Carnegie Mellon University
The Robotics Institute

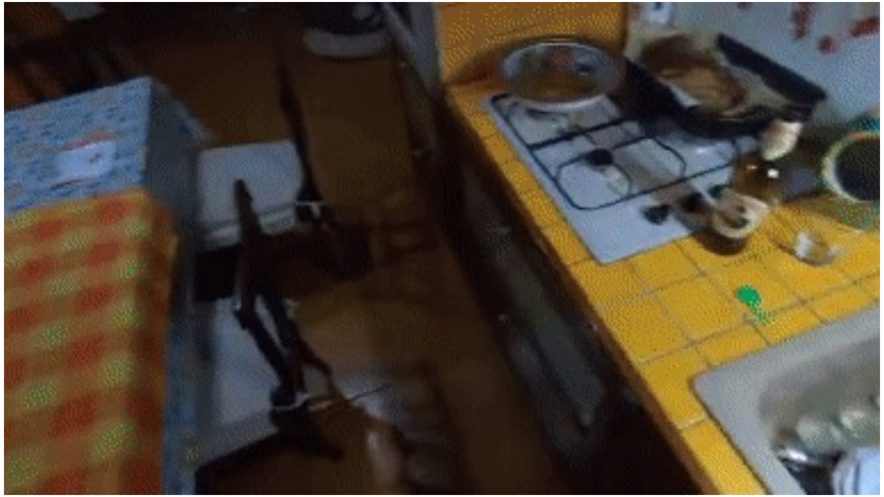
Learning State-Aware Visual Representations from Audible Interactions




Himangi Mittal Pedro Morgado Unnat Jain Abhinav Gupta

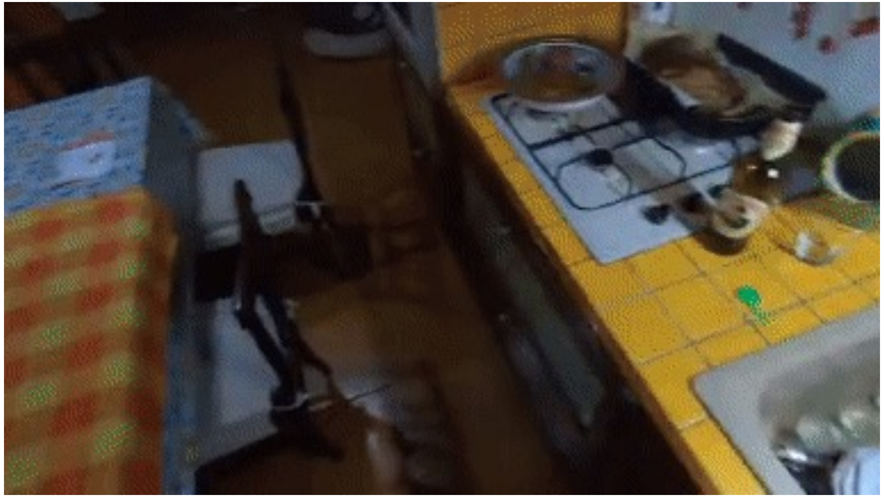
Neural Information Processing Systems (NeurIPS), 2022.

Real-world videos consists of long, untrimmed, egocentric daily activities



Damen, Dima, et al. "Rescaling egocentric vision." *arXiv preprint arXiv:2006.13256* (2020).

Real-world videos consists of long, untrimmed, egocentric daily activities



Learning representations from real-world videos can be quite challenging

Damen, Dima, et al. "Rescaling egocentric vision." *arXiv preprint arXiv:2006.13256* (2020).

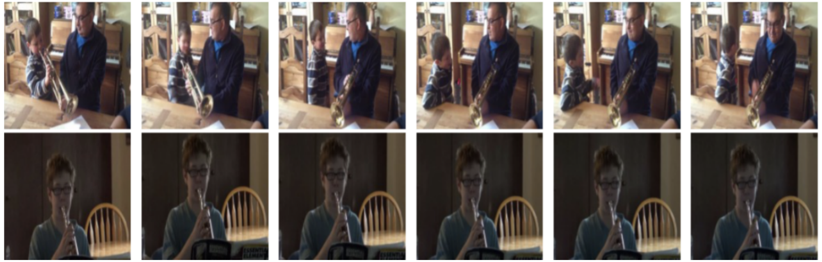
Traditional Pipelines (1)

Work on curated, trimmed videos

Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).

