# EgoAdapt: A Joint Distillation and Policy Learning Framework for Efficient Multisensory Egocentric Perception

📍Wed 22 Oct | 11:15 a.m. HST — 1:15 p.m. HST | Poster Session 3 | Exhibit Hall I #983

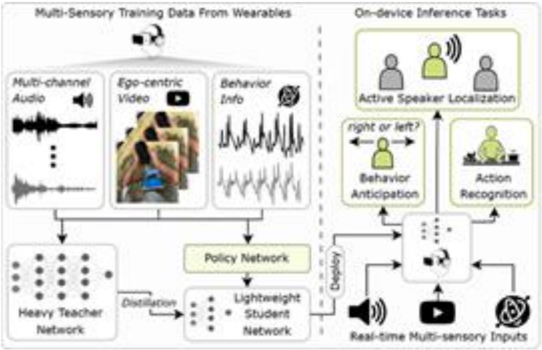

Sanjoy Chowdhury[1,2], Subrata Biswas[2,3], Sayan Nag[4], Tushar Nagarajan[2], Calvin Murdock[2], Ishwarya Ananthabhotla[2], Yijun Qian[2], Vamsi Krishna Ithapu[2], Dinesh Manocha[1], Ruohan Gao[1,2]

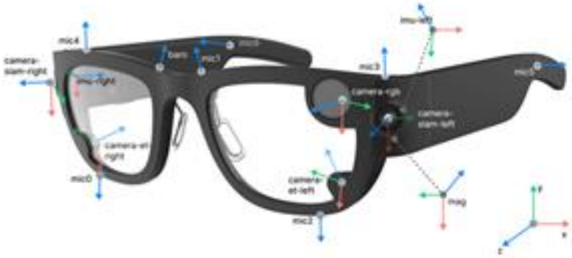
[1]University of Maryland, College Park, [2]Meta, [3]Worcester Polytechnic Institute, [4]University of Toronto


Oct 19th 2025

# Motivation

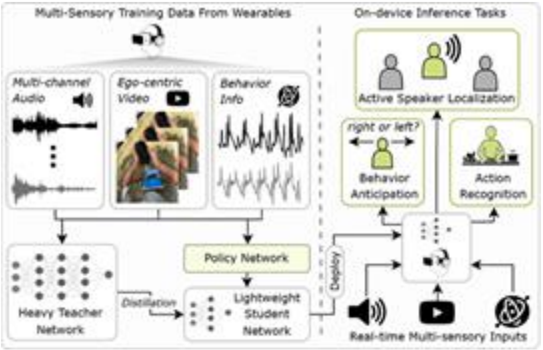
- Egocentric multi-modal perception is challenging because:
  - Limited compute in wearable devices
  - Needs real-time processing

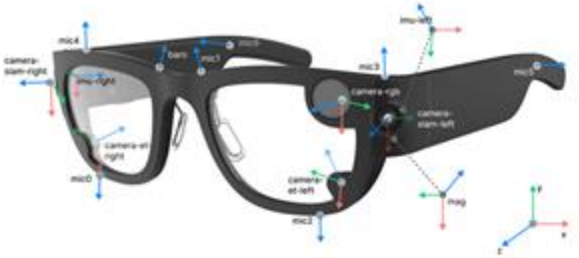


# Motivation

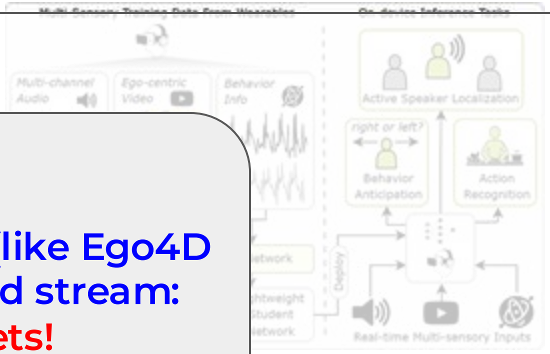
- Egocentric multi-modal perception is challenging because:
    - Limited compute in wearable devices
    - Needs real-time processing
    - Wait, but performance is important too!!



# Motivation

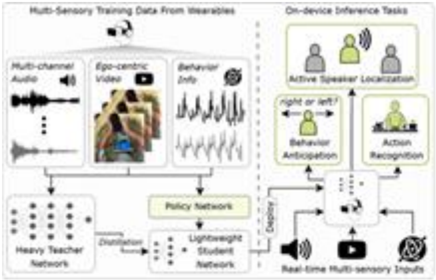
- Egocentric multi-modal perception is challenging because:
  - Limited c
  - Needs rea
  - Wait, but

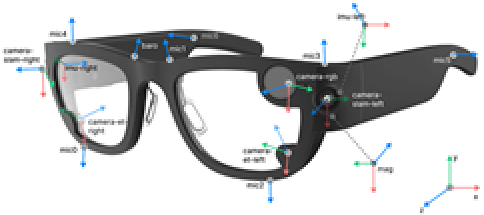



Typical 3D CNN-based egocentric models (like Ego4D baselines) require >100 GMACs per second stream: Infeasible for on-device AR headsets!

# Motivation

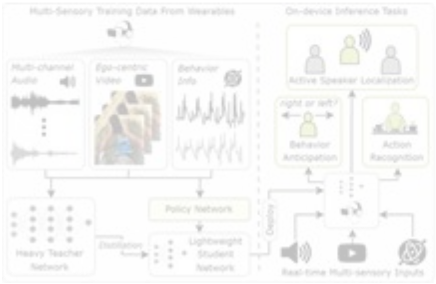
- Egocentric multi-modal perception is challenging because:
  - Limited compute in wearable devices
  - Needs real-time processing
  - Wait, but performance is important too!!



- Key Questions?
  - Are all of them important at all the time?
  - What all modalities do we leverage for optimal efficiency?
  - What is the smartest way to switch between the modalities at our disposal?

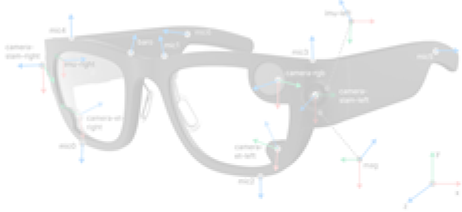
# Motivation



- Egocentric multi-modal perception is challenging because:
    - Limited compute in wearable devices
    - Ne
    - W

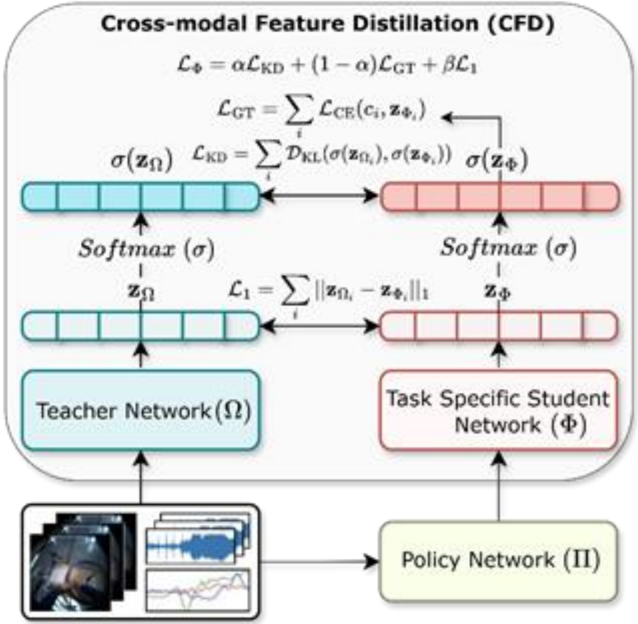
Can we combine model distillation with policy learning?

- Key Questions?
    - Are all of them important at all the time?
    - What all modalities do we leverage for optimal efficiency?
    - What is the smartest way to switch between the modalities at our disposal?

