

EGOADAPT: Adaptive Multisensory Distillation and Policy Learning for Efficient Egocentric Perception **Supplementary**

We add the following details in this supplementary:

- 1 Supplementary Video
- 2 Model and Implementation Details
- 3 More Related Works
- 4 Selection of Frames
- 5 Comparison with Other Teacher Models
- 6 Discussions on SOTA
- 7 Details on Efficiency Metrics
- 8 Performance in Noisy Condition
- 9 More Qualitative Examples
- 10 Occlusion and Conflicting signals results
- 11 Failure Modes
- 12 Details on Evaluation Metrics
- 13 Ablation on Choice of Audio Channels
- 14 Model Component Ablations
- 15 CFD Loss Ablations
- 16 Combined Algorithm
- 17 Frame Sampling and Policy Activation

1. Supplementary Video

In our supplementary video, we have a brief overview of EgoAdapt outlining all the training stages. We also add illustrative results for egocentric action recognition and active speaker localization tasks.

2. Model and Implementation Details

Active Speaker Localization & Behaviour Anticipation.

The policy submodule is designed to efficiently process and integrate various data streams by using three lightweight modality encoders, one for video, one for audio, and one for sensor inputs. Each encoder has a compact architecture with 2 convolutional layers, each followed by a max pooling operation and a ReLU activation. For video and audio encoder, we use 2-D convolution and MaxPooling layer, whereas 1-D convolution and max pool are used for the sensor encoder. The first convolutional layer extracts essential low-level features from the input data, while the subsequent max pooling layer reduces spatial dimensions, thereby enhancing computational efficiency. The second convolutional layer further refines these features, and the ReLU activation introduces the non-linearity needed to cap-

ture complex patterns. This streamlined yet effective design enables the policy submodule to rapidly and accurately extract modality-specific features, ensuring that critical temporal and spatial information is retained for downstream decision-making processes regarding “*which modality to use*.” We use K parallel FC layers on top of LSTM outputs to generate the binary decision policy for each modality or the chosen frame.

The Student module features 2 multi-head attention (MHA) layers followed by a fully connected (FC) layer to efficiently capture and synthesize information. In the MHA module, we use 8 heads. The MHA layers allow the module to learn complex relationships by focusing on different parts of the input simultaneously, ensuring that various patterns and dependencies are recognized from multiple perspectives. Once these rich, context-aware representations are achieved, the FC layer of dimension 512 integrates and consolidates the information into a final output, supporting decision-making or subsequent processing tasks. This design not only enhances the model’s ability to capture nuanced patterns but also improves performance and efficiency.

Activity Recognition. The policy submodule extracts time-aware audio features for action recognition by integrating three processing stages into one cohesive framework. Initially, 3 multi-head attention layers with 8 heads, enable the network to focus on different segments of the audio simultaneously, computing relationships between every pair of time steps and yielding refined features that capture global dependencies. These features are then complemented by RCNN layers, which combine convolutional operations to extract local acoustic patterns with recurrent processing to capture short-term dynamics. Each RCNN layer consists of a convolutional layer using a 3×3 kernel with 64 filters, a stride of 1, and “*same*” padding to maintain spatial dimensions, followed by batch normalization and a ReLU activation. We use bidirectional LSTM with 128 hidden units as the recurrent layer in RCNN. Information from both the MHA and RCNN layers is shared through a handshaking mechanism that ensures seamless integration of global and local insights. Finally, an LSTM layer of dimension 256 synthesizes the combined long- and short-term features to

determine the region of interest for action recognition accurately.

We follow the FasterNet [4] architecture as our student model due to its capacity to reduce FLOPS and enhance efficiency. FasterNet is a hierarchical convolutional neural network comprising four stages that employs the Partial Convolution (PConv) operation to minimize computational redundancy and memory access. Each stage initiates with either an embedding or merging layer to down-sample spatial dimensions and increase channel depth, followed by stacks of FasterNet blocks. Each FasterNet block first applies a 3×3 PConv that selectively processes the input channels, while the remaining channels bypass this operation. Next, a 1×1 pointwise convolution (PWConv) expands the channel dimension, and this is immediately followed by Batch Normalization and a ReLU activation function. This architecture, which resembles an inverted residual structure with shortcut connections, is designed to be straightforward, hardware-friendly, and efficient, making it well-suited for rapid inference within computational budgets.

Teacher Models. We employ TIM [2] as the teacher model for the action recognition and MUST for [59] active speaker localization and behavior anticipation tasks, respectively.

Hyper-parameters. We set $\alpha = 0.90$ and $\beta = 0.85$ based on validation data. For EpicKitchens, we set $\tau = 1.0$ and train for 150 epochs. For EasyCom we set $\tau = 10.0$ and train the model for 50 epochs.

3. More Related Works

Egocentric Video Understanding. In the field of egocentric video understanding, numerous studies have demonstrated that incorporating additional modalities can greatly enhance performance [22, 27, 34, 43, 60]. The hypothesis is straightforward: certain actions are more effectively understood through specific modalities. For instance, identifying that a person is ‘waving their hand’ can be determined using motion trajectories alone [34, 45]. However, these studies assume that all modalities used during training are also accessible during inference and that the computational resources are sufficient to process modalities beyond the RGB frames. Consequently, some work [40, 53] have effectively used Faster-RCNN [46] and other object detection-specific models at inference time. In contrast, we argue that dynamically computing additional modalities for egocentric video understanding may be impractical. Thus, we propose a distillation and policy learning-based approach that learns how to optimally use various downstream modalities for efficient inference.

Egocentric Action Recognition. Incorporating contextual information, such as the motions of human body parts and details about active objects, offers a promising approach to egocentric action recognition. Several methods have been

developed to leverage hand information, as the actor’s hands provide crucial contextual cues [25, 51]. Some studies [23, 37] have addressed action recognition by incorporating information about the actor’s gaze, focusing on where they look during actions. Given that many actions in first-person videos involve interactions between the actor and objects, various methods have been developed to leverage information about the active objects [18, 33]. Similar to our approach, several existing methods [16, 30, 61] integrate multiple types of contextual information for this task. While feature fusion is a promising strategy for boosting recognition performance, the associated increase in computational costs (e.g., larger model size) during inference is often overlooked. In contrast, our method learns to optimally use resources without compromising performance during inference.

