[Appendix] Spherical World-Locking for Audio-Visual Localization in Egocentric Videos

A More Implementation Details

In this section, we provide a detailed explanation of two essential components for reproducing experiments reported in §5: implicit spherical world-locking for multisensory inputs and configurations used for each benchmark. Algorithm 1 describes the pseudocode for computing each position and semantic embedding of k-th multisensory input at time t, which are equivalent to p_k and x_k in Fig. 3–4. Table 4 reports an exhaustive list of hyperparameters involved in training MuST for each benchmark.

Algorithm 1 Implicit spherical world-locking

$\tilde{x}_k \leftarrow k$-th multisensory input at time t
if isExocentric($\tilde{x}_k$) **then** `// Exocentric inputs (e.g., motion trajectory)`
  $p_k \leftarrow \tilde{x}_k / \left\|\tilde{x}_k\right\|_2$
  $x_k \leftarrow \text{Linear}(p_k)$
else
  $q \leftarrow \text{readPose}(t)$ `// A unit quaternion`
  if isAudio($\tilde{x}_k$) **then** `// Audio patches`
    $(g^0, g^1, g^2, g^3) \leftarrow q(0, 0, 0, 1)q^{-1}$
    $x_k \leftarrow \text{Linear}(\tilde{x}_k)$
  else if isVisual($\tilde{x}_k$) **then** `// Visual patches`
    $i, j \leftarrow \text{columnIdx}(\tilde{x}_k), \text{rowIdx}(\tilde{x}_k)$
    $\bar{g}_k \leftarrow (\tan(\theta_{\text{HF}} \times (j/W_v - 0.5)), \tan(\theta_{\text{VF}} \times (i/H_v - 0.5)), 1)$
    $\bar{g}_k \leftarrow \bar{g}_k / \left\|\bar{g}_k\right\|_2$
    $(g^0, g^1, g^2, g^3) \leftarrow q(0, \bar{g}_k^0, \bar{g}_k^1, \bar{g}_k^2)q^{-1}$
    $x_k \leftarrow \text{ResNet18}(\tilde{x}_k)$
  else `// Egocentric behaviors with direction`
    $\bar{g}_k \leftarrow \tilde{x}_k^t / \left\|\tilde{x}_k^t\right\|_2$
    $(g^0, g^1, g^2, g^3) \leftarrow q(0, \bar{g}_k^0, \bar{g}_k^1, \bar{g}_k^2)q^{-1}$
    $x_k \leftarrow \text{Linear}((g^1, g^2, g^3))$
  end if
  $p_k \leftarrow (g^1, g^2, g^3)$ `// A 3D point on a unit sphere`
end if
return p_k, x_k

Table 4: List of hyperparameters and configurations used in all experiments.

	EasyCom (§5.1)	RLR-CHAT (§5.2)	Aria Pilot (§5.3)
Epochs	10	10	10
Batch size	16	16	16
Optimizer	Adam	Adam	Adam
Learning rate	0.0001	0.0001	0.0001
Learning rate decay	None	None	None
Weight decay	0	0	0
Encoder depth	12	12	12
Embedding dimension	384	384	384
Self-attention heads	6	6	6
Dropout	0.1	0.1	0.1
CLS embeddings	5×10	1×20	5×10
Audio length	300ms	200ms	700ms
STFT window length, n_fft	199	199	199
STFT hop length	24	24	96
Audio patch size	100×8	100×8	100×8
Video resolution	160×320	-	224×224

B Additional Qualitative Examples

We provide more qualitative examples of our model's prediction in different benchmarks as well as the baselines'. In Fig. 9 concerning audio-visual active-speaker localization, our full model can perform accurate localization in common scenarios (row 1, 3) as well as more tricky cases like zero active speakers (row 2) and field-of-view occlusion (row 4). Fig. 10 visualizes audio-based spherical source localization, where we can observe a similar tendency in detection and localization capability among different models. Lastly, in egocentric behavior anticipation, the prediction from our model in Fig. 11 is fairly accurate for behaviors with fixated or relatively simple paths, but the prediction could become less accurate when more complex interactions are involved.