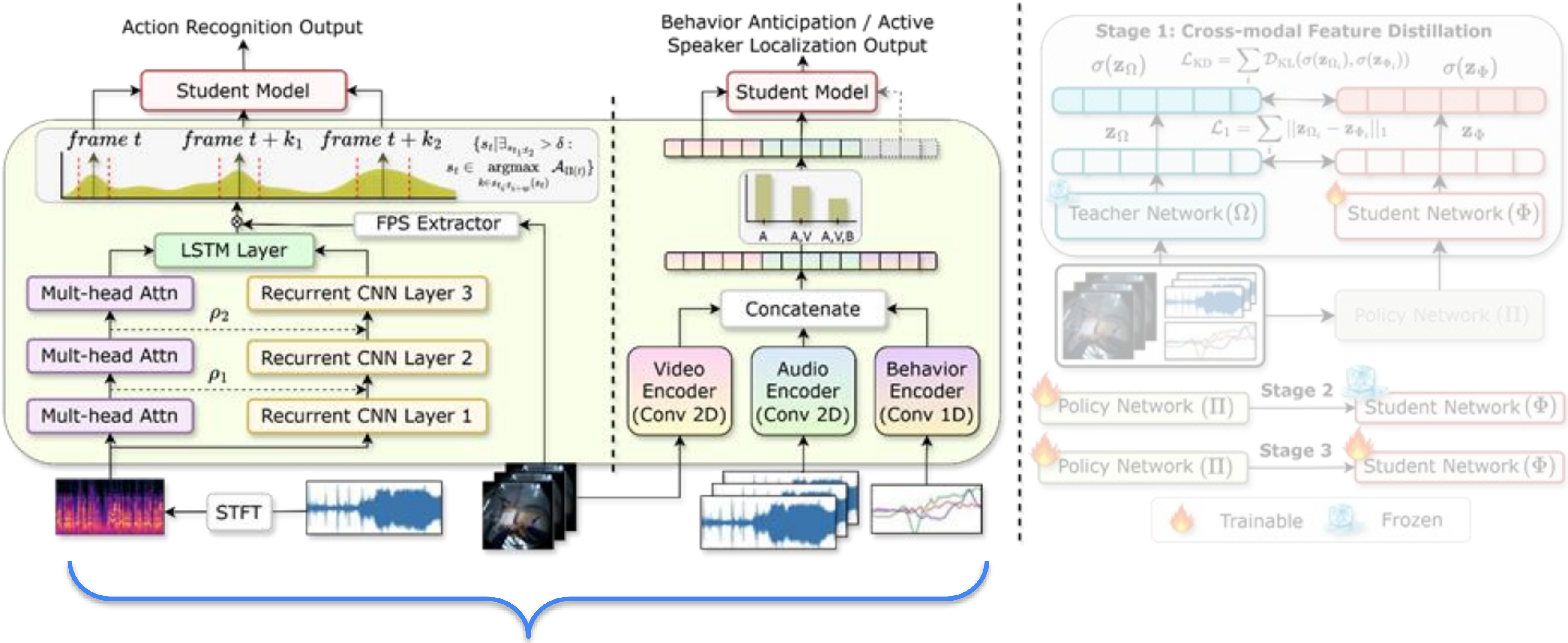


# Distillation module

- $\mathcal{L}_{KD}$ Aligns soft targets (semantic space)
- $\mathcal{L}_1$ Aligns intermediate representations (feature space)
- $\mathcal{L}_{GT}$ Anchors final task objective (logit space).

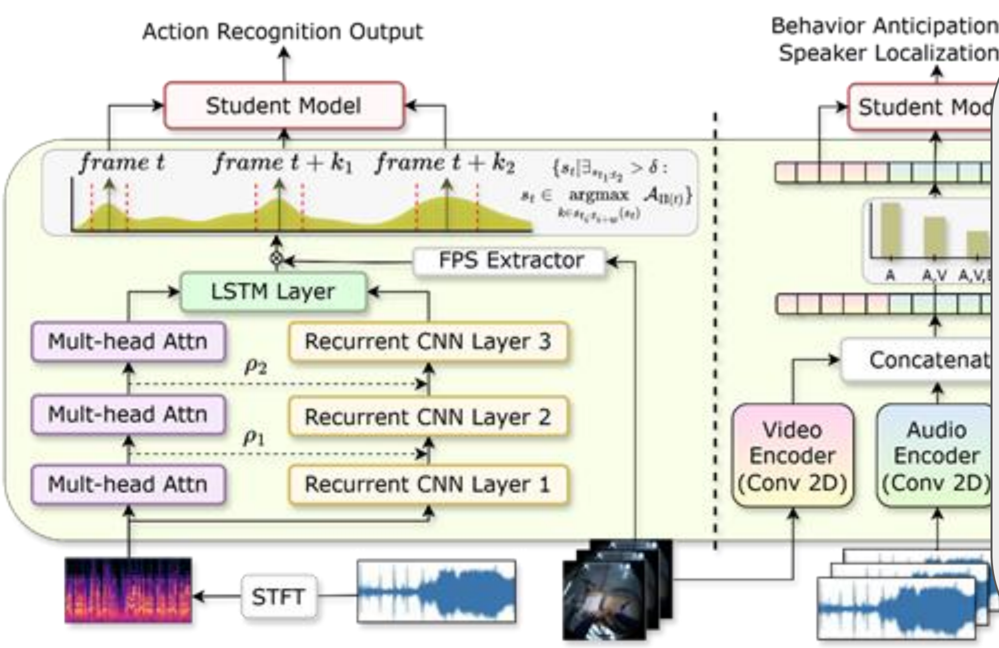


# Policy module




Task-specific policy modules

# Policy module

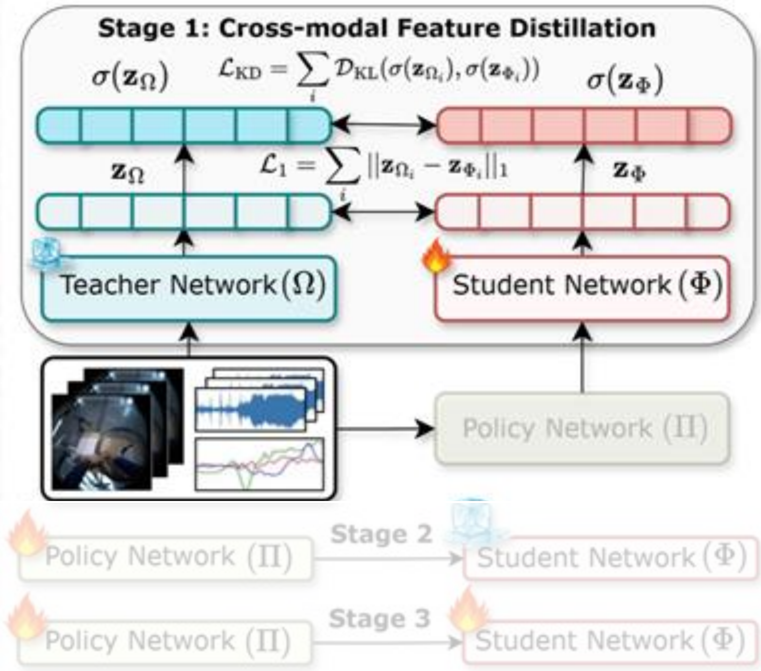



$$\mathcal{P}_k = \mathtt{Softmax}((\log z_k + G_k)/\tau)$$

Unify continuous (distillation) and discrete (policy) learning streams via joint optimization over differentiable sampling and backpropagation through modality gates.

Task-specific policy modules

## Stage 1: Distillation - stabilizes KT

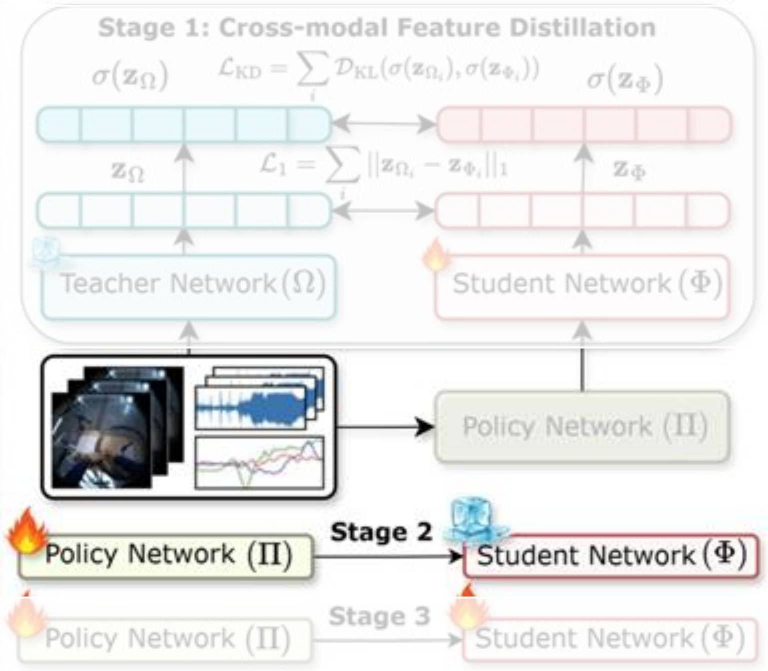



Task specific student model is distilled from the heavy teacher model. The distillation objective optimizes three loss functions $\mathcal{L}$, $\mathcal{L}_{KD}$, $\mathcal{L}_{GT}$

Trainable

Frozen

# Stage 2: Policy learning with frozen student - isolates policy gradients, avoiding feature drift

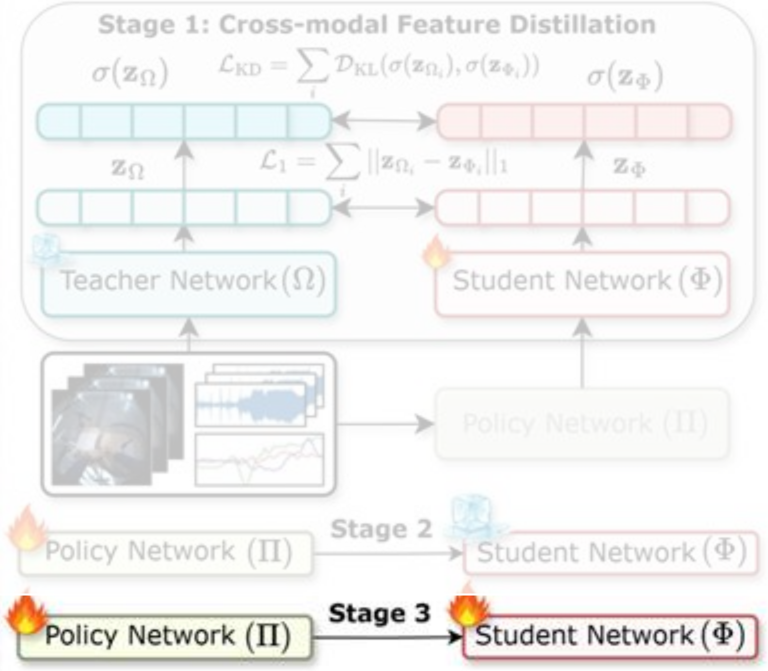



In this stage the policy module is trained keeping the distilled student model fixed. This results in stable training and quicker convergence!