Multi-modal Learning. Multimodal learning, in contrast to traditional single-modality approaches [44, 57], has made significant strides in areas such as cross-modal generation [8, 11, 12, 35, 49], audio-visual representation learning [6, 7, 19, 47], multimodal large language models [10, 13, 14, 41], and cross-modal integration [9, 17, 38, 39, 42, 43]. Recent studies have advanced cross-modal generation by utilizing visual and/or language context to produce coherent, complex audio [8, 11]. Work on active audio-visual separation and embodied agents emphasizes the significance of motion and egocentric perception in developing robust representations. These concepts naturally extend to audio-visual LLMs [3, 17], where perceptually grounded models engage with dynamic environments. In vision-language integration, recent research highlights the effectiveness of alignment across modalities [9]. Collectively, these efforts illustrate the importance of dynamic, embodied perception in creating versatile multimodal systems.

4. Selection of Frames

Tab. 1 compares the performance under different frame selection strategies. Upon randomly selecting a frame from potentially distinct activity regions (as determined by previewing audio) results in much inferior ‘Noun’ detection performance which subsequently results in poor action recognition results. Similar performances are observed when the *first* and *middle* frames are chosen while there is a slight improvement in performance when all the underlying frames are chosen within a region of interest albeit at a much higher compute cost. We note, that EgoAdapt can achieve a great balance between performance and efficiency by optimally choosing the most informative frame.

5. Comparison with Other Teacher Models

In this section, we compare the performance of EgoAdapt while distilled from other teacher models for each task as

Model	Verb \uparrow	Noun \uparrow	Action \uparrow	GMACs \downarrow
Fixed stride sampling	72.32	59.25	46.81	7.14
Random frame	73.93	60.86	48.63	7.14
First frame	76.16	61.63	52.25	7.14
Middle frame	76.24	64.06	52.98	7.14
All frames	76.84	66.96	56.97	<u>16.51</u>
EgoAdapt	<u>76.65</u>	<u>66.83</u>	<u>56.74</u>	7.14

Table 1. **Effect of visual frame selection strategy.** We compare the performance under different modes of key frame selection for action recognition task on Epic-Kitchens dataset.

outlined below.

5.1. Action Recognition

Similar to ASL, we employ MoViNet [28], MeMViT [55], MBT [40] as our teacher model as reported in Tab. 2. The distilled student model coupled with TeMPLe is able to closely emulate the teacher model’s performance by operating at up to $\sim 28\times$ less GMACs.

5.2. Active Speaker Localization

We systematically employ TalkNet [50] and MAVASL [24] in addition to MUST [59] as our teacher model to compare the performance of EgoAdapt as reported in Tab. 3. Experimental results demonstrate that in all the cases our proposed approach is able to closely replicate the teacher model’s performance while operating at a very low computation budget.

5.3. Behavior Anticipation

For behavior anticipation, the selected teacher models are SOTA approaches GazeMLE [31] and GLC [29] (Tab. 4). Experimental results indicate that EgoAdapt is able to achieve similar performance as the heavy teachers (GMACs compared in the main paper) at a very lower compute cost.

The performance on these three tasks with varying teacher models underlines the efficiency of our proposed training paradigm. We claim that EgoAdapt is teacher model agnostic and is able to replicate the heavy teacher model’s performance while operating under a constrained compute setting thereby maintaining a good balance between performance and efficiency.

6. Discussions on SOTA

6.1. Active Speaker Localization

For a more comprehensive study, we compare our method against SOTA methods. MAVASL [24] combine audio-only and audio-visual networks to perform spherical and inner field-of-view active speaker localization while LoCoNet [54] learns a long-Short context network. Sync-TalkNet [58] models cross-modal information with complex attention modules. ASD-Trans [15] employs a ResNet-18 to ex-

tract audio features. LW-ASD [32] proposes a GRU based active speaker detection model.

6.2. Action Recognition

MoViNet [28] proposes a three-step approach to improve computational efficiency while substantially reducing the peak memory usage of 3D CNNs. TBN [26] introduces an audio-visual temporal binding network for egocentric action recognition. AdaFuse [36] presents an adaptive temporal fusion network for action recognition tasks. More recently, Ego-only [52] proposed an approach that enables action detection on egocentric videos without any form of exocentric (third-person) transferring.

6.3. Egocentric Behavior Anticipation.

Most existing works on egocentric gaze modeling target at egocentric gaze estimation rather than anticipation. We adapt gaze estimation models [29, 31] to compare our method against them. We also report the performance of a competitive baseline of Multitask Gaussian Process [20].

7. Details on Efficiency Metrics

7.1. GMACs

We use the native PyTorch FLOP counter to get the total FLOP count in the forward pass. We convert this to GMACs (approximately 2 FLOPs = 1 MAC) by dividing with 10^9 .

7.2. Energy Consumption

Accurately assessing the energy consumption of models is essential for their deployment in AR/VR devices [1, 5]. The energy expenditure arises from a complex interaction of factors, including sensors, computation, communication, data processing, memory transfers (SRAM and DRAM), and leakage—many of which are often overlooked when designing *efficient* models, despite their significant contribution to overall energy use.

Following prior work [21, 48], we consider three key factors when modeling energy consumption: (1) The energy required for each model forward pass, determined by the number of operations (MACs). (2) The energy cost of

Method	Input resolution ↓	Verb↑	Noun↑	Action↑	GMACs↓
MoViNet-A6 [28]	320 × 320	72.24	57.31	47.79	79.35
MeMViT [56]	224 × 224	71.4	60.3	48.4	161.90
MBT [40]	224 × 224	64.8	58.0	43.4	201.64
TIM AV [2]	224 × 224	77.19	67.22	57.57	26.62
EgoAdapt w/MoViNet	224 × 224	71.46	56.32	46.14	7.14
EgoAdapt w/MeMViT	224 × 224	70.67	59.91	48.25	7.14
EgoAdapt w/MBT	224 × 224	63.66	57.34	42.53	7.14
EgoAdapt w/TIM	224 × 224	76.65	66.83	56.74	7.14

Table 2. **Comparison with other models as teachers on EPIC-Kitchens.** We report the top-1 accuracy for verb, noun, and action (%).