Traditional Pipelines (1)

Work on curated, trimmed videos



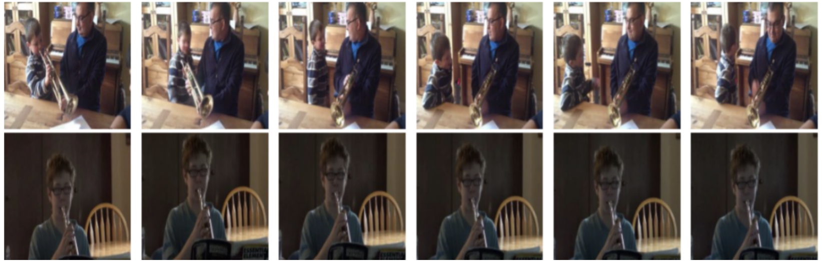
Playing trumpet

Dribbling basketball

Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).

Traditional Pipelines (1)

Work on curated, trimmed videos



Playing trumpet

Dribbling basketball

Long, untrimmed, real-world videos consists of multiple actions as well as **no-activity segments**

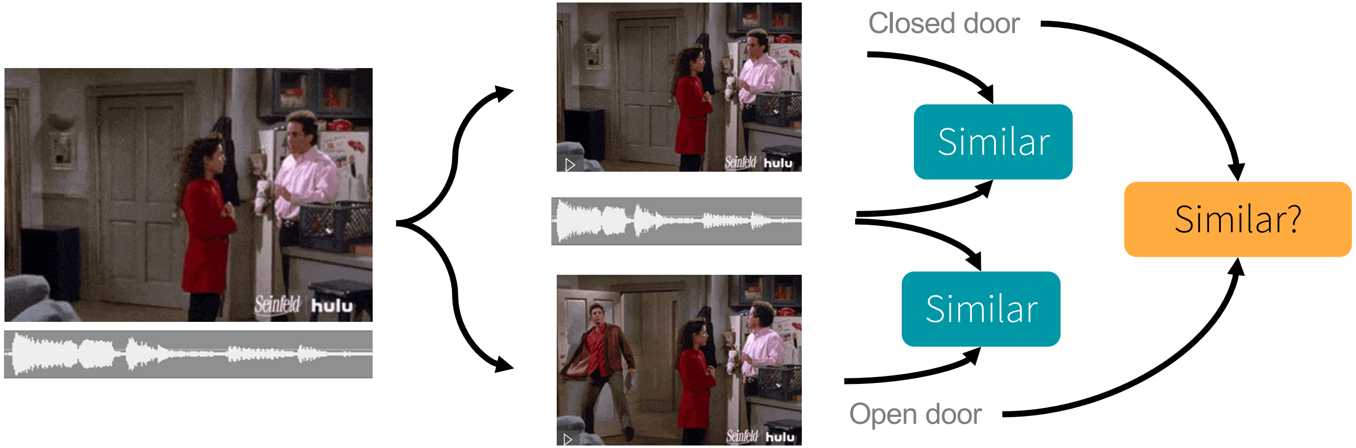
Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).

Traditional Pipelines (2)

Morgado, Pedro, Nuno Vasconcelos, and Ishan Misra. "Audio-visual instance discrimination with cross-modal agreement." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Traditional Pipelines (2)

Invariant to state-changes in the environment



Morgado, Pedro, Nuno Vasconcelos, and Ishan Misra. "Audio-visual instance discrimination with cross-modal agreement." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Can we learn meaningful representations from interaction-rich, untrimmed, and multi-modal egocentric data?

Two key components/contributions:

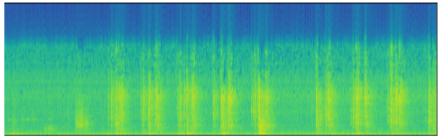
Focus learning on moments of interaction

Learning from audible state changes

Video

Cut celery

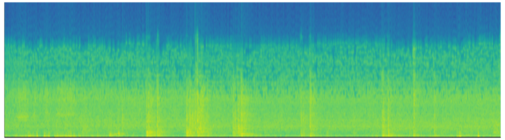
Spectrogram



Video

Peel onion

Spectrogram

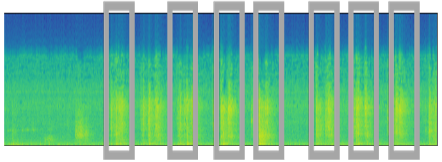


Damen, Dima, et al. "Rescaling egocentric vision." *arXiv preprint arXiv:2006.13256* (2020).