Trainable

Frozen

# Stage 3: Policy and distillation update - $\mathcal{L}_{\Theta} = \eta_1 \mathcal{L}_{\Pi} + \eta_2 \mathcal{L}_{\Phi}$



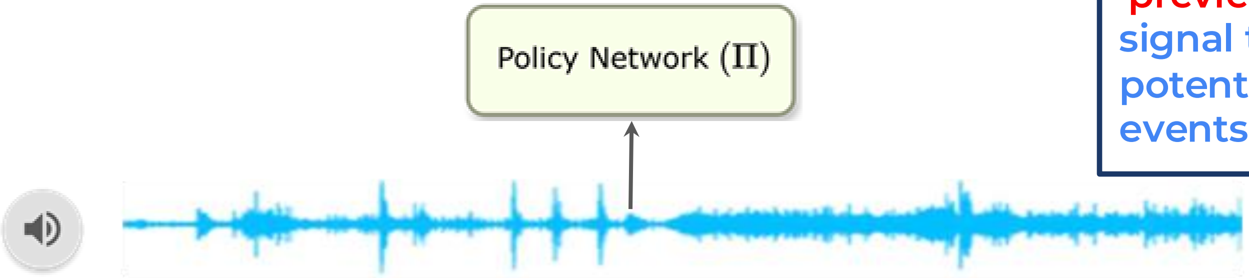

In this stage both the policy module and the distilled student is trained in tandem. This is the final training stage!