Method	mAP↑	GMACs↓	Params (M)↓	Energy (J)↓
TalkNet [50]	69.13	3.17	15.7	0.518
MAVSL [24]	86.32	6.85	16.13	0.698
MUST [59]	89.88	0.642	2.17	0.029
EgoAdapt w/TalkNet	68.90	0.070	0.39	0.003
EgoAdapt w/MAVSL	85.74	0.070	0.39	0.003
EgoAdapt w/MUST	89.74	0.070	0.39	0.003

Table 3. **Performance of active speaker localization on EasyCom with other teachers.**

memory read-write operations, including storing intermediate activations and model outputs. (3) The energy involved in activating, deactivating, and continuously operating sensors (e.g., camera, audio, IMU).

We use GPUs as our processing device and employ the PyTorch memory profiler to capture a list of all operations performed during the forward pass (`model.forward()` call) along with their corresponding GPU memory usage. The total memory consumption is calculated as the sum of the memory costs for each individual operation.

For each modality, we track its active duration by counting the number of observations sampled that include the modality. We ensure that the sensors capture a minimum of 1 second’s worth of samples.

7.3. Trainable Parameters

These are simply the learnable parameters that are adjusted during training to minimize the overall loss function.

8. Performance in noisy conditions

Fig. 1 illustrates the performance of the Active Speaker Localization (ASL) model under varying audio noise levels, quantified by signal-to-noise ratio (SNR). As the SNR decreases from -5 dB to -20 dB (indicating progressively noisier audio conditions), the model adaptively increases its reliance on visual modalities, with video usage rising from 20.71% to 88.89%. Despite the degraded audio quality, the model maintains robust performance, evidenced by only a marginal decline in mAP from 84.27% to 80.21%. This underscores the effectiveness of the EgoAdapt framework in dynamically leveraging complementary modalities through

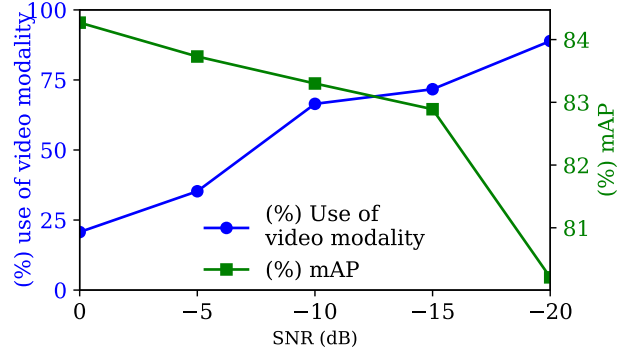


Figure 1. **Adaptive Modality Utilization** of EgoAdapt in the noisy scenario. EgoAdapt increases video modality reliance to compensate for degraded audio, maintaining stable model performance under varying noise conditions.

policy learning with the guidance of TeMPLe: under high noise, heightened video utilization compensates for unreliable audio cues, ensuring stable localization accuracy.

9. More Qualitative Examples

Fig. 2 compares action recognition performance between upto stage 2 training vs full training. As seen from the examples: the joint training of the two modules (CFD and TeMPLe) indeed improves the performance by providing more accurate and tighter action prediction results. In the first and second example, after full finetuning, EgoAdapt is able to choose a correct frame resulting in accurate action recognition.

Fig. 3 compares the performance of EgoAdapt with and

Method	Gaze			Orientation			Trajectory		
	$T_{300\text{ ms}} \downarrow$	$T_{500\text{ ms}} \downarrow$	$T_{700\text{ ms}} \downarrow$	$T_{300\text{ ms}} \downarrow$	$T_{500\text{ ms}} \downarrow$	$T_{700\text{ ms}} \downarrow$	$T_{300\text{ ms}} \downarrow$	$T_{500\text{ ms}} \downarrow$	$T_{700\text{ ms}} \downarrow$
GazeMLE [31]	10.74	14.37	18.14	4.68	9.11	12.03	14.33	16.02	18.64
GLC [29]	10.21	14.66	17.80	4.76	8.98	11.70	13.15	15.39	17.41
MuST _{AVB} [59]	<u>9.17</u>	<u>12.15</u>	14.75	4.78	7.36	9.90	9.96	12.38	13.95
EgoAdapt w/GazeMLE	10.91	14.61	18.78	4.74	9.89	12.60	14.85	16.36	18.98
EgoAdapt w/GLC	10.87	14.76	19.32	5.10	9.24	12.09	13.72	15.80	17.94
EgoAdapt w/MUST	8.53	11.93	14.58	4.61	<u>7.39</u>	<u>9.91</u>	9.58	11.97	13.36

Table 4. Comparison of behavior anticipation errors on the AEA Dataset with other teachers.

without TeMPLe for ASL task. We note that by learning to choose the useful combination of channels and sporadically making use of the visual modality our method is able to achieve improved performance when compared to without policy training.

Fig. 4 illustrates more qualitative samples from egocentric behavior anticipation task. We contrast the performance with and without the policy learning module. As seen from the examples, after employing TeMPLe the gaze prediction regions are more accurate and densely aligned to the GT.

10. Occlusion and Conflicting signals results

Refer to Fig. 3 – last row column 3: EgoAdapt can identify active speakers under heavy occlusion. The active speaker is looking away from the camera and his face is partially visible. As seen in the example, our approach is able to correctly identify the active speaker.

In the same figure, the person is pretending to talk by taking notes. However, our model correctly relies on audio modality to eliminate such cases.

11. Failure Modes

In this section we discuss failure modes of EgoAdapt. In Fig. 5 we note that since our model is not equipped with fine grained understanding of the actions and not supervised by language models, it predicts ‘spread butter’ while the original class label is ‘spread more butter’.

Frame 3 in Fig. 6 represents a case where the person (with the hat) is actually non-active while holding his pose. EgoAdapt due to its gaze bias is focusing towards this speaker resulting in wrong prediction results. While in frame 5, although the same speaker is active, our model is not able to detect him as he is partially visible.

12. Details on Evaluation Metrics

12.1. Action Recognition

To measure action recognition performance, we report the per-video top-1 accuracy on the validation set. We densely sample clips from each video and average their predictions to compute accuracy.