C Discussion

C.1 Verifying Annotations with Spherical World-Locking

One of the most straightforward applications of spherical world-locking is to validate different streams of modalities for seated conversations. For example, as illustrated in Fig. 12-(a), if the pose information is not correctly synchronized (*e.g.*, lagged by 100ms), we can intuitively visualize the deviation of participants in a scene or quantify it with respect to temporal sliding windows for a correct alignment. Another example is the bounding box annotations in field-of-view images. Although more than 99% of the bounding boxes are correctly assigned with the rigorous annotation pipeline in EasyCom, a couple of failure cases can be detected automatically with our world-locked sphere. Fig. 12-(b) visualizes some of the representative failure cases in the test split, where drastic motion

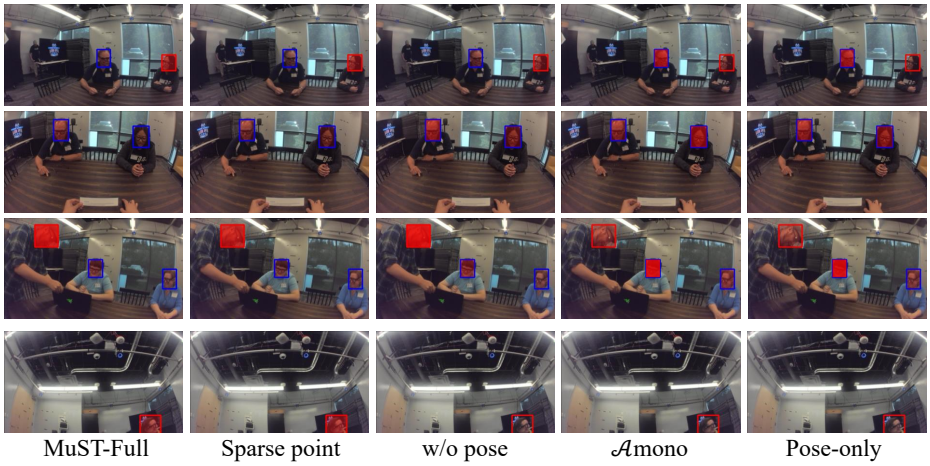


Fig. 9: Additional qualitative examples of egocentric active speaker localization on EasyCom [2].

blur (row 1), hand occlusion (row 2), and field-of-view limitations (row 3) hinder precise bounding box annotations.

C.2 Usage of Pose Information as Inputs

One conceivable limitation of our framework is that we use the wearer's pose information as the model's input. Although the wearer's pose information is quite efficient to obtain (*e.g.*, IMUs) and commonly available in multiple datasets and devices [1, 3–5], there are videos in the wild without that modality. In this case, several methods can potentially be utilized to infer poses from other modalities. For example, as outlined in Fig. 13, there is a good correlation between the wearer's head rotation on a world-locked sphere parametrized by (θ, φ) and dense optical flow parametrized by the X and Y axes. This implies that optical flow in these videos could be exploited to approximate IMUs for spherical world-locking, which we leave for future work.

C.3 Responsibility and Broader Impact

Spherical world-locking and the MuST framework could help improve people's quality of life through multiple technological applications, *e.g.*, hearing enhancement or smart homes. Still, using egocentric videos or multisensory observations from the wearer might be a source of privacy infringement or personal data abuse. All datasets we use in experiments are IRB-approved and collected with the consent of all participants, where personally identifiable information is further filtered out. Moreover, considering that the performance of our model without visual inputs outperforms prior arts by a large margin, our proposed framework could serve as a more privacy-aware alternative for audio-visual localization than prior arts.