# Egocentric Action Recognition Results

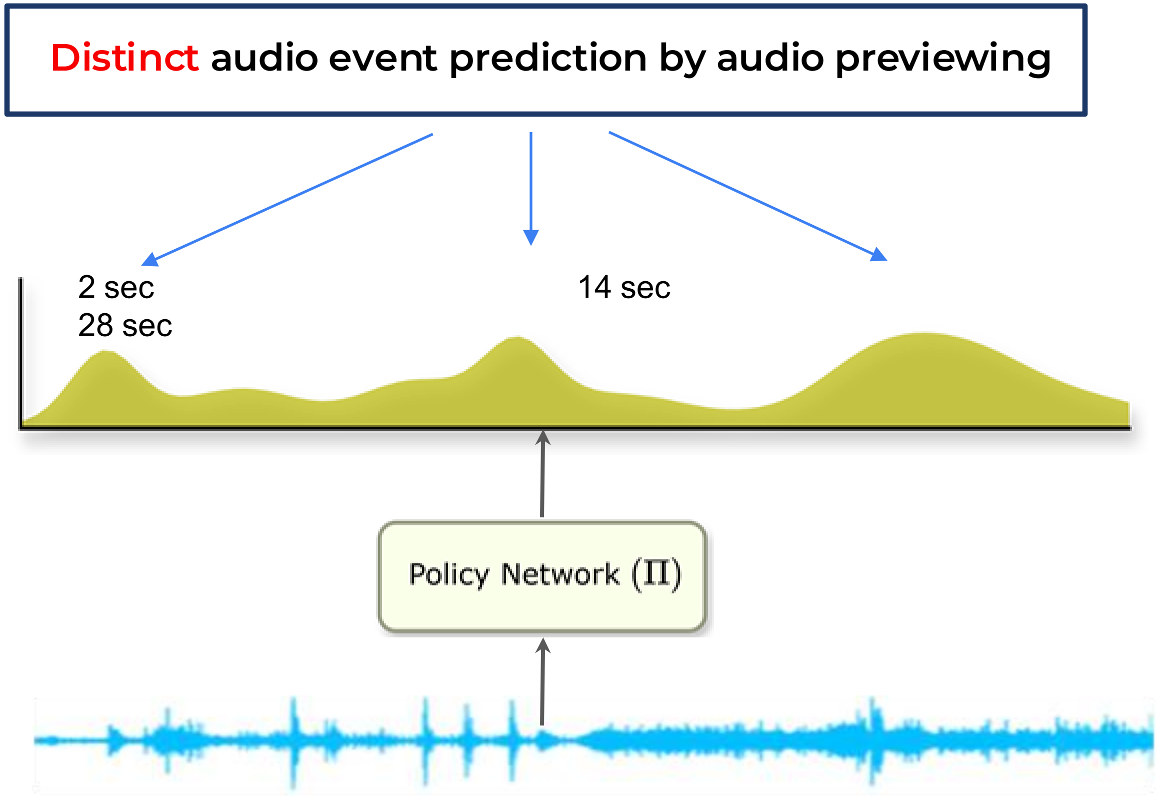
EPIC-Kitchens 100 example

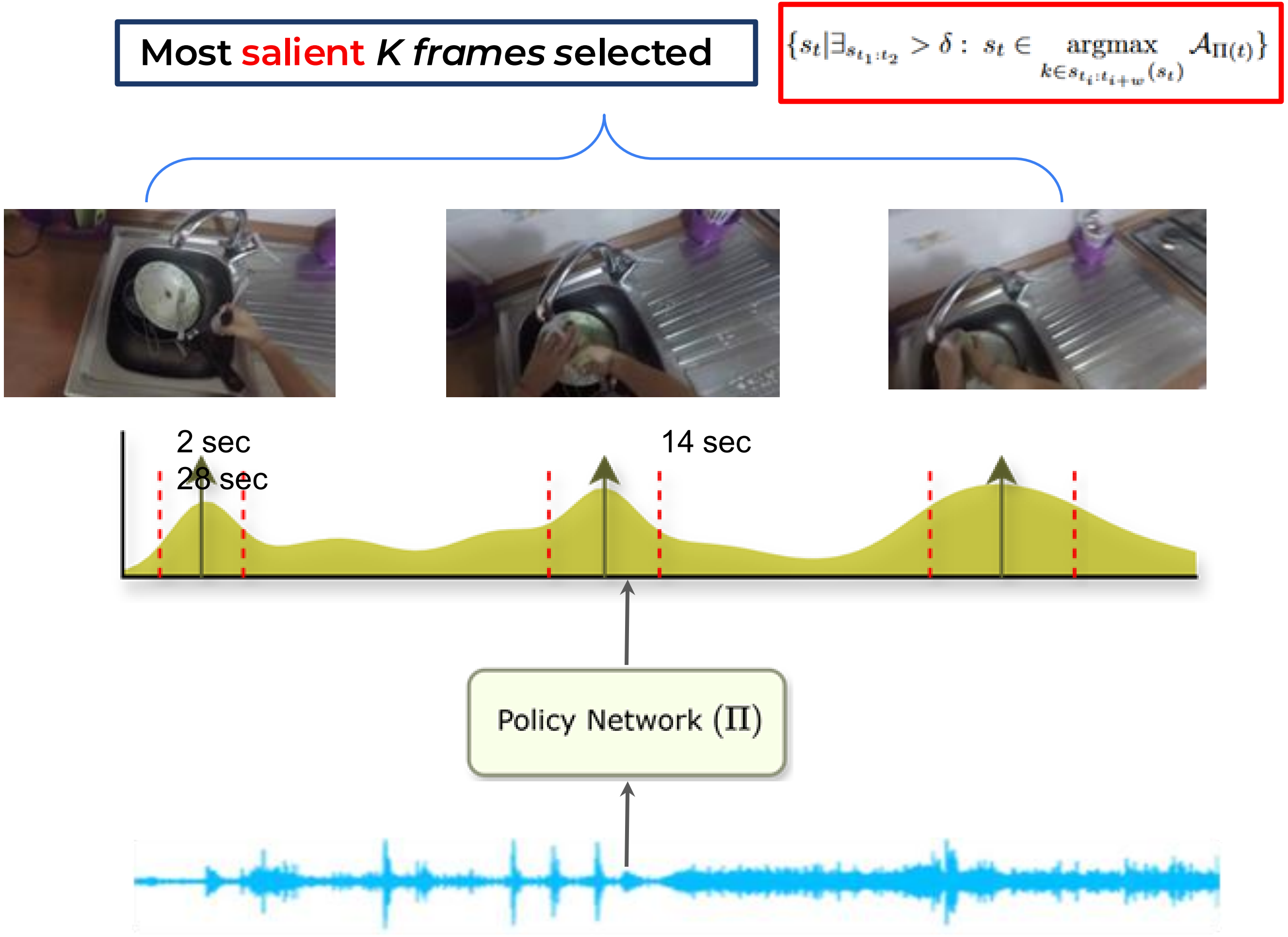
# Audio Preview

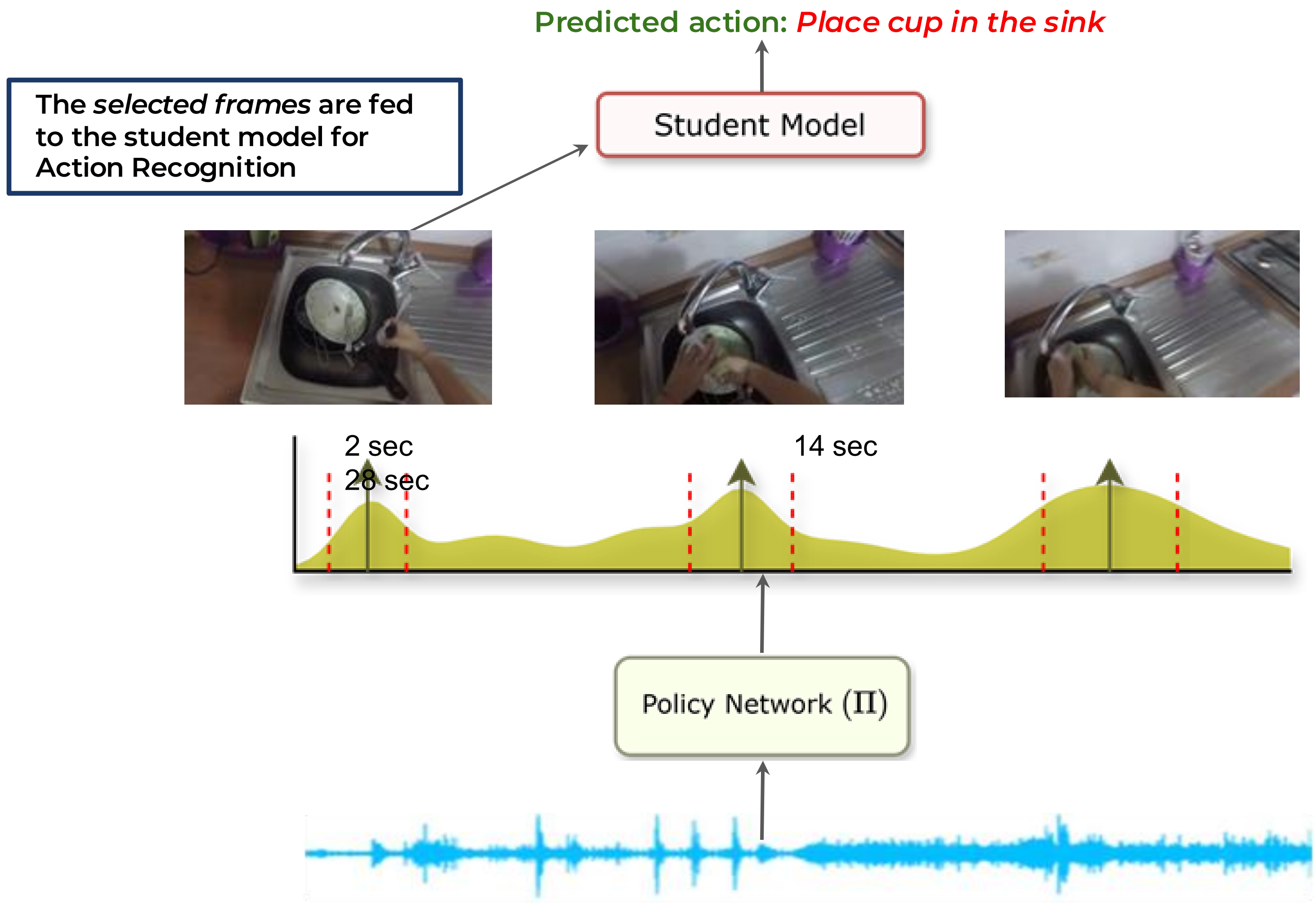
A person doing kitchen chores. Opens the tap, washes dishes, cleans counter.

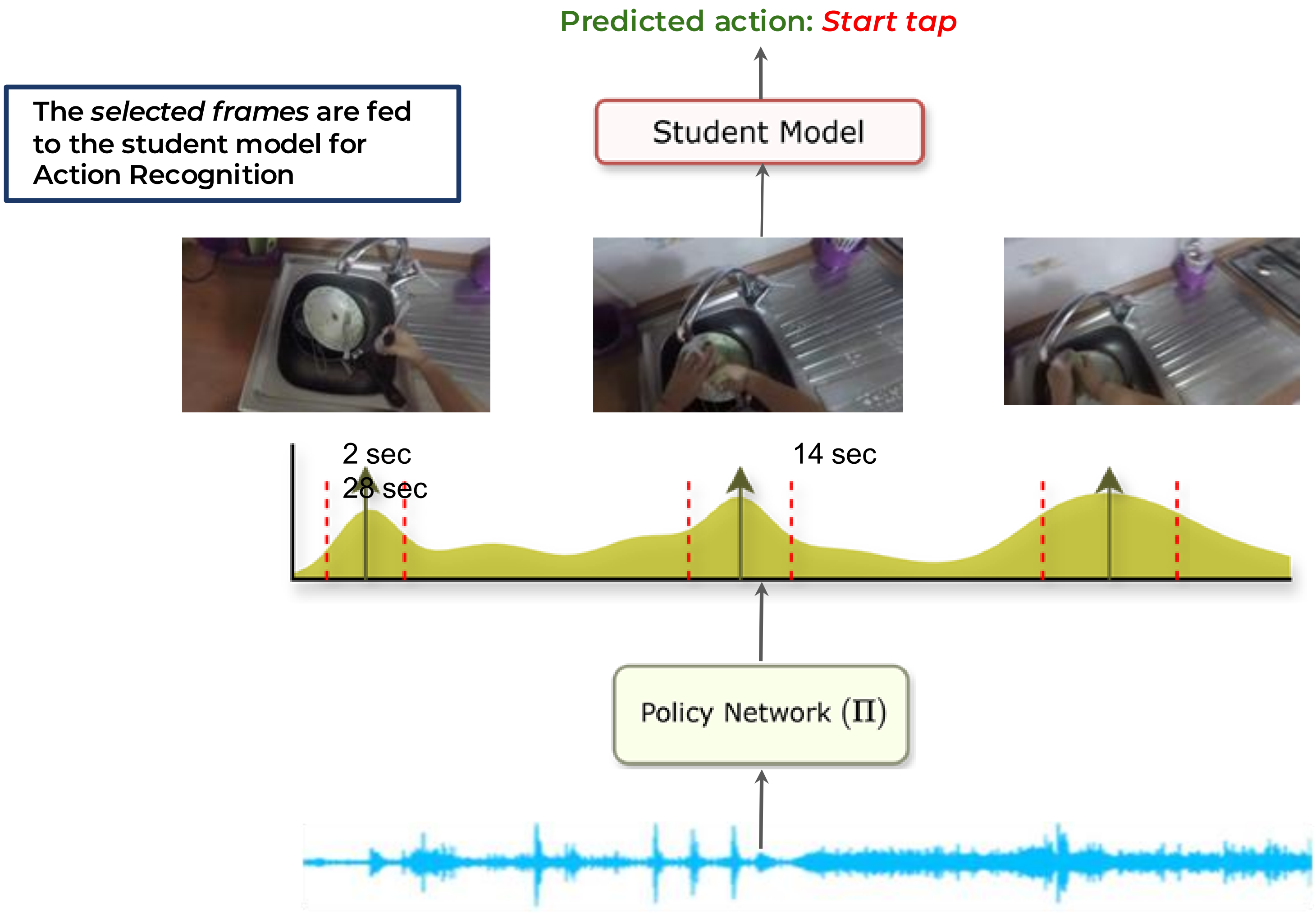



Only the audio is passed to the policy network. Which 'previews' the audio signal to identify potentially distinct events in the video!