12.2. Active Speaker Localization

We follow prior works’ [24, 59] experimental settings for a fair comparison and report the mean average precision (mAP), which captures both spatial and temporal localization of speech activity inside the camera’s FOV. The mAP scores of all models are computed by pooling the maximum logit value within the corresponding head bounding boxes.

12.3. Behavior Anticipation

Our goal is to anticipate future behaviors in various reaction times (300/500/700ms) given the current audio-visual observations and previous behavioral contexts. To evaluate localization performance, we use Mean Angular Errors (MAE) of behaviors by comparing the argmax coordinate of the model’s prediction with ground truth behaviors following [59]. MAE between prediction to ground truth reflects how far the model’s prediction deviates from the ground truth source direction on a sphere.

13. Ablation on Choice of Audio Channels

We compare the performance of EgoAdapt under different predefined choices of audio channels in Tab. 5. Although selecting all 4 channels results in strong ASL performance, the best performance is achieved when strategically all the downstream modalities are leveraged. As we employ very lightweight modules for each modality, the overall GMAC values remain within considerable limits.

14. Model Components Ablations

In Tab. 6 we compare the contributions of the two main model components, i.e. TeMPLe (Π) and CFD (Φ). We observe that lower values of η_1 and η_2 imply lesser weightage on both the components resulting in suboptimal performance. As we gradually add more weightage to the policy component the performance improves, however, since the task module is learned by CFD, higher values of η_2 boosts the overall performance for both the tasks.

15. CFD Loss Ablations

We ablate α , β in Tab. 7 to compare the contributions of the loss components in the distillation module. While α directly controls the relative weightage for the \mathcal{L}_{KD} ; $(1 - \alpha)$

Algorithm 1 EgoAdapt: Training

Input: Video/Frames: \mathcal{V} ; Single/Multi-Channel Audio: \mathcal{A} ; Behavior (IMU, Gaze) info: \mathcal{B} ; Pre-trained Teacher: Θ_Ω ; Cross-modal Feature Distillation Module: **CFD**; Policy Module: **TeMPLe**; Ground Truth Labels: GT ; Task Type: \mathcal{T} ; Loss Hyperparameters: η_1, η_2 ;

Output: Trained Policy (combination of sub-networks): $\Theta_\Pi = \{\Theta_{\text{LSTM}}, \Theta_{\text{FC}_i}, \Theta_{\text{X}_i}\}, \forall i \in \{1, 2, \dots, k\}$; Trained Student: Θ_Φ .

- 1: $\Theta_\Phi \leftarrow \text{CFD}(\Theta_\Omega, \mathcal{V}, \mathcal{A}, \mathcal{B})$ \triangleright Distillation (Stage 1.) using Eq. (4)
 - 2: **while** not converged **do**
 - 3: $\Theta_\Pi \leftarrow \text{TeMPLe}(\Theta_\Phi, \mathcal{V}, \mathcal{A}, \mathcal{B}, \mathcal{T}, GT)$ \triangleright Policy Training (Stage 2) using Eq. (10).
 - 4: $\mathcal{L}_\Theta \leftarrow \eta_1 \mathcal{L}_\Pi + \eta_2 \mathcal{L}_\Phi$ \triangleright Combined Training (Stage 3) using Eq. (11).
 - 5: Optimize for Θ to reduce \mathcal{L}_Θ until convergence.
 - 6: **return** $\Theta = \{\Theta_\Phi, \Theta_\Pi\}$.
-

Audio Channels	Policy driven?	mAP \uparrow	GMACs \downarrow
[1, 2]	✗	71.66	0.059
[1, 3]	✗	73.38	0.059
[1, 4]	✗	68.25	0.059
[2, 3]	✗	69.57	0.059
[2, 4]	✗	66.19	0.059
[3, 4]	✗	51.87	0.059
[1, 2, 3]	✗	82.69	0.062
[1, 2, 4]	✗	83.62	0.062
[1, 3, 4]	✗	80.48	0.062
[2, 3, 4]	✗	80.22	0.062
[1, 2, 3, 4]	✗	87.91	0.068
EgoAdapt	✓	89.74	0.073

Table 5. **Effect of choice of audio channels.** We systematically study the effect of various combinations of audio channels for ASL tasks on the EasyCom dataset.

η_1	η_2	EK-100 (acc) \uparrow	EC (mAP) \uparrow
0.10	0.20	45.22	68.19
0.50	0.20	46.21	70.35
0.75	0.20	47.80	71.48
0.95	0.20	49.52	72.81
0.95	0.50	52.68	75.95
0.95	0.80	56.08	88.28
0.95	1.2	56.74	89.74

Table 6. η_1 and η_2 ablation results.

controls the weightage for the \mathcal{L}_{GT} and β controls that of \mathcal{L}_1 loss. We note that when α is small due to less weightage on \mathcal{L}_{KD} the student model is not able to replicate the teacher model’s performance resulting in inferior performance. Higher values of α improve performance. Similarly, as \mathcal{L}_1 helps to bring the feature level representations between the *teacher* and *student* models closer, higher values of β result in better performance. However, we observe that the optimal results are obtained when α and β are set to 0.90 and 0.85 respectively which helps to maintain a good

balance between different loss components.

α	β	EK-100 (acc) \uparrow	EC (mAP) \uparrow
0.10	0.50	43.63	78.09
0.40	0.50	46.15	80.23
0.70	0.50	48.73	82.49
0.90	0.50	51.41	85.57
0.90	0.75	53.54	87.80
0.90	0.85	56.74	89.74

Table 7. α and β ablation results.

16. Combined Algorithm

We provide the overall EgoAdapt training algorithm in Algorithm 1. Our key novelty is jointly training a policy and a distillation modules detailed in Line 3-4. Our approach can learn task-specific action spaces and coupled with distilled student modules achieve a great balance between performance and efficiency as shown through extensive experiments.

17. Frame Sampling and Policy Activation

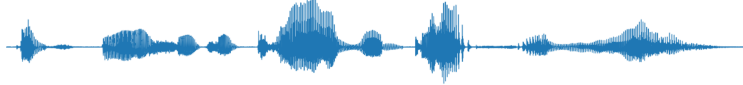
We perform sparse frame sampling for AR through audio previewing to avoid processing the heavy \mathcal{V} modality. The semantic information obtained through sparsely sampled frames is sufficient for AR task. For the other two tasks, our policy module learns to sample optimal choice of modalities. Hence the visual modality is being used in all the cases, albeit in different ways.

