanding and applicable for assistive systems in augmented or mixed reality scenarios. Since there is no public benchmark that jointly anticipates a set of cohesive egocentric behaviors with multisensory observations to the best of our knowledge, we organize a suite of tasks for holistic egocentric behavior anticipation with AEA.

Experiment Setting. We tackle three egocentric behavioral targets on the world-locked sphere: eye gaze, head orientation, and motion trajectory. Considering the typical behavioral reaction time of humans, our goal is to anticipate future behaviors in 300/500/700ms given the current audio-visual observations and previous behavioral contexts of 700ms. To evaluate localization performance, we use Mean Angular Errors (MAE) of behaviors at different timestamps by comparing the argmax coordinate of the model's prediction with ground truth behaviors, similar to §5.2. We report the average performance from a five-fold cross-validation using five different scenes in the dataset. As the problem of egocentric behavior anticipation in this dataset has not been addressed previously, we report the performance of a competitive baseline of Multitask Gaussian Process [19] as well as selected variants of our model for an extensive analysis.

Performance Analysis. As visualized in Fig. 8, egocentric behavior is quite complex and often challenging to predict. Still, our model achieves consistent

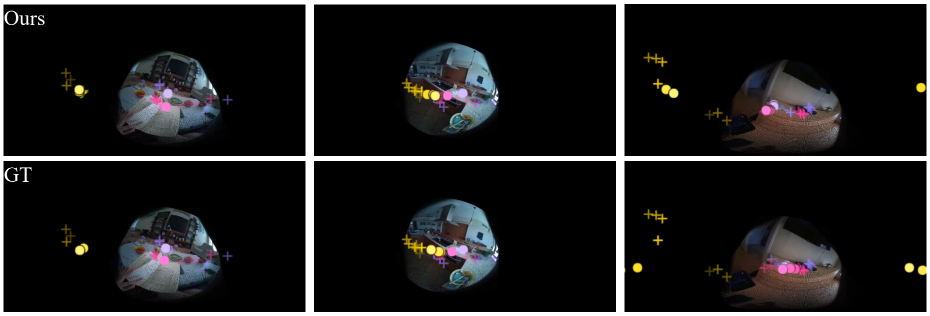


Fig. 8: Qualitative examples of egocentric behavior anticipation on the AEA Dataset [45] where gaze, orientation, and trajectory are color coded. Cross/circle symbols denote previous/anticipated behaviors. Our model reasonably anticipates future behaviors in common scenarios like long-term fixation and human-object interaction.

performance improvements over the baseline except for short-term head orientation anticipation, reducing the angular error by 20.5%, 12.8%, and 29.5% for gaze, orientation, and trajectory respectively. It is noteworthy that our audio-visual model without previous gaze context (MuST$_{\mathcal{AV}}$) displays compelling performance in gaze anticipation task but is poor at predicting future head trajectories. Such tendency suggests that, unlike exocentric behaviors, egocentric behaviors like gaze can be anticipated from current audio-visual observations to a meaningful extent without previous behavioral context. Different sets of input modalities are often more proficient for a specific task than others. For example, audio inputs are closely tied with previous pose information, achieving better performance in anticipating orientation than others. Lastly, our model trained to jointly anticipate all behaviors outperforms single-task counterparts in all tasks, meaning that MuST is properly leveraging coherence across different behavioral contexts.

6 Conclusion

We presented the Spherical World-Locking, a new framework for audio-visual localization in egocentric videos that leverages the wearer's pose information to offset challenges in self-motion. Powered by implicit SWL, our MuST architecture facilitates cross-modal interaction on a world-locked sphere by means of rotation quaternions and modality-wise operations, enabling learning better multisensory scene representation. It also provides fine-grained and flexible decoding for localization with multiple spatial classification tokens. We have conducted extensive experiments on three different multisensory egocentric localization benchmarks. Our results demonstrate significant improvement both quantitatively and qualitatively over prior arts. As future work, we plan to extend our framework to exploit other modalities like optical flow as a proxy of pose information, which is not always available, and scale to more large-scale egocentric video datasets.