Video

Cut celery

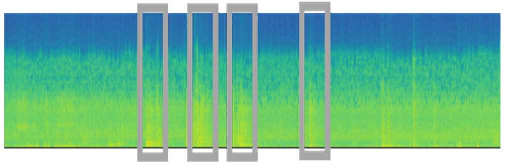
Spectrogram



Video

Peel onion

Spectrogram

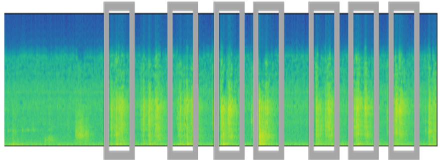


Damen, Dima, et al. "Rescaling egocentric vision." *arXiv preprint arXiv:2006.13256* (2020).

Video

Cut celery

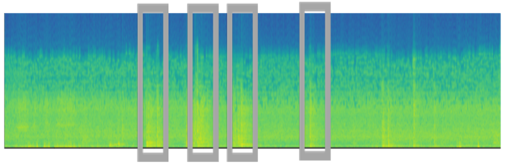
Spectrogram



Video

Peel onion

Spectrogram



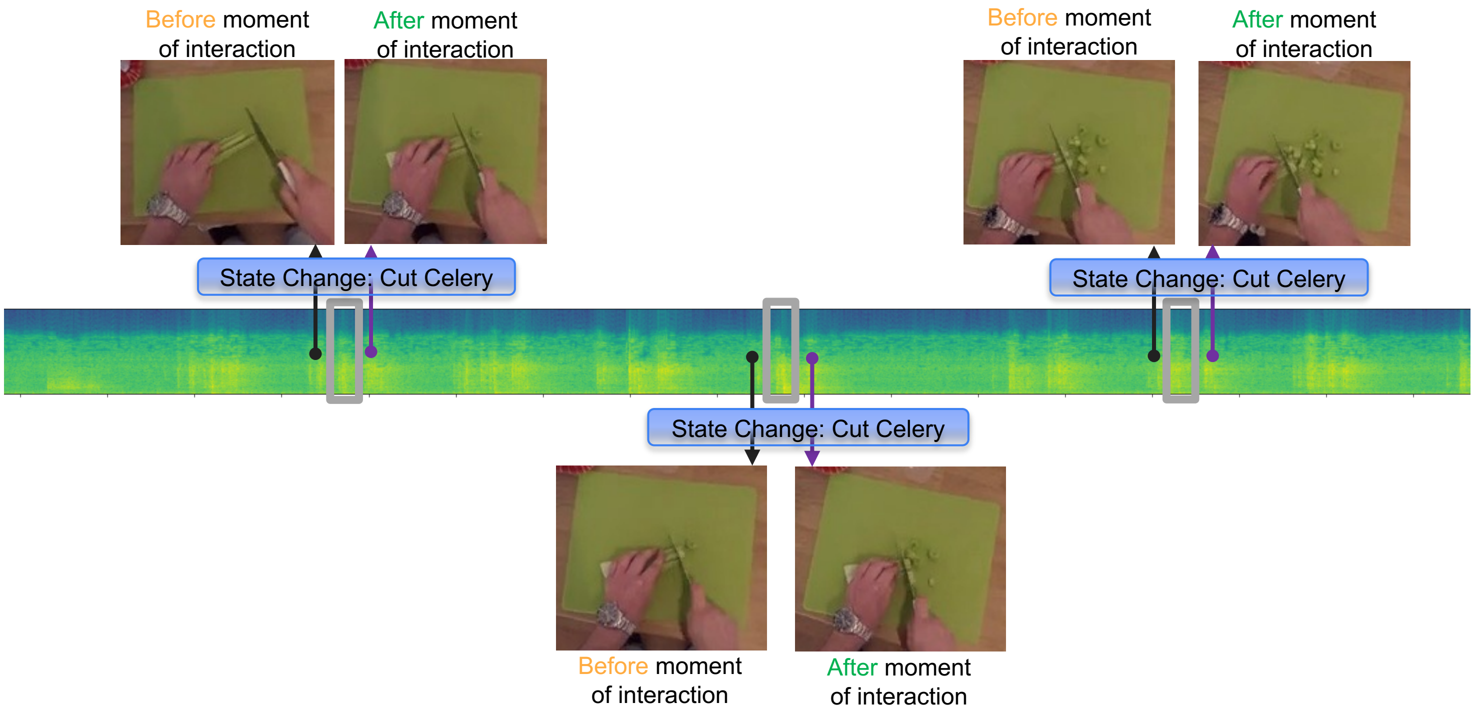
We refer these as **Moments of Interaction (MoI)**

Damen, Dima, et al. "Rescaling egocentric vision." *arXiv preprint arXiv:2006.13256* (2020).