The policy module is activated depending on the task, and both the subnetworks can run concurrently.

**Selected
Frames**



Audio



GT Action

Open Fridge Put Grapes Grab Salmon Put Salmon

GT Frame

253-311 307-469 528-589 575-655

Upto Stage 2

PRED Frame

261 302 581 627

PRED Action

Open Fridge Open Fridge Put Salmon Put Salmon

EgoAdapt

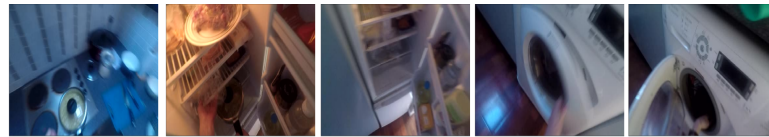
PRED Frame

278 319 553 611

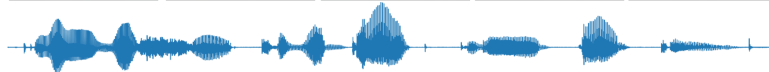
PRED Action

Open Fridge Put Grapes Grab Salmon Put Salmon

**Selected
Frames**



Audio



GT Action

Grab Saucepan Put Saucepan in the Fridge Close Fridge Open Washing Machine Check Dryness

GT Frame

780-810 936-1070 1241-1322 1441-1518 1523-1612

Upto Stage 2

PRED Frame

801 1011 1301 1489 1511

PRED Action

Open Fridge Put Saucepan in the Fridge Close Fridge Open Washing Machine Open Washing Machine

EgoAdapt

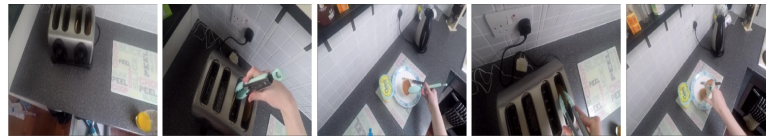
PRED Frame

789 1012 1277 1471 1601

PRED Action

Grab Saucepan Put Saucepan in the Fridge Close Fridge Open Washing Machine Check Dryness

**Selected
Frames**



Audio



GT Action

Grab Tongs Remove Pancake Place Pancake Remove Pancake Place Pancake

GT Frame

274-318 423-561 583-642 684-1275 1269-1345

Upto Stage 2

PRED Frame

333 312 630 982 1301

PRED Action

Take Glass Remove Pancake Place Pancake Remove Pancake Place Pancake

EgoAdapt

PRED Frame

312 492 617 1073 1337

PRED Action

Place Pancake Remove Pancake Place Pancake Remove Pancake Place Pancake

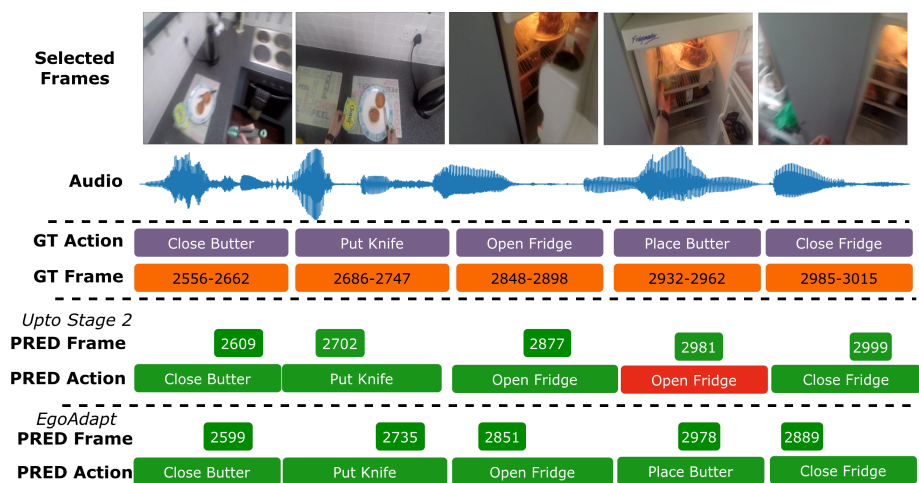
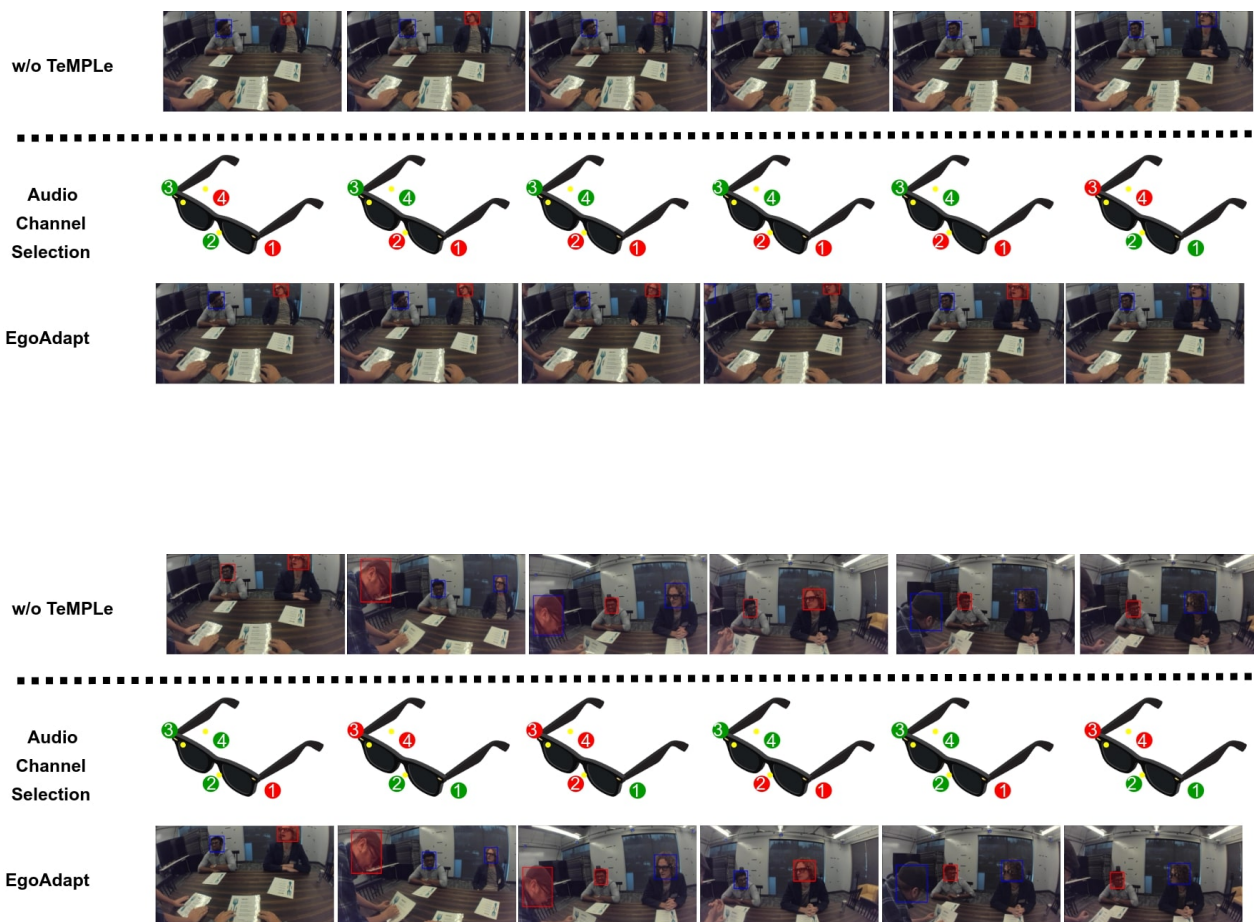