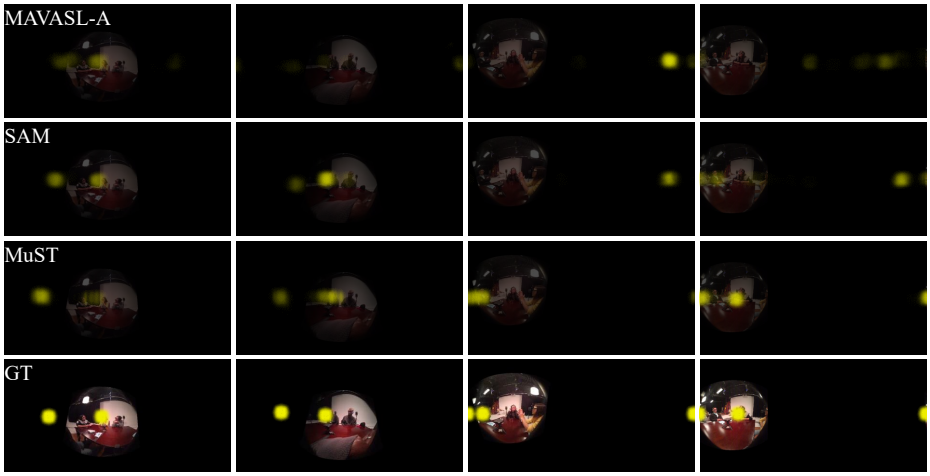
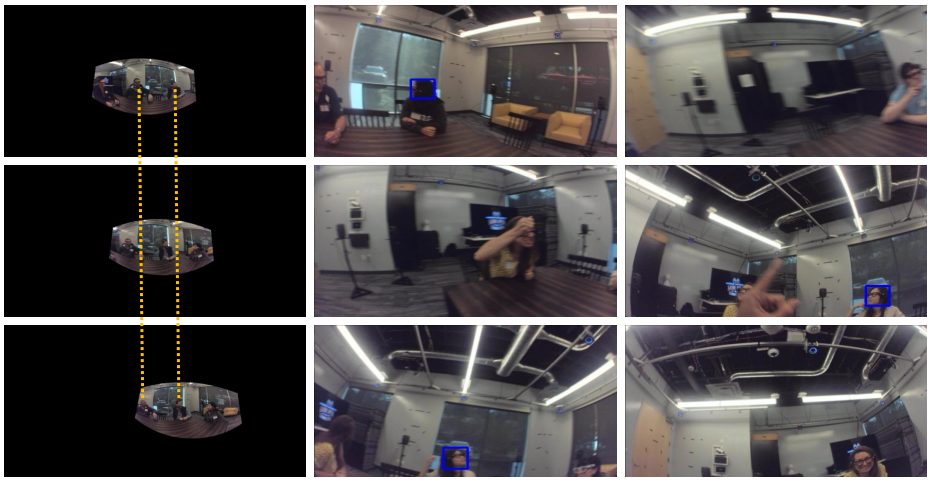


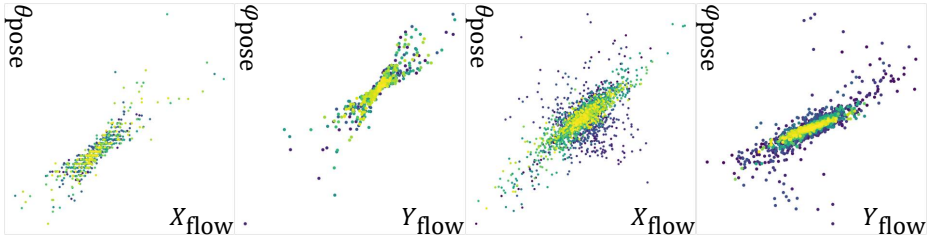
Fig. 10: Additional qualitative examples of audio-based spherical source localization on RLR-CHAT [6].

Fig. 11: Additional qualitative examples of egocentric behavior anticipation on Aria Pilot [5].



(a) Temporal verification (b) Spatial verification

Fig. 12: Spatial and temporal verification of annotations in EasyCom [2].



(a) Pose-flow correlation in EasyCom (b) Pose-flow correlation in Aria Pilot

Fig. 13: Correlation analysis between head rotation on a world-locked sphere and dense optical flow in a randomly sampled video from each dataset. The Pearson correlation coefficients are 0.8903 and 0.7325, respectively, with the p-value less than 1e-3.

References

1. 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)
2. 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)
3. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)
4. 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)
5. Somasundaram, K., Dong, J., Tang, H., Straub, J., Yan, M., Goesele, M., Engel, J.J., De Nardi, R., Newcombe, R.: Project Aria: A new tool for egocentric multi-modal ai research. arXiv:2308.13561 (2023)
6. 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)