References

1. Afouras, T., Owens, A., Chung, J.S., Zisserman, A.: Self-supervised learning of audio-visual objects from video. In: ECCV (2020)
2. Ahonen, J., Kallinger, M., Küch, F., Pulkki, V., Schultz-Amling, R.: Directional analysis of sound field with linear microphone array and applications in sound reproduction. In: Audio Engineering Society Convention (2008)
3. Arandjelovic, R., Zisserman, A.: Objects that sound. In: ECCV (2018)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016)
5. Bansal, S., Arora, C., Jawahar, C.: My view is the best view: Procedure learning from egocentric videos. In: ECCV (2022)
6. Bottini, G., Karnath, H.O., Vallar, G., Sterzi, R., Frith, C.D., Frackowiak, R.S., Paulesu, E.: Cerebral representations for egocentric space: functional–anatomical evidence from caloric vestibular stimulation and neck vibration. Brain (2001)
7. Brimijoin, W.O., Boyd, A.W., Akeroyd, M.A.: The contribution of head movement to the externalization and internalization of sounds. PloS one (2013)
8. Chang, H.S., Sun, R.Y., Ricci, K., McCallum, A.: Multi-CLS BERT: An efficient alternative to traditional ensembling. In: ACL (2023)
9. Chen, Z., Qian, S., Owens, A.: Sound localization from motion: Jointly learning sound direction and camera rotation. In: ICCV (2023)
10. Choi, C., Kim, S.M., Kim, Y.M.: Balanced spherical grid for egocentric view synthesis. In: CVPR (2023)
11. Cohen, T.S., Geiger, M., Köhler, J., Welling, M.: Spherical CNNs. In: ICLR (2018)
12. 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: ECCV (2018)
13. 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:2107.04174 (2021)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
15. Eder, M., Shvets, M., Lim, J., Frahm, J.M.: Tangent images for mitigating spherical distortion. In: CVPR (2020)
16. Esteves, C., Allen-Blanchette, C., Makadia, A., Daniilidis, K.: Learning SO(3) equivariant representations with spherical CNNs. In: ECCV (2018)
17. Gao, R., Chen, C., Al-Halah, Z., Schissler, C., Grauman, K.: VisualEchoes: Spatial image representation learning through echolocation. In: ECCV (2020)
18. Gao, R., Grauman, K.: 2.5 d visual sound. In: CVPR (2019)
19. Gardner, J., Pleiss, G., Weinberger, K.Q., Bindel, D., Wilson, A.G.: GPyTorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. NeurIPS (2018)
20. Gauthier, G.M., Vercher, J.L., Blouin, J.: Egocentric visual target position and velocity coding: role of ocular muscle proprioception. Annals of biomedical engineering (1995)
21. Gong, Y., Chung, Y.A., Glass, J.: AST: Audio spectrogram transformer. In: InterSpeech (2021)
22. 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: CVPR (2022)

23. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives. arXiv:2311.18259 (2023)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
25. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: CVPR (2019)
26. Huang, C., Tian, Y., Kumar, A., Xu, C.: Egocentric audio-visual object localization. In: CVPR (2023)
27. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. NIPS (2015)
28. Jang, H., Meuleman, A., Kang, D., Kim, D., Richardt, C., Kim, M.H.: Egocentric scene reconstruction from an omnidirectional video. Transactions on Graphics (2022)
29. Jia, W., Liu, M., Jiang, H., Ananthabhotla, I., Rehg, J.M., Ithapu, V.K., Gao, R.: The audio-visual conversational graph: From an egocentric-exocentric perspective. In: CVPR (2024)
30. Jia, W., Liu, M., Rehg, J.M.: Generative adversarial network for future hand segmentation from egocentric video. In: ECCV (2022)
31. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3D body pose from egocentric video. In: CVPR (2017)
32. Jiang, H., Ithapu, V.K.: Egocentric pose estimation from human vision span. In: ICCV (2021)
33. Jiang, H., Murdock, C., Ithapu, V.K.: Egocentric deep multi-channel audio-visual active speaker localization. In: CVPR (2022)
34. Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., Damen, D.: With a little help from my temporal context: Multimodal egocentric action recognition. In: BMVC (2021)
35. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-Fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019)
36. Kim, Y.J., Heo, H.S., Choe, S., Chung, S.W., Kwon, Y., Lee, B.J., Kwon, Y., Chung, J.S.: Look who's talking: Active speaker detection in the wild. In: InterSpeech (2021)
37. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014)
38. Lai, B., Liu, M., Ryan, F., Rehg, J.M.: In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. IJCV (2023)
39. Lee, S., Sung, J., Yu, Y., Kim, G.: A memory network approach for story-based temporal summarization of 360 videos. In: CVPR (2018)
40. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: Ego-Exo: Transferring visual representations from third-person to first-person videos. In: CVPR (2021)
41. Lin, Y.B., Sung, Y.L., Lei, J., Bansal, M., Bertasius, G.: Vision transformers are parameter-efficient audio-visual learners. In: CVPR (2023)
42. Liu, M., Ma, L., Somasundaram, K., Li, Y., Grauman, K., Rehg, J.M., Li, C.: Egocentric activity recognition and localization on a 3D map. In: ECCV (2022)
43. 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: ECCV (2020)
44. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. NeurIPS (2021)