Figure 2. More qualitative examples of Action Recognition on the Epic-Kitchens Dataset.



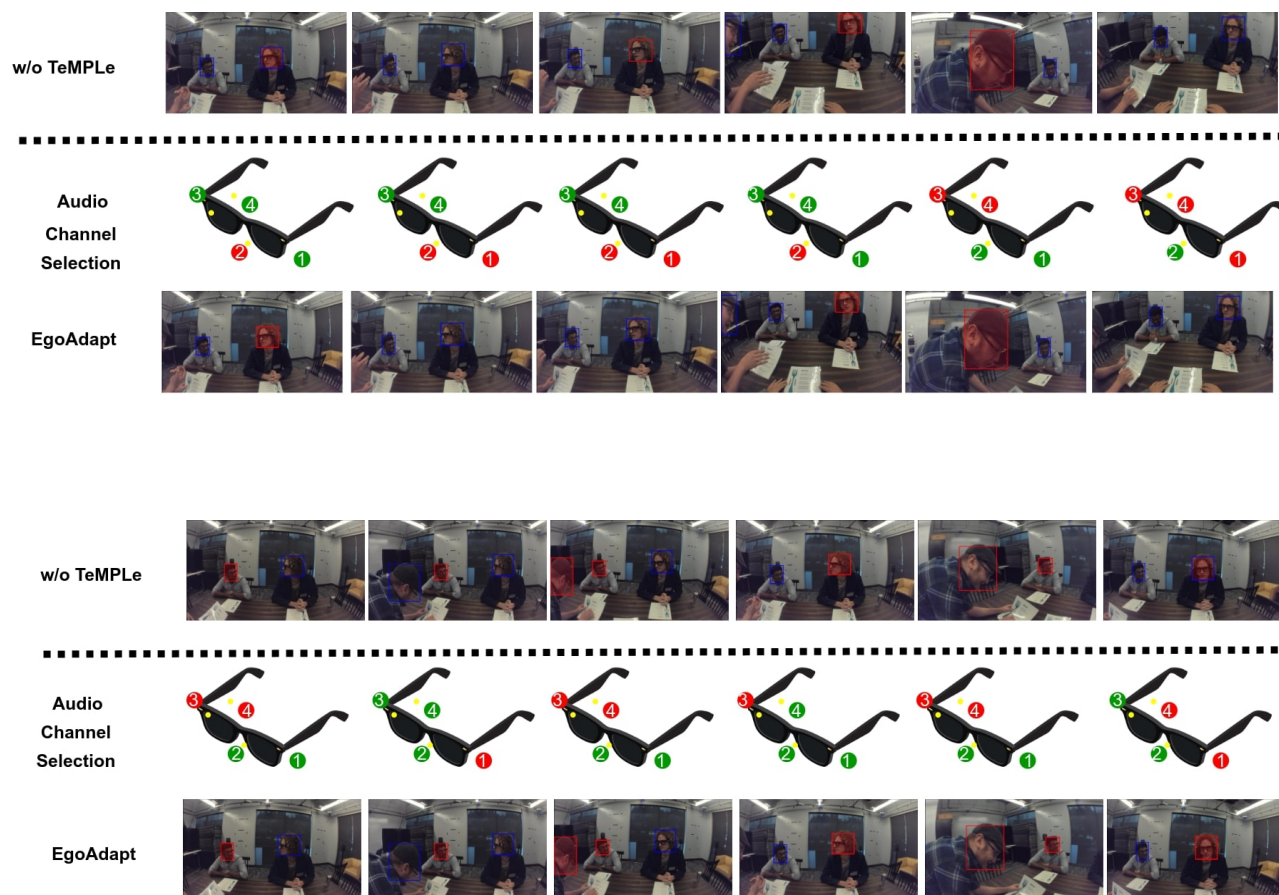


Figure 3. More qualitative examples of Active Speaker Localization on the EasyCom Dataset.