Two key components/contributions:

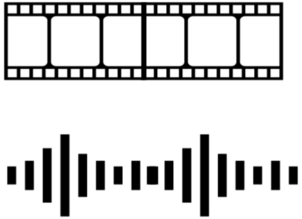
Focus learning on moments of interaction

Learning from audible state changes



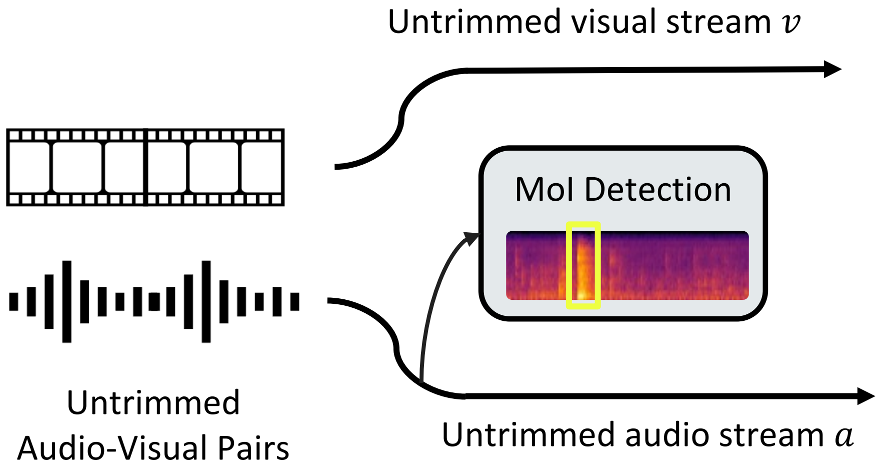
Transition between object states is often marked by characteristic sounds