45. Lv, Z., Charron, N., Moulon, P., Gamino, A., Peng, C., Sweeney, C., Miller, E., Tang, H., Meissner, J., Dong, J., Somasundaram, K., Pesqueira, L., Schwesinger, M., Parkhi, O., Gu, Q., Nardi, R.D., Cheng, S., Saarinen, S., Baiyya, V., Zou, Y., Newcombe, R., Engel, J.J., Pan, X., Ren, C.: Aria everyday activities dataset. arXiv:2402.13349 (2024)
46. Mai, J., Hamdi, A., Giancola, S., Zhao, C., Ghanem, B.: EgoLoc: Revisiting 3D object localization from egocentric videos with visual queries. In: ICCV (2023)
47. Mo, S., Tian, Y.: Audio-visual grouping network for sound localization from mixtures. In: CVPR (2023)
48. Moon, S., Madotto, A., Lin, Z., Dirafzoon, A., Saraf, A., Bearman, A., Damavandi, B.: IMU2CLIP: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. In: EMNLP Findings (2023)
49. Mudigonda, M., Kim, S., Mahesh, A., Kahou, S., Kashinath, K., Williams, D., Michalski, V., O'Brien, T., Prabhat, M.: Segmenting and tracking extreme climate events using neural networks. In: Deep Learning for Physical Sciences (DLPS) Workshop, held with NIPS Conference (2017)
50. Murdock, C., Ananthabhotle, I., Lu, H., Ithapu, V.K.: Self-motion as supervision for egocentric audiovisual localization. In: ICASSP (2024)
51. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: EGO-TOPO: Environment affordances from egocentric video. In: CVPR (2020)
52. Nagarajan, T., Ramakrishnan, S.K., Desai, R., Hillis, J., Grauman, K.: EgoEnv: Human-centric environment representations from egocentric video. NeurIPS (2023)
53. Ohmi, M.: Egocentric perception through interaction among many sensory systems. Cognitive Brain Research (1996)
54. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)
55. Pulkki, V.: Directional audio coding in spatial sound reproduction and stereo upmixing. In: Audio Engineering Society Conference (2006)
56. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
57. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., et al.: AVA active speaker: An audio-visual dataset for active speaker detection. In: ICASSP (2020)
58. Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., Ithapu, V.K.: Egocentric auditory attention localization in conversations. In: CVPR (2023)
59. Su, Y.C., Grauman, K.: Learning spherical convolution for 360° recognition. IEEE TPAMI (2021)
60. Su, Y.C., Jayaraman, D., Grauman, K.: Pano2Vid: Automatic cinematography for watching 360 videos. In: ACCV (2016)
61. Sun, W., Zhang, J., Wang, J., Liu, Z., Zhong, Y., Feng, T., Guo, Y., Zhang, Y., Barnes, N.: Learning audio-visual source localization via false negative aware contrastive learning. In: CVPR (2023)
62. Tan, R., Ray, A., Burns, A., Plummer, B.A., Salamon, J., Nieto, O., Russell, B., Saenko, K.: Language-guided audio-visual source separation via trimodal consistency. In: CVPR (2023)
63. Tan, S., Nagarajan, T., Grauman, K.: EgoDistill: Egocentric head motion distillation for efficient video understanding. NeurIPS (2023)
64. 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: ACM MM (2021)

65. Thurlow, W.R., Mangels, J.W., Runge, P.S.: Head movements during sound localization. The Journal of the Acoustical society of America (1967)
66. Tourbabin, V., Donley, J., Rafaely, B., Mehra, R.: Direction of arrival estimation in highly reverberant environments using soft time-frequency mask. In: WASPAA (2019)
67. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
68. Truong, T.D., Duong, C.N., Pham, H.A., Raj, B., Le, N., Luu, K., et al.: The right to talk: An audio-visual transformer approach. In: ICCV (2021)
69. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
70. Wallach, H.: The role of head movements and vestibular and visual cues in sound localization. Journal of Experimental Psychology (1940)
71. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: ICCV (2021)
72. Wu, X., Wu, Z., Ju, L., Wang, S.: Binaural audio-visual localization. In: AAAI (2021)
73. Xiong, J., Wang, G., Zhang, P., Huang, W., Zha, Y., Zhai, G.: CASP-Net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective. In: CVPR (2023)
74. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: CVPR (2022)
75. Xue, Z.S., Grauman, K.: Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. NeurIPS (2023)
76. Yin, Y., Ananthabhotla, I., Ithapu, V.K., Petridis, S., Wu, Y.H., Miller, C.: Hearing loss detection from facial expressions in one-on-one conversations. In: ICASSP (2024)
77. Yun, H., Lee, S., Kim, G.: Panoramic vision transformer for saliency detection in 360° videos. In: ECCV (2022)
78. Yun, H., Na, J., Kim, G.: Dense 2D-3D indoor prediction with sound via aligned cross-modal distillation. In: ICCV (2023)
79. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-AVQA: Grounded audio-visual question answering on 360deg videos. In: ICCV (2021)
80. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: ICCV (2019)
81. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: ECCV (2018)