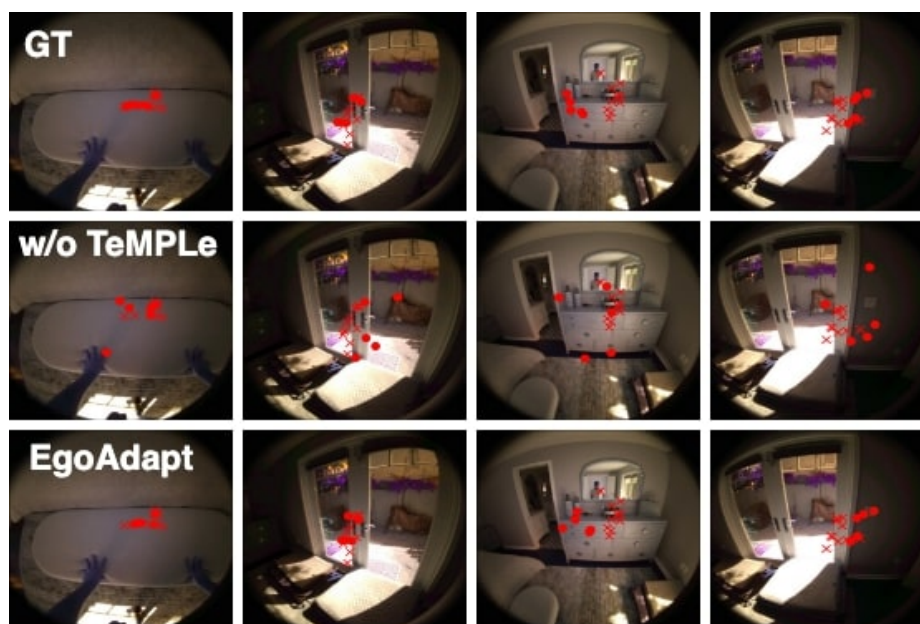




Figure 4. **Qualitative examples of egocentric behavior anticipation on the AEA Dataset.** Cross/circle symbols denote previous/anticipated behaviors.

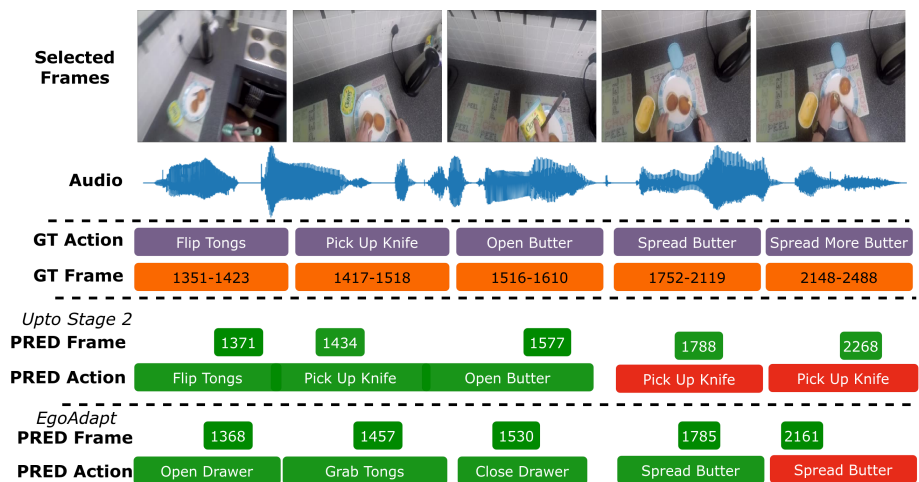


Figure 5. **Failure case for egocentric action recognition on EPIC-Kitchens dataset.**

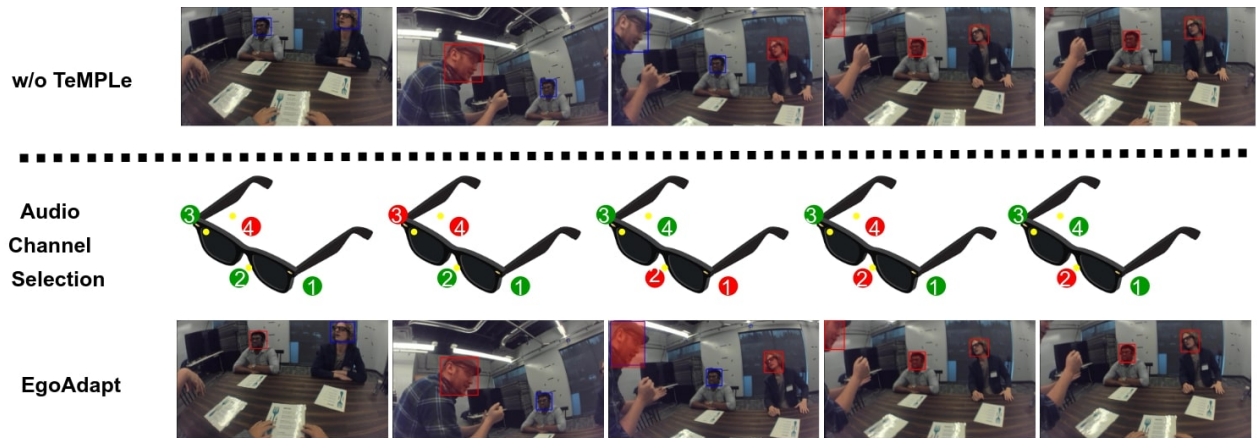


Figure 6. Failure case for egocentric ASL on EasyCom Dataset.