Method

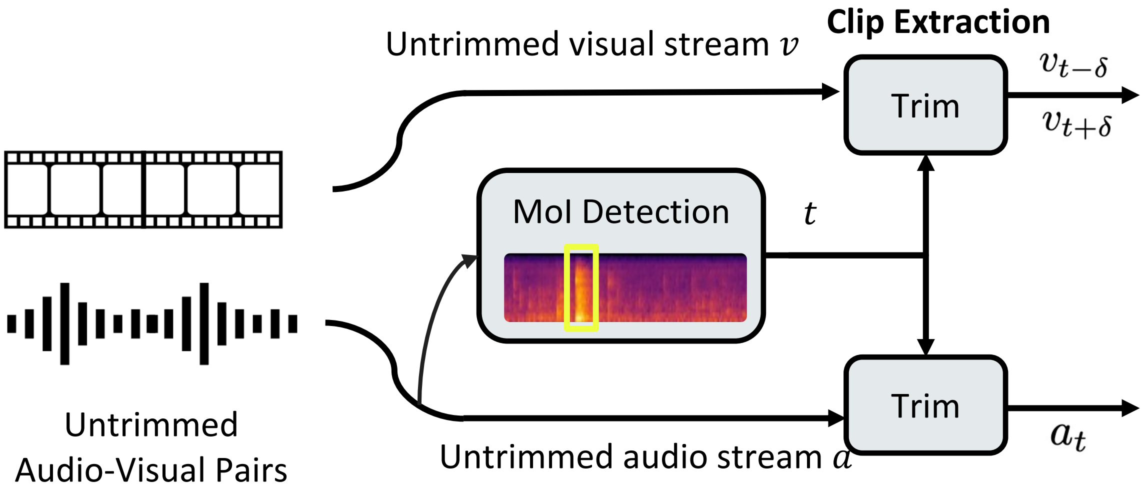


Untrimmed
Audio-Visual Pairs

Method

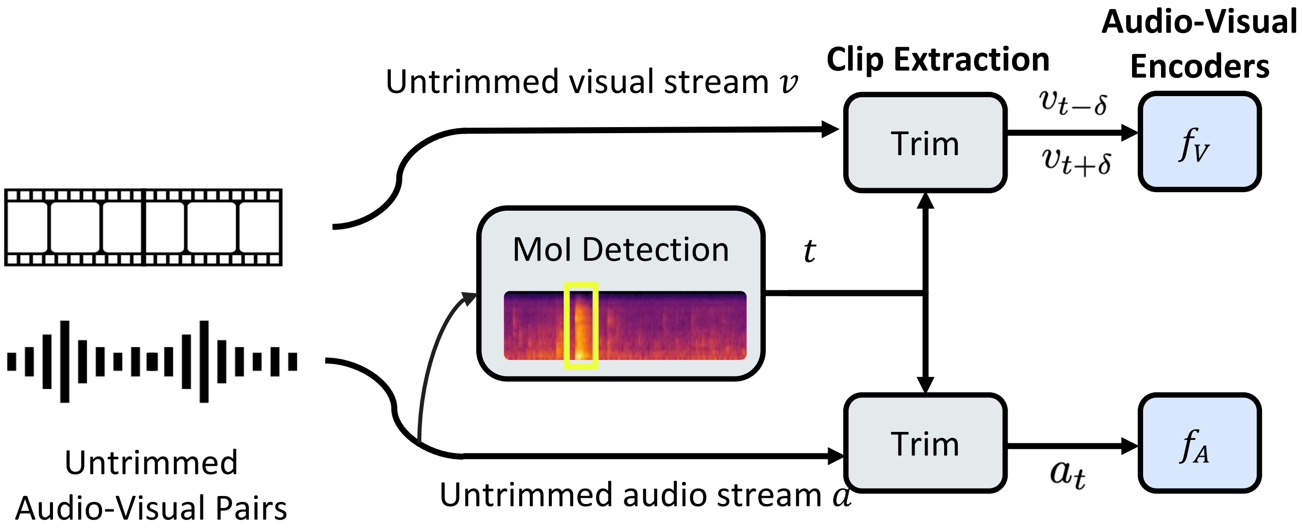

Untrimmed visual stream v
MoI Detection
Untrimmed
Audio-Visual Pairs
Untrimmed audio stream a

Method



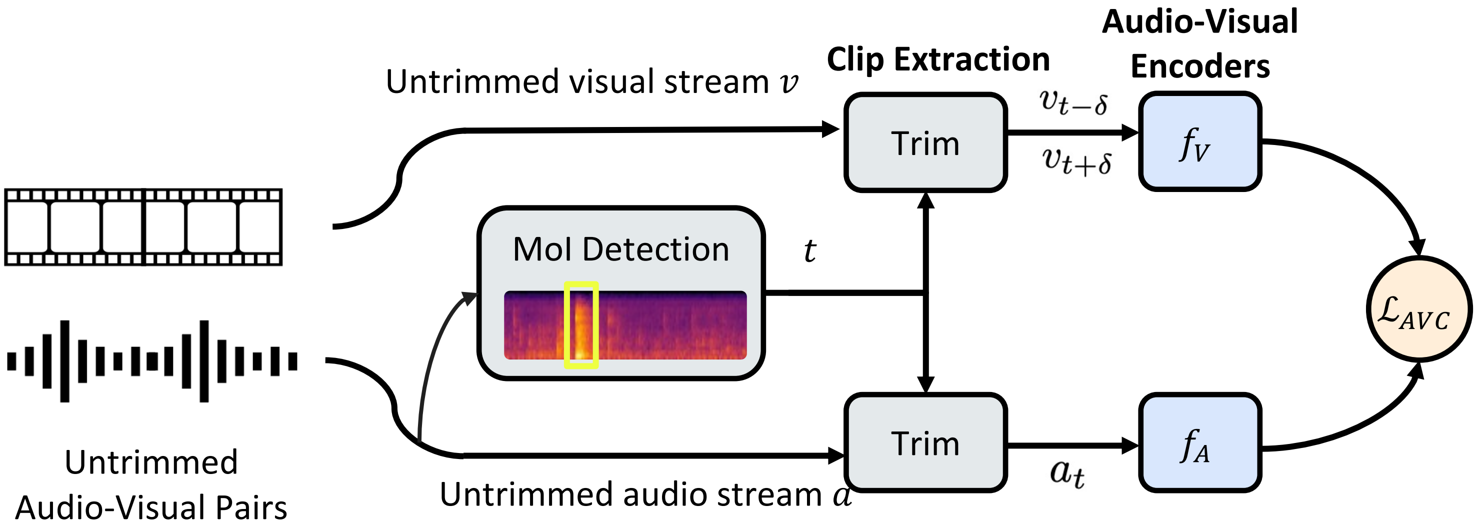
Clip Extraction
Untrimmed visual stream v
Trim
$v_{t-\delta}$
$v_{t+\delta}$
MoI Detection
t
Trim
a_t
Untrimmed audio stream a
Untrimmed
Audio-Visual Pairs

Method

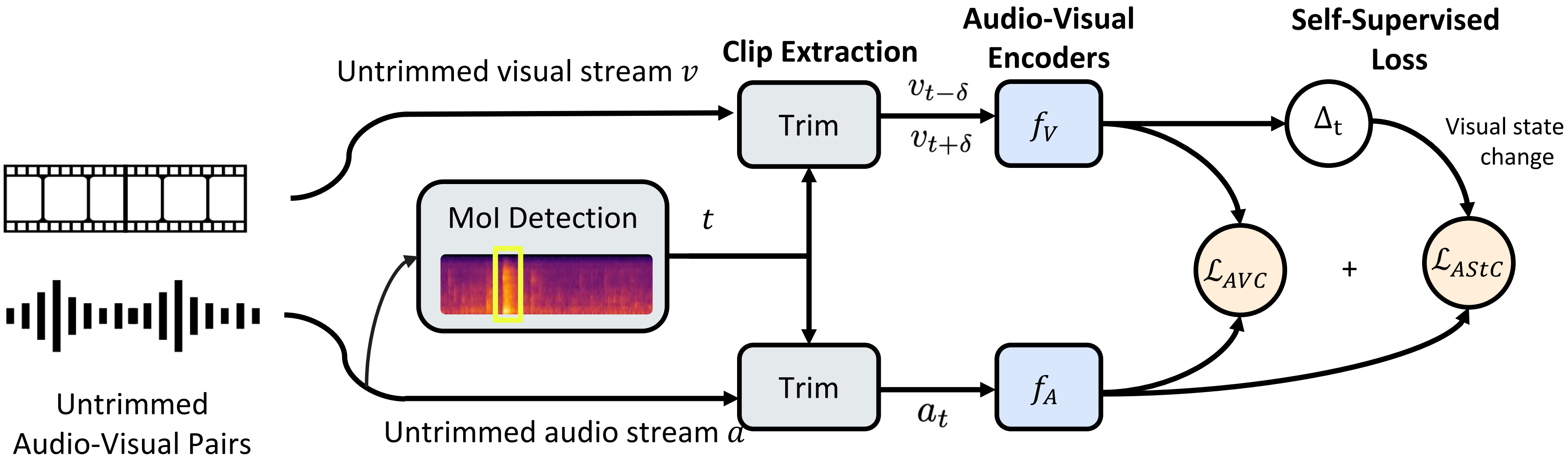


Audio-Visual Encoders
Clip Extraction
Untrimmed visual stream v
Trim
$v_{t-\delta}$
$v_{t+\delta}$
f_V
MoI Detection
t
Trim
f_A
a_t
Untrimmed Audio-Visual Pairs
Untrimmed audio stream a

Method


Untrimmed visual stream v
Clip Extraction
Audio-Visual Encoders
Trim
$v_{t-\delta}$
$v_{t+\delta}$
f_V
MoI Detection
t
$\mathcal{L}_{AVC}$
Trim
a_t
f_A
Untrimmed Audio-Visual Pairs
Untrimmed audio stream a

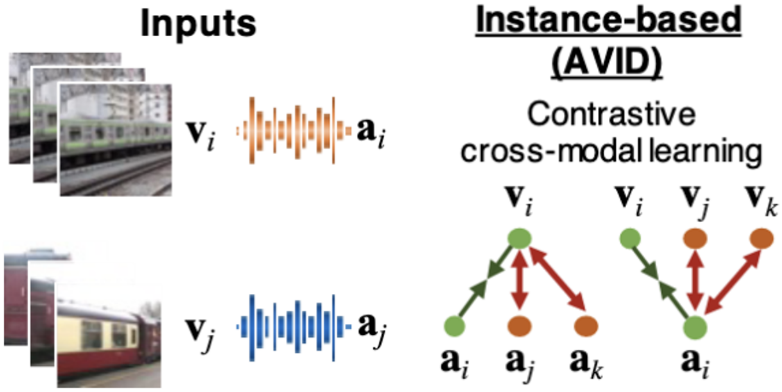
Method



Untrimmed visual stream v
Clip Extraction
Audio-Visual Encoders
Self-Supervised Loss
Trim
$v_{t-\delta}$
$v_{t+\delta}$
f_V
Δ_t
Visual state change
MoI Detection
t
$\mathcal{L}_{AVC}$
+
$\mathcal{L}_{AStC}$
Trim
a_t
f_A
Untrimmed Audio-Visual Pairs
Untrimmed audio stream a

Audio-Visual Correspondence Loss ($\mathcal{L}_{AVC}$)

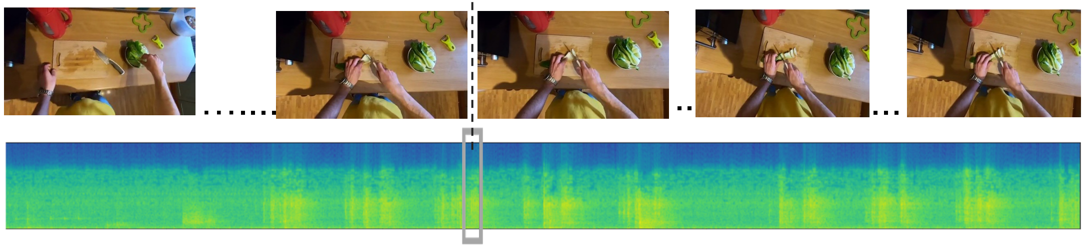
Learns audio-visual representations by contrasting visual representations from multiple audios.



Morgado, Pedro, Nuno Vasconcelos, and Ishan Misra. "Audio-visual instance discrimination with cross-modal agreement." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

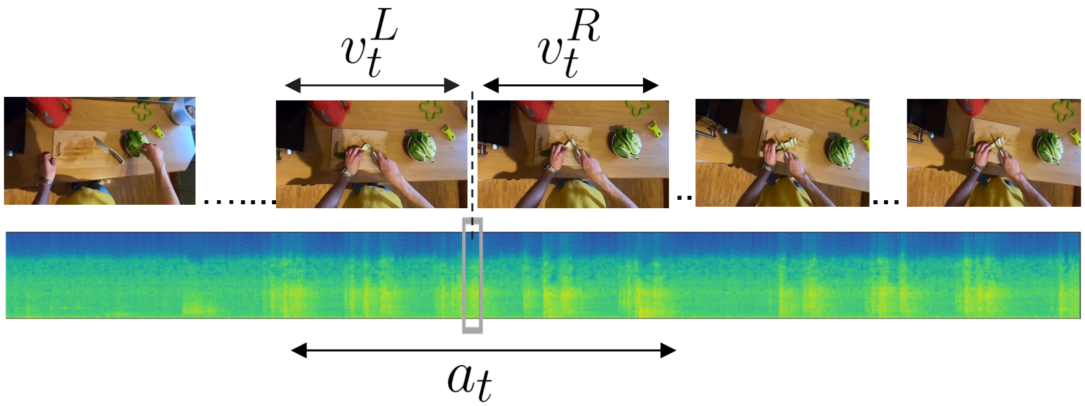
Audio-Visual Correspondence Loss ($\mathcal{L}_{AVC}$)

Learns audio-visual representations by contrasting visual representations from multiple audios.



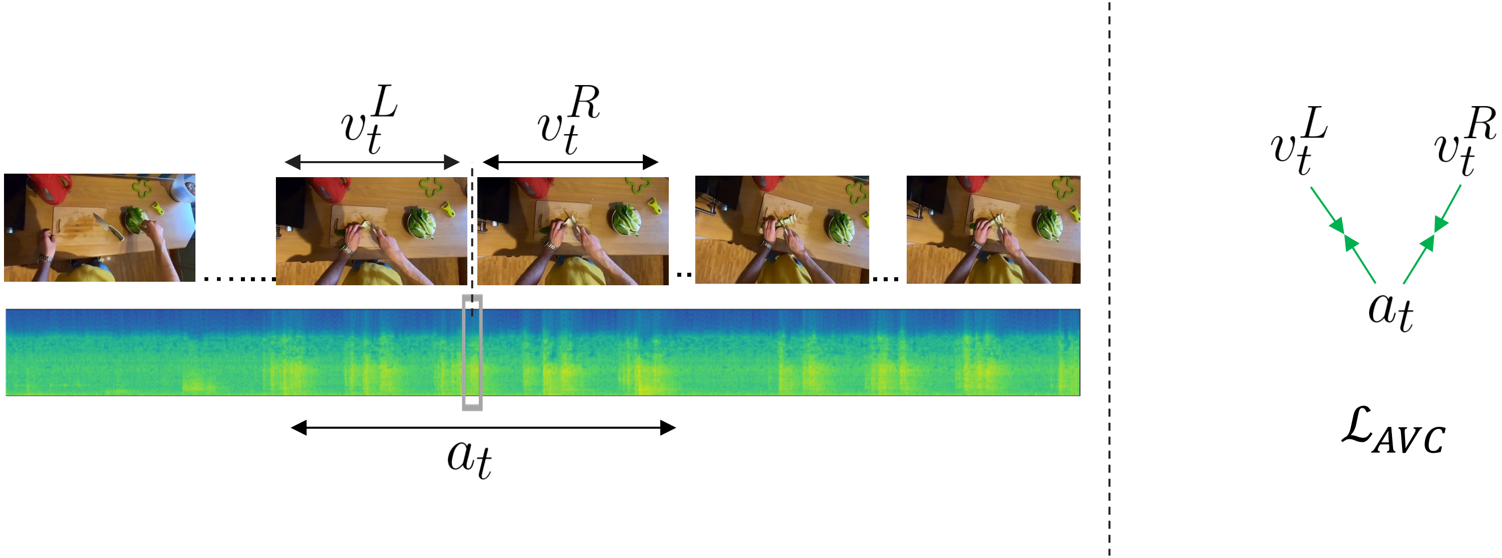
Audio-Visual Correspondence Loss ($\mathcal{L}_{AVC}$)

Learns audio-visual representations by contrasting visual representations from multiple audios.



Audio-Visual Correspondence Loss ($\mathcal{L}_{AVC}$)

Learns audio-visual representations by contrasting visual representations from multiple audios.

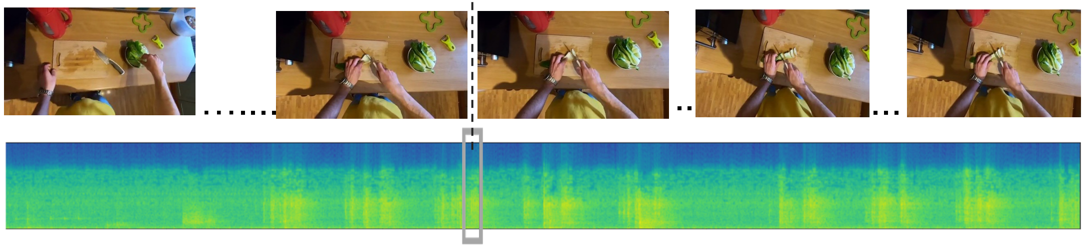


Audible State Change Loss ($\mathcal{L}_{AStC}$)

Our proposed objective function tries to *associate the audio with changes in the visual state* during a moment of interaction.

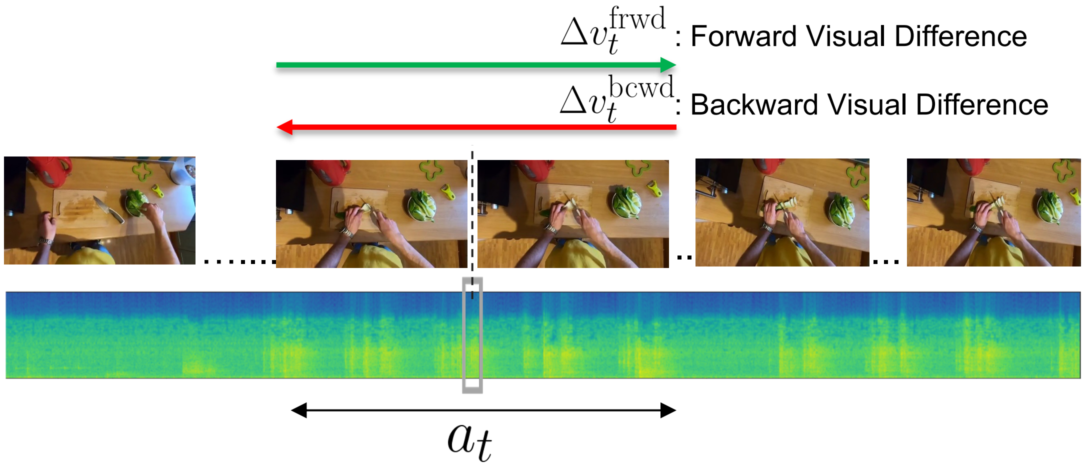
Audible State Change Loss ($\mathcal{L}_{AStC}$)

Our proposed objective function tries to *associate the audio with changes in the visual state* during a moment of interaction.



Audible State Change Loss ($\mathcal{L}_{AStC}$)

Our proposed objective function tries to *associate the audio with changes in the visual state* during a moment of interaction.