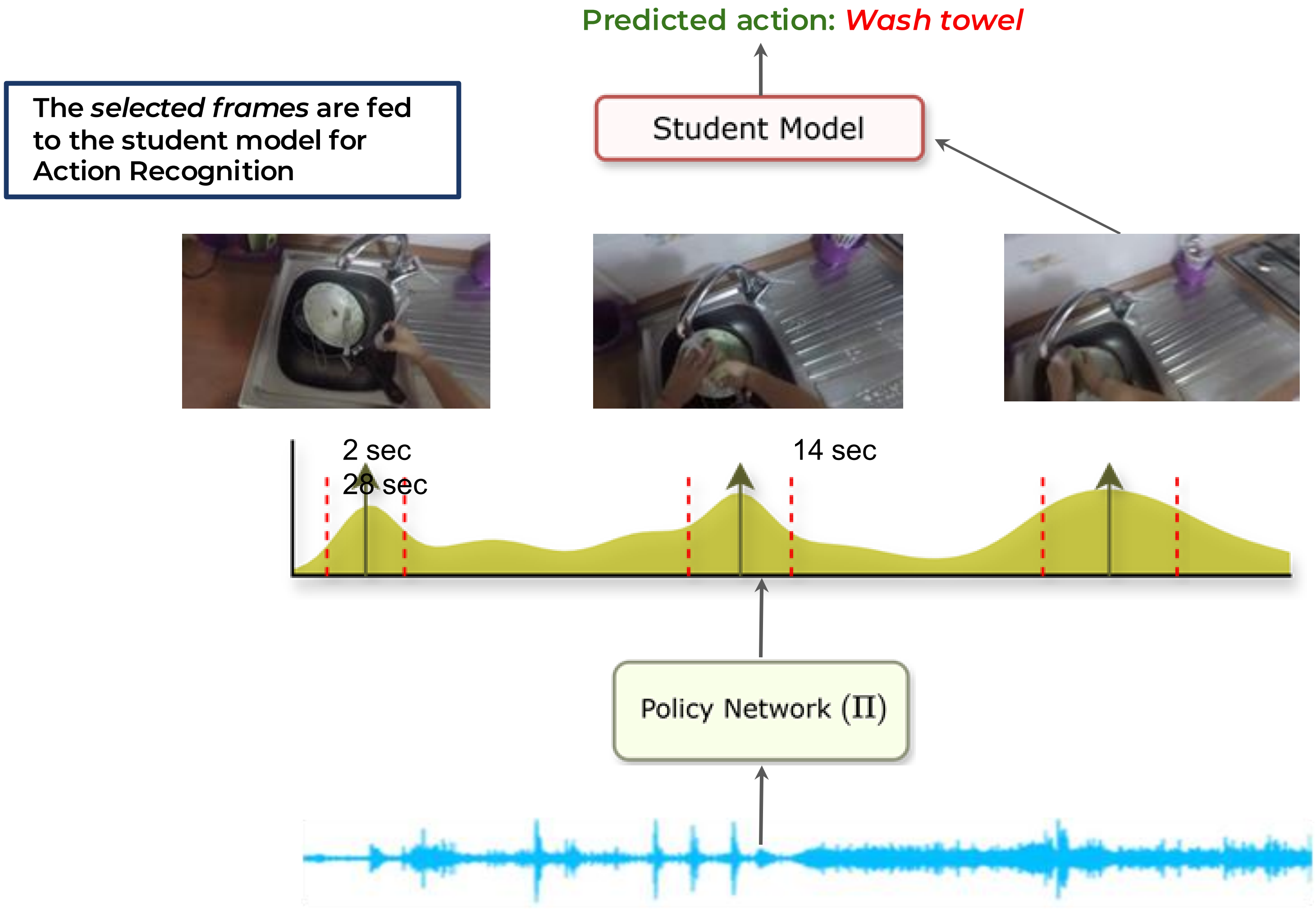

Distinct audio event prediction by audio previewing
2 sec
28 sec
14 sec
Policy Network (Π)


Most salient K frames selected
$\{s_t | \exists_{s_{t_1:t_2}} > \delta : s_t \in \underset{k \in s_{t_i:t_{i+w}}(s_t)}{\operatorname{argmax}} \mathcal{A}_{\Pi(t)}\}$
2 sec
28 sec
14 sec
Policy Network (Π)


Predicted action: Place cup in the sink
The selected frames are fed to the student model for Action Recognition
Student Model
2 sec
28 sec
14 sec
Policy Network (Π)



Predicted action: Start tap
The selected frames are fed to the student model for Action Recognition
Student Model
2 sec
28 sec
14 sec
Policy Network (Π)


Predicted action: *Wash towel*
The *selected frames* are fed to the student model for Action Recognition
Student Model
2 sec
28 sec
14 sec
Policy Network (Π)

# Active Speaker Localization Results

# Frame-wise Comparison Results


Non-active speaker

Active speaker

Model prediction

# Results: Performance

| Method | Input resolution ↓ | Verb↑ | Noun↑ | Action↑ | GMACs↓ |
|---|---|---|---|---|---|
| MoViNet-A6 [37] | 320 × 320 | 72.24 | 57.31 | 47.79 | 79.35 |
| TBN [36] | 224 × 224 | 66.03 | 47.24 | 36.72 | 75.73 |
| AdaFuse [48] | 224 × 224 | 65.52 | 55.75 | 50.16 | 95.84 |
| Ego-only [68] | 224 × 224 | 73.33 | 59.48 | 52.59 | 507.39 |
| ListenToLook [18] | 224 × 224 | 61.27 | 52.52 | 39.85 | 380.46 |
| AdaMML [50] | 224 × 224 | 64.95 | 55.27 | 41.73 | 277.76 |
| VS-VIO [78] | 224 × 224 | 61.37 | 52.21 | 38.07 | 106.97 |
| TIM AV [5] | 224 × 224 | **77.19** | **67.22** | **57.57** | 26.62 |
| EGOADAPT w/o TeMPLe | 224 × 224 | 68.34 | 59.02 | 50.88 | **5.79** |
| **EGOADAPT** | **224 × 224** | 76.65 | 66.83 | 56.74 | **7.14** |

**Table 1. Egocentric action recognition performance of baselines and other other SOTA on EPIC-Kitchens.** We report the top-1 accuracy for verb, noun, and action (%).

| Method | mAP↑ | GMACs↓ | Params (M)↓ | Energy (J)↓ |
|---|---|---|---|---|
| MAVSL$_{C+R}$[35] | 86.32 | 6.852 | 16.13 | 0.698 |
| LocoNet [72] | 71.83 | 3.364 | 34.30 | 1.104 |
| Sync-TalkNet [77] | 65.86 | 3.788 | 32.91 | 0.985 |
| ASD-Trans [9] | 70.13 | 3.621 | 15.03 | 0.482 |
| LW-ASD [42] | 71.60 | 1.280 | 5.36 | 0.145 |
| ListenToLook [18] | 71.28 | 12.452 | 17.34 | 1.032 |
| AdaMML [50] | 76.90 | 10.681 | 13.61 | 0.913 |
| VS-VIO [78] | 72.31 | 7.873 | 5.97 | 0.266 |
| MUST [80] | 89.88 | 0.642 | 2.17 | 0.029 |
| EGOADAPT w/o TeMPLe | 78.59 | 0.077 | 0.36 | 0.003 |
| EGOADAPT | 89.74 | 0.070 | 0.39 | 0.003 |

**Table 2. Performance of active speaker localization on EasyCom.** We compare the mAPs (in %) of various baselines in the visual field of view. Most of these tests use 4-channel audio. EGOADAPT can dynamically choose optimal number of channels.

| Method | Gaze | | | Orientation | | | Trajectory | | | Energy (J) ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $T_{300\text{ ms}}$ | $T_{500\text{ ms}}$ | $T_{700\text{ ms}}$ | $T_{300\text{ ms}}$ | $T_{500\text{ ms}}$ | $T_{700\text{ ms}}$ | $T_{300\text{ ms}}$ | $T_{500\text{ ms}}$ | $T_{700\text{ ms}}$ | |
| MultitaskGP [20] | 11.42 | 15.59 | 18.40 | 4.70 | 9.28 | 12.27 | 13.75 | 17.86 | 20.02 | 0.056 |
| GazeMLE [40] | 10.74 | 14.37 | 18.14 | 4.68 | 9.11 | 12.03 | 14.33 | 16.02 | 18.64 | 1.371 |
| GLC [39] | 10.21 | 14.66 | 17.80 | 4.76 | 8.98 | 11.70 | 13.15 | 15.39 | 17.41 | 0.972 |
| ListenToLook [18] | 13.68 | 17.24 | 19.02 | 5.47 | 8.92 | 11.36 | 13.54 | 15.11 | 17.02 | 0.512 |
| AdaMML [50] | 12.16 | 16.70 | 18.31 | 5.41 | 8.76 | 11.24 | 13.27 | 14.10 | 16.28 | 0.296 |
| VS-VIO [78] | 14.83 | 19.27 | 20.54 | 8.41 | 12.44 | 13.19 | 15.71 | 16.92 | 18.53 | 0.097 |
| MuST$_{AVB}$[80] | 9.17 | 12.15 | 14.75 | 4.78 | **7.36** | **9.90** | 9.96 | 12.38 | 13.95 | 0.029 |
| EGOADAPT w/o TeMPLe | 10.95 | 14.69 | 16.18 | 5.20 | 7.88 | 10.81 | 11.50 | 13.66 | 12.98 | 0.003 |
| EGOADAPT | **8.53** | **11.93** | **14.58** | **4.61** | 7.39 | 9.91 | **9.58** | **11.97** | **13.36** | **0.003** |

**Table 3. Comparison of behavior anticipation errors on the AEA Dataset.** The energy values (in J) are reported by aggregating over three time windows ($T_{300\text{ ms}}$, $T_{500\text{ ms}}$, and $T_{700\text{ ms}}$).

# Results: Performance

| Method | Input resolution ↓ | Verb↑ | Noun↑ | Action↑ | GMACs↓ |
|---|---|---|---|---|---|
| MoViNet-A6 [37] | 320 × 320 | 72.24 | 57.31 | 47.79 | 79.35 |
| TBN [36] | 224 × 224 | 66.03 | 47.24 | 36.72 | 75.73 |
| AdaFuse [48] | 224 × 224 | 65.52 | 55.75 | 50.16 | 95.84 |
| Ego-only [68] | 224 × 224 | 73.33 | 59.48 | 52.59 | 507.39 |
| ListenToLook [18] | 224 × 224 | 61.27 | 52.52 | 39.85 | 380.46 |
| AdaMML [50] | 224 × 224 | 64.95 | 55.27 | 41.73 | 277.76 |
| VS-VIO [78] | 224 × 224 | 61.37 | 52.21 | 38.07 | 106.97 |
| TIM AV [5] | 224 × 224 | **77.19** | **67.22** | **57.57** | 26.62 |
| EGOADAPT w/o TeMPLe | 224 × 224 | 68.34 | 59.02 | 50.88 | **5.79** |
| EGOADAPT | 224 × 224 | 76.65 | 66.83 | 56.74 | 7.14 |

Table 1. **Egocentric action recognition performance of baselines and other other SOTA on EPIC-Kitchens.** We report the top-1 accuracy for verb, noun, and action (%).

| Method | mAP↑ | GMACs↓ | Params (M)↓ | Energy (J)↓ |
|---|---|---|---|---|
| MAVSL$_{C+R}$[35] | 86.32 | 6.852 | 16.13 | 0.698 |
| LocoNet [72] | 71.83 | 3.364 | 34.30 | 1.104 |
| Sync-TalkNet [77] | 65.86 | 3.788 | 32.91 | 0.985 |
| ASD-Trans [9] | 70.13 | 3.621 | 15.03 | 0.482 |
| LW-ASD [42] | 71.60 | 1.280 | 5.36 | 0.145 |
| ListenToLook [18] | 71.28 | 12.452 | 17.34 | 1.032 |
| AdaMML [50] | 76.90 | 10.681 | 13.61 | 0.913 |
| VS-VIO [78] | 72.31 | 7.873 | 5.97 | 0.266 |
| MUST [80] | **89.88** | 0.642 | 2.17 | 0.029 |
| EGOADAPT w/o TeMPLe | 78.59 | 0.077 | 0.36 | 0.003 |
| EGOADAPT | 89.74 | **0.070** | **0.39** | **0.003** |

Table 2. **Performance of active speaker localization on Easy-Com.** We compare the mAPs (in %) of various baselines in the visual field of view. Most of these tests use 4-channel audio. EGOADAPT can dynamically choose optimal number of channels.

| Method | Gaze | | | Orientation | | | Trajectory | | | Energy (J) ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $T_{300\ \text{ms}}$ | $T_{500\ \text{ms}}$ | $T_{700\ \text{ms}}$ | $T_{300\ \text{ms}}$ | $T_{500\ \text{ms}}$ | $T_{700\ \text{ms}}$ | $T_{300\ \text{ms}}$ | $T_{500\ \text{ms}}$ | $T_{700\ \text{ms}}$ | |
| MultitaskGP [20] | 11.42 | 15.59 | 18.40 | 4.70 | 9.28 | 12.27 | 13.75 | 17.86 | 20.02 | 0.056 |
| GazeMLE [40] | 10.74 | 14.37 | 18.14 | 4.68 | 9.11 | 12.03 | 14.33 | 16.02 | 18.64 | 1.371 |
| GLC [39] | 10.21 | 14.66 | 17.80 | 4.76 | 8.98 | 11.70 | 13.15 | 15.39 | 17.41 | 0.972 |
| ListenToLook [18] | 13.68 | 17.24 | 19.02 | 5.47 | 8.92 | 11.36 | 13.54 | 15.11 | 17.02 | 0.512 |
| AdaMML [50] | 12.16 | 16.70 | 18.31 | 5.41 | 8.76 | 11.24 | 13.27 | 14.10 | 16.28 | 0.296 |
| VS-VIO [78] | 14.83 | 19.27 | 20.54 | 8.41 | 12.44 | 13.19 | 15.71 | 16.92 | 18.53 | 0.097 |
| MuST$_{AVB}$[80] | 9.17 | 12.15 | 14.75 | 4.78 | 7.36 | 9.90 | 9.96 | 12.38 | 13.95 | 0.029 |
| EGOADAPT w/o TeMPLe | 10.95 | 14.69 | 16.18 | 5.20 | 7.88 | 10.81 | 11.50 | 13.66 | 12.98 | 0.003 |
| EGOADAPT | **8.53** | **11.93** | **14.58** | **4.61** | 7.39 | 9.91 | **9.58** | **11.97** | **13.36** | **0.003** |

Table 3. **Comparison of behavior anticipation errors on the AEA Dataset.** The energy values (in J) are reported by aggregating over three time windows ($T_{300\ \text{ms}}$, $T_{500\ \text{ms}}$, and $T_{700\ \text{ms}}$).

# Results: Performance

| Method | Input resolution ↓ | Verb↑ | Noun↑ | Action↑ | GMACs↓ |
|---|---|---|---|---|---|
| MoViNet-A6 [37] | 320 × 320 | 72.24 | 57.31 | 47.79 | 79.35 |
| TBN [36] | 224 × 224 | 66.03 | 47.24 | 36.72 | 75.73 |
| AdaFuse [48] | 224 × 224 | 65.52 | 55.75 | 50.16 | 95.84 |
| Ego-only [68] | 224 × 224 | 73.33 | 59.48 | 52.59 | 507.39 |
| ListenToLook [18] | 224 × 224 | 61.27 | 52.52 | 39.85 | 380.46 |
| AdaMML [50] | 224 × 224 | 64.95 | 55.27 | 41.73 | 277.76 |
| VS-VIO [78] | 224 × 224 | 61.37 | 52.21 | 38.07 | 106.97 |
| TIM AV [5] | 224 × 224 | **77.19** | **67.22** | **57.57** | 26.62 |
| EGOADAPT w/o TeMPLe | 224 × 224 | 68.34 | 59.02 | 50.88 | 5.79 |
| **EGOADAPT** | **224 × 224** | 76.65 | 66.83 | 56.74 | **7.14** |

Table 1. **Egocentric action recognition performance of baselines and other other SOTA on EPIC-Kitchens.** We report the top-1 accuracy for verb, noun, and action (%).

| Method | mAP↑ | GMACs↓ | Params (M)↓ | Energy (J)↓ |
|---|---|---|---|---|
| $\text{MAVSL}_{C+E}$ [35] | 86.32 | 6.852 | 16.13 | 0.698 |
| LocoNet [72] | 71.83 | 3.364 | 34.30 | 1.104 |
| Sync-TalkNet [77] | 65.86 | 3.788 | 32.91 | 0.985 |
| ASD-Trans [9] | 70.13 | 3.621 | 15.03 | 0.482 |
| LW-ASD [42] | 71.60 | 1.280 | 5.36 | 0.145 |
| ListenToLook [18] | 71.28 | 12.452 | 17.34 | 1.032 |
| AdaMML [50] | 76.90 | 10.681 | 13.61 | 0.913 |
| VS-VIO [78] | 72.31 | 7.873 | 5.97 | 0.266 |
| MUST [80] | **89.88** | 0.642 | 2.17 | 0.029 |
| EGOADAPT w/o TeMPLe | 78.59 | 0.077 | 0.36 | 0.003 |
| **EGOADAPT** | 89.74 | **0.070** | **0.39** | **0.003** |

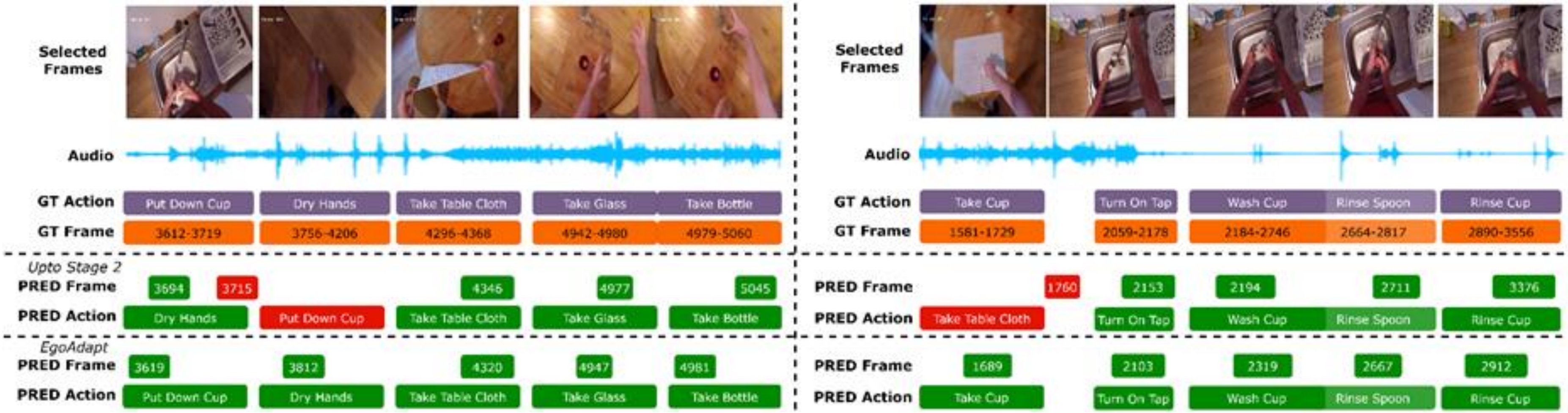
Table 2. **Performance of active speaker localization on Easy-Com.** We compare the mAPs (in %) of various baselines in the visual field of view. Most of these tests use 4-channel audio. EGOADAPT can dynamically choose optimal number of channels.

| Method | Gaze | | | Orientation | | | Trajectory | | | Energy (J) ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $T_{300\text{ ms}}$ | $T_{500\text{ ms}}$ | $T_{700\text{ ms}}$ | $T_{300\text{ ms}}$ | $T_{500\text{ ms}}$ | $T_{700\text{ ms}}$ | $T_{300\text{ ms}}$ | $T_{500\text{ ms}}$ | $T_{700\text{ ms}}$ | |
| MultitaskGP [20] | 11.42 | 15.59 | 18.40 | 4.70 | 9.28 | 12.27 | 13.75 | 17.86 | 20.02 | 0.056 |
| GazeMLE [40] | 10.74 | 14.37 | 18.14 | 4.68 | 9.11 | 12.03 | 14.33 | 16.02 | 18.64 | 1.371 |
| GLC [39] | 10.21 | 14.66 | 17.80 | 4.76 | 8.98 | 11.70 | 13.15 | 15.39 | 17.41 | 0.972 |
| ListenToLook [18] | 13.68 | 17.24 | 19.02 | 5.47 | 8.92 | 11.36 | 13.54 | 15.11 | 17.02 | 0.512 |
| AdaMML [50] | 12.16 | 16.70 | 18.31 | 5.41 | 8.76 | 11.24 | 13.27 | 14.10 | 16.28 | 0.296 |
| VS-VIO [78] | 14.83 | 19.27 | 20.54 | 8.41 | 12.44 | 13.19 | 15.71 | 16.92 | 18.53 | 0.097 |
| $\text{MuST}_{AVB}$[80] | 9.17 | 12.15 | 14.75 | 4.78 | **7.36** | **9.90** | 9.96 | 12.38 | 13.95 | 0.029 |
| EGOADAPT w/o TeMPLe | 10.95 | 14.69 | 16.18 | 5.20 | 7.88 | 10.81 | 11.50 | 13.66 | 12.98 | 0.003 |
| **EGOADAPT** | **8.53** | **11.93** | **14.58** | **4.61** | 7.39 | 9.91 | **9.58** | **11.97** | **13.36** | **0.003** |

Table 3. **Comparison of behavior anticipation errors on the AEA Dataset.** The energy values (in J) are reported by aggregating over three time windows ($T_{300\text{ ms}}$, $T_{500\text{ ms}}$, and $T_{700\text{ ms}}$).

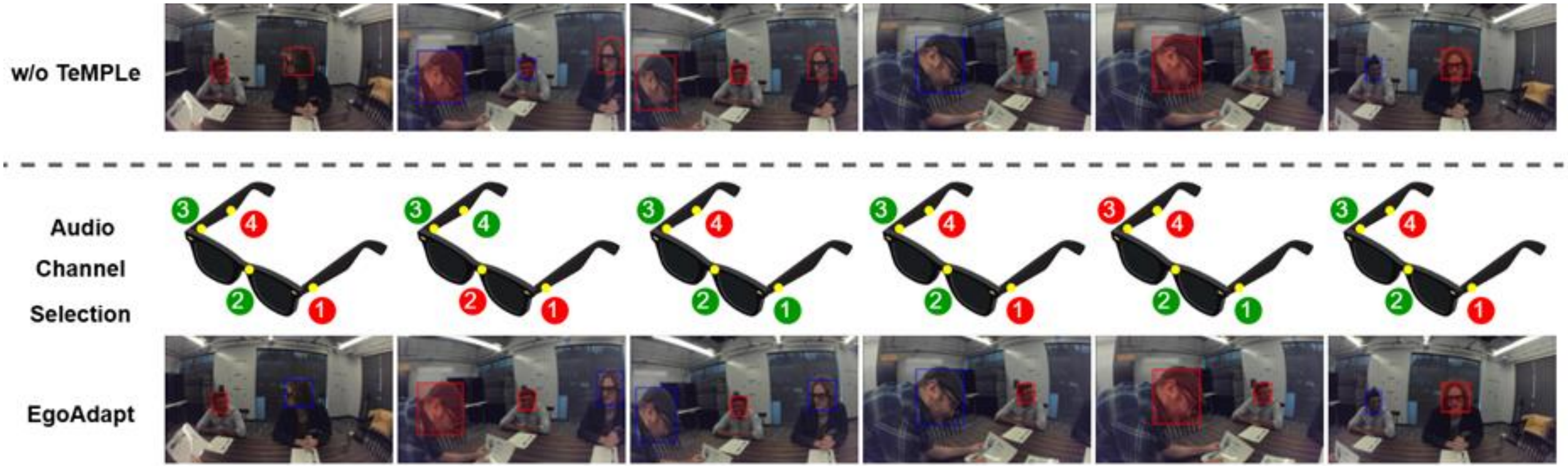
# Ablation Results

# Qualitative Results: Action Recognition



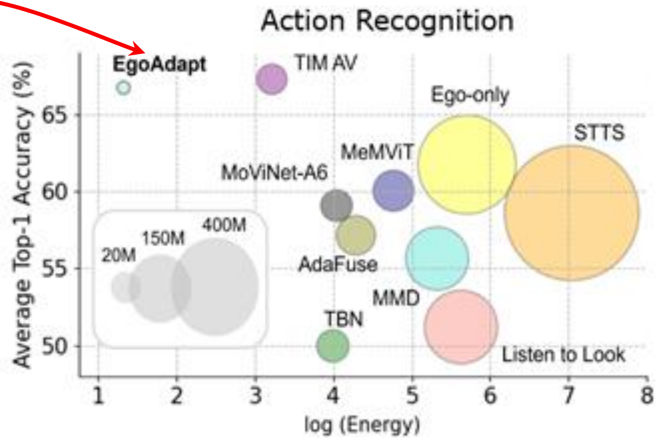

The green and red boxes represent correct and incorrect predictions, respectively. EgoAdapt picks the most informative frame to predict the ‘Noun’ classes, which is subsequently used to predict the action

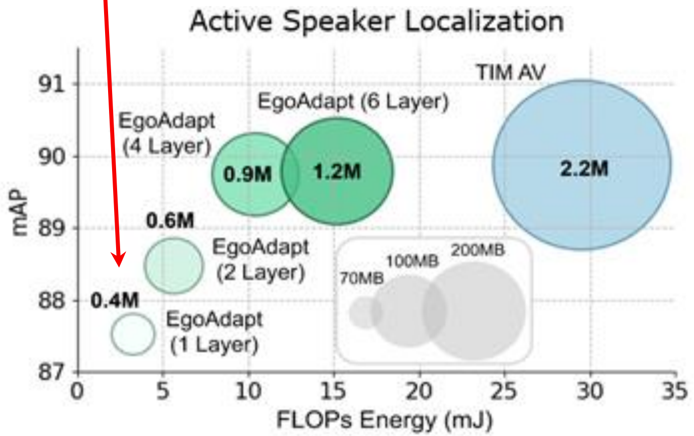
# Qualitative Results: Active Speaker Localization




The red/blue boxes indicate active/non-active speakers, and the red heatmap indicates model prediction. EgoAdapt can make correct predictions for scenes with motion blur (col. 4), partial vision (col. 5), and multi-speakers (col. 2, 5). The red/green circles represent the discarded and selected audio channels.

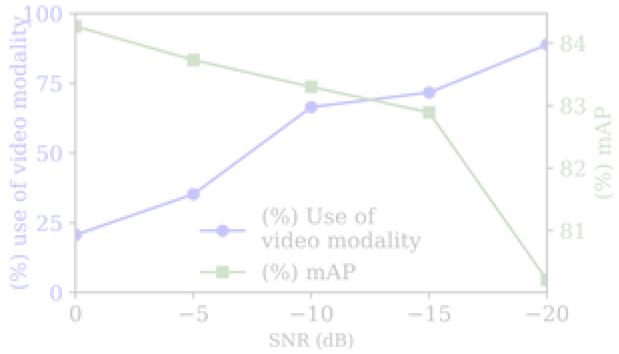
# Results: Efficiency



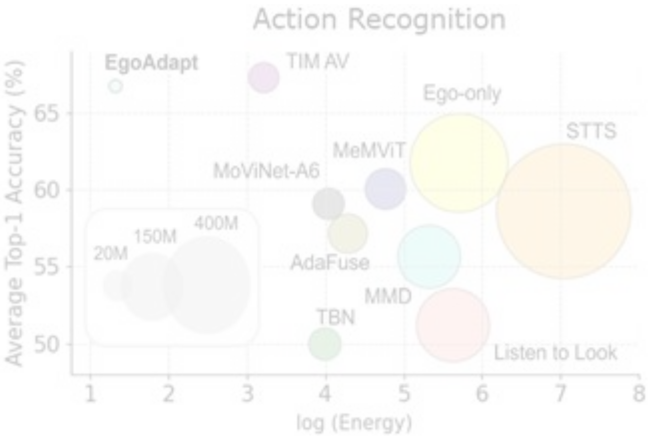



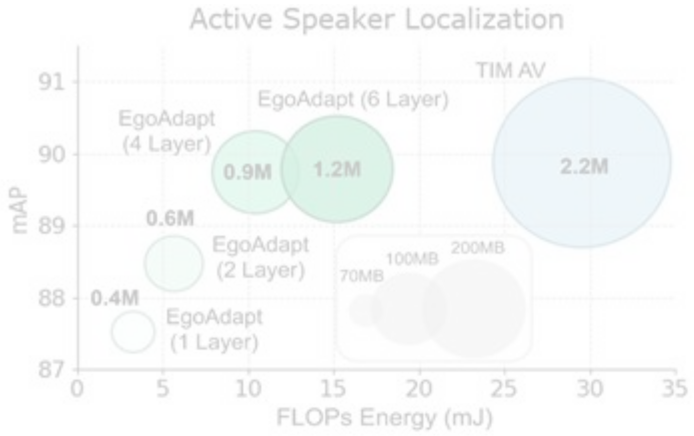

| Precision Level | Modality A | Modality V | mAP ↑ | Power (mW) ↓ | Exec. Time ↓ |
|---|---|---|---|---|---|
| 4 bit | ✓ | ✗ | 77.14 | **7.38** | **0.12** |
| | ✓ | ✓ | 78.92 | 9.94 | 0.21 |
| 8 bit | ✓ | ✗ | 80.56 | 11.37 | 0.33 |
| | ✓ | ✓ | 81.13 | 14.90 | 0.42 |
| 16 bit | ✓ | ✗ | 84.39 | 19.11 | 0.59 |
| | ✓ | ✓ | 85.74 | 23.06 | 0.68 |
| 32 bit | ✓ | ✗ | 83.22 | 29.71 | 0.89 |
| | ✓ | ✓ | **89.74** | 34.39 | 1.00 |

On device implementation



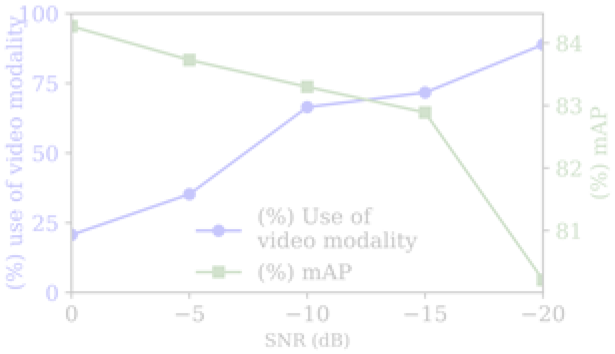
# Results: Efficiency



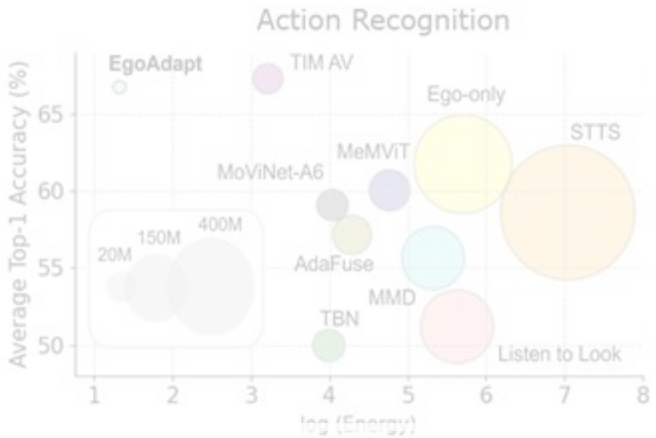



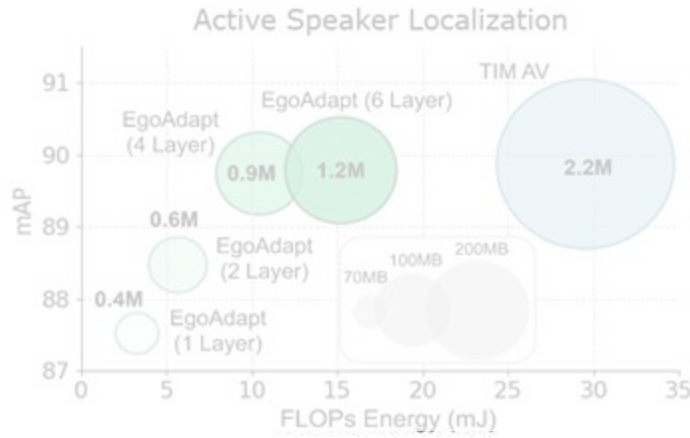

| Precision Level | Modality A | Modality V | mAP ↑ | Power (mW) ↓ | Exec. Time ↓ |
|---|---|---|---|---|---|
| 4 bit | ✓ | ✗ | 77.14 | **7.38** | **0.12** |
| | ✓ | ✓ | 78.92 | 9.94 | 0.21 |
| 8 bit | ✓ | ✗ | 80.56 | 11.37 | 0.33 |
| | ✓ | ✓ | 81.13 | 14.90 | 0.42 |
| 16 bit | ✓ | ✗ | 84.39 | 19.11 | 0.59 |
| | ✓ | ✓ | 85.74 | 23.06 | 0.68 |
| 32 bit | ✓ | ✗ | 83.22 | 29.71 | 0.89 |
| | ✓ | ✓ | **89.74** | 34.39 | 1.00 |

On device implementation



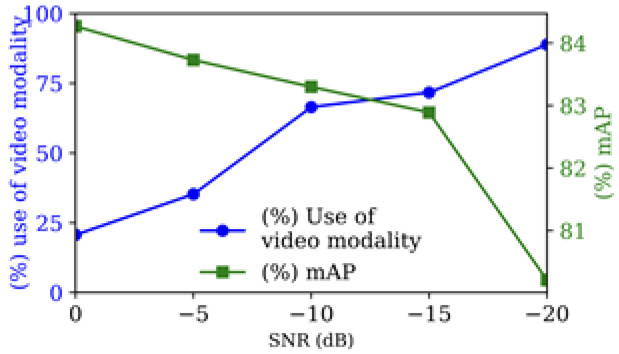
# Results: Efficiency



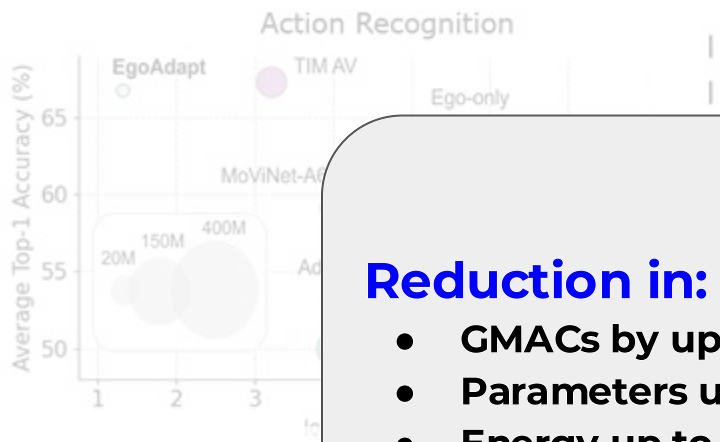



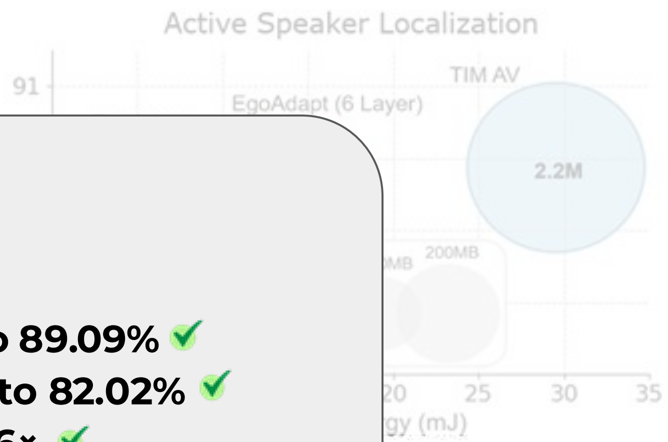

| Precision Level | Modality A | Modality V | mAP ↑ | Power (mW) ↓ | Exec. Time ↓ |
|---|---|---|---|---|---|
| 4 bit | ✓ | ✗ | 77.14 | **7.38** | **0.12** |
| | ✓ | ✓ | 78.92 | 9.94 | 0.21 |
| 8 bit | ✓ | ✗ | 80.56 | 11.37 | 0.33 |
| | ✓ | ✓ | 81.13 | 14.90 | 0.42 |
| 16 bit | ✓ | ✗ | 84.39 | 19.11 | 0.59 |
| | ✓ | ✓ | 85.74 | 23.06 | 0.68 |
| 32 bit | ✓ | ✗ | 83.22 | 29.71 | 0.89 |
| | ✓ | ✓ | **89.74** | 34.39 | 1.00 |

On device implementation



# Results: Efficiency



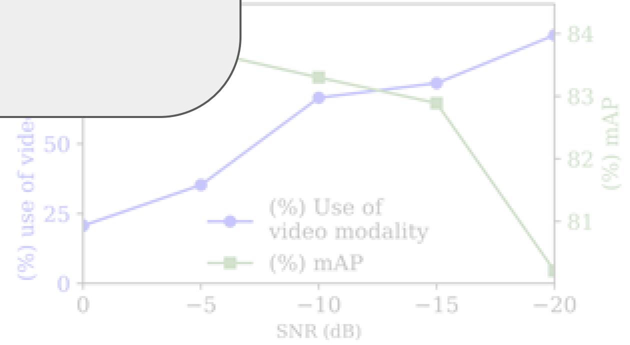




**Reduction in:**

- GMACs by up to 89.09% ✓
- Parameters up to 82.02% ✓
- Energy up to 9.6× ✓

| Precision Level | Modality A | Modality V | mAP | [illegible] | [illegible] |
|---|---|---|---|---|---|
| 4 bit | ✓ | ✗ | 77.1[illegible] | [illegible] | [illegible] |
| | ✓ | ✓ | 78.92 | [illegible] | [illegible] |
| 8 bit | ✓ | ✗ | 80.56 | 11.37 | 0.33 |
| | ✓ | ✓ | 81.13 | 14.90 | 0.42 |
| 16 bit | ✓ | ✗ | 84.39 | 19.11 | 0.59 |
| | ✓ | ✓ | 85.74 | 23.06 | 0.68 |
| 32 bit | ✓ | ✗ | 83.22 | 29.71 | 0.89 |
| | ✓ | ✓ | **89.74** | 34.39 | 1.00 |

On device implementation



**Project Page**

# Questions?

Thank you!