References

- [1] Michael Abrash. Creating the future: Augmented reality, the next human-machine interface. In *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021.
- [2] Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. Tim: A time interval machine for audio-visual action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18153–18163, 2024.
- [3] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021.
- [4] Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S-H Gary Chan. Run, don’t walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12021–12031, 2023.
- [5] Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, and Vivienne Sze. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2019.
- [6] Sanjoy Chowdhury, Subhrajyoti Dasgupta, Sudip Das, and Ujjwal Bhattacharya. Listen to the pixels. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2568–2572. IEEE, 2021.
- [7] Sanjoy Chowdhury, Aditya Patra, Subhrajyoti Dasgupta, and Ujjwal Bhattacharya. Audvisum: Self-supervised deep reinforcement learning for diverse audio-visual summary generation. In *BMVC*, page 315, 2021.
- [8] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7884–7896, 2023.
- [9] Sanjoy Chowdhury, Sayan Nag, and Dinesh Manocha. Apollo: unified adapter and prompt learning for vision language models. *arXiv preprint arXiv:2312.01564*, 2023.
- [10] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision*, pages 52–70. Springer, 2024.
- [11] Sanjoy Chowdhury, Sayan Nag, KJ Joseph, Balaji Vasani, Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26826–26835, 2024.
- [12] Sanjoy Chowdhury, Mohamed Elmoghany, Yohan Abeyasinghe, Junjie Fei, Sayan Nag, Salman Khan, Mohamed Elhoseiny, and Dinesh Manocha. Magnet: A multi-agent framework for finding audio-visual needles by reasoning over multi-video haystacks. *arXiv preprint arXiv:2506.07016*, 2025.
- [13] Sanjoy Chowdhury, Hanan Gani, Nishit Anand, Sayan Nag, Ruohan Gao, Mohamed Elhoseiny, Salman Khan, and Dinesh Manocha. Aurelia: Test-time reasoning distillation in audio-visual llms. *arXiv preprint arXiv:2503.23219*, 2025.
- [14] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. *arXiv preprint arXiv:2501.02135*, 2025.
- [15] Gourav Datta, Tyler Etchart, Vivek Yadav, Varsha Hedau, Pradeep Natarajan, and Shih-Fu Chang. Asd-transformer: Efficient active speaker detection using self and multimodal transformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4568–4572. IEEE, 2022.
- [16] Eadom Dessalene, Chinmaya Devaraj, Michael Maynard, Cornelia Fermüller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6703–6714, 2021.
- [17] Vishnu Sashank Dorbala, Sanjoy Chowdhury, and Dinesh Manocha. Can llms generate human-like wayfinding instructions? towards platform-agnostic embodied instruction synthesis. *arXiv preprint arXiv:2403.11487*, 2024.
- [18] Alireza Fathi and James M Rehg. Modeling actions through state changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2586, 2013.
- [19] Junyu Gao, Hao Yang, Maoguo Gong, and Xuelong Li. Audio-visual representation learning for anomaly events detection in crowds. *Neurocomputing*, 582:127489, 2024.
- [20] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.
- [21] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.
- [22] Roei Herzig, Elad Ben-Avraham, Kartikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3148–3159, 2022.
- [23] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.
- [24] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10552, 2022.
- [25] Georgios Kapidis, Ronald Poppe, Elsbeth Van Dam, Lucas Noldus, and Remco Veltkamp. Multitask learning to improve egocentric action recognition. In *Proceedings of*

- the *IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [26] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019.
 - [27] Tae Soo Kim, Jonathan Jones, and Gregory D Hager. Motion guided attention fusion to recognize interactions from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13076–13086, 2021.
 - [28] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16020–16030, 2021.
 - [29] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *arXiv preprint arXiv:2208.04464*, 2022.
 - [30] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *CVPR*, 2015.
 - [31] Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45(6): 6731–6747, 2021.
 - [32] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22932–22941, 2023.
 - [33] Yang Liu, Ping Wei, and Song-Chun Zhu. Jointly recognizing object fluents and tasks in egocentric videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2924–2932, 2017.
 - [34] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1049–1059, 2020.
 - [35] Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaoheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. Foleygen: Visually-guided audio generation. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2024.
 - [36] Yue Meng, Rameswar Panda, Chung-Ching Lin, Prasanna Sattigeri, Leonid Karlinsky, Kate Saenko, Aude Oliva, and Rogerio Feris. Adafuse: Adaptive temporal fusion network for efficient action recognition. *arXiv preprint arXiv:2102.05775*, 2021.
 - [37] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021.
 - [38] Trisha Mittal, Sanjoy Chowdhury, Pooja Guhan, Snikitha Chelluri, and Dinesh Manocha. Towards determining perceived audience intent for multimodal social media posts using the theory of reasoned action. *Scientific Reports*, 14(1): 10606, 2024.
 - [39] Sayan Nag, Koustava Goswami, and Srikrishna Karanam. Safari: Adaptive sequence transformer for weakly supervised referring expression segmentation. In *European Conference on Computer Vision*, pages 485–503. Springer, 2024.
 - [40] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.
 - [41] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. *arXiv preprint arXiv:2312.12423*, 2023.
 - [42] Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik J Shah, Yann LeCun, and Rama Chellappa. Volta: Vision-language transformer with weakly-supervised local-feature alignment. *Transactions on Machine Learning Research*, 2023.
 - [43] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.
 - [44] BH Prasad, Lokesh R Boregowda, Kaushik Mitra, Sanjoy Chowdhury, et al. V-desirr: Very fast deep embedded single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2390–2399, 2021.
 - [45] Gorjan Radevski, Marie-Francine Moens, and Tinne Tuytelaars. Revisiting spatio-temporal layouts for compositional action recognition. *arXiv preprint arXiv:2111.01936*, 2021.
 - [46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
 - [47] Parthasaarathy Sudarsanam, Irene Martín-Morató, and Tuomas Virtanen. Representation learning for semantic alignment of language, audio, and visual modalities. *arXiv preprint arXiv:2505.14562*, 2025.
 - [48] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. How to evaluate deep neural network processors: Tops/w (alone) considered harmful. *IEEE Solid-State Circuits Magazine*, 2020.
 - [49] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434, 2024.
 - [50] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3927–3935, 2021.
 - [51] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 4511–4520, 2019.

- [52] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5250–5261, 2023.
- [53] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.
- [54] Xizi Wang, Feng Cheng, and Gedas Bertasius. Loconet: Long-short context network for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18462–18472, 2024.
- [55] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.
- [56] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.
- [57] Jiaye Wu, Sanjoy Chowdhury, Hariharmano Shanmugaraja, David Jacobs, and Soumyadip Sengupta. Measured albedo in the wild: Filling the gap in intrinsics evaluation. In *2023 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2023.
- [58] Abudukelimu Wuerkaixi, You Zhang, Zhiyao Duan, and Changshui Zhang. Rethinking audio-visual synchronization for active speaker detection. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 01–06. IEEE, 2022.
- [59] Heeseung Yun, Ruohan Gao, Ishwarya Ananthabhotla, Anurag Kumar, Jacob Donley, Chao Li, Gunhee Kim, Vamsi Krishna Ithapu, and Calvin Murdock. Spherical world-locking for audio-visual localization in egocentric videos. In *European Conference on Computer Vision (ECCV)*, 2024.
- [60] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Is an object-centric video representation beneficial for transfer? In *Proceedings of the Asian Conference on Computer Vision*, pages 1976–1994, 2022.
- [61] Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, and Qi Tian. Cascaded interactional targeting network for egocentric video analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1904–1913, 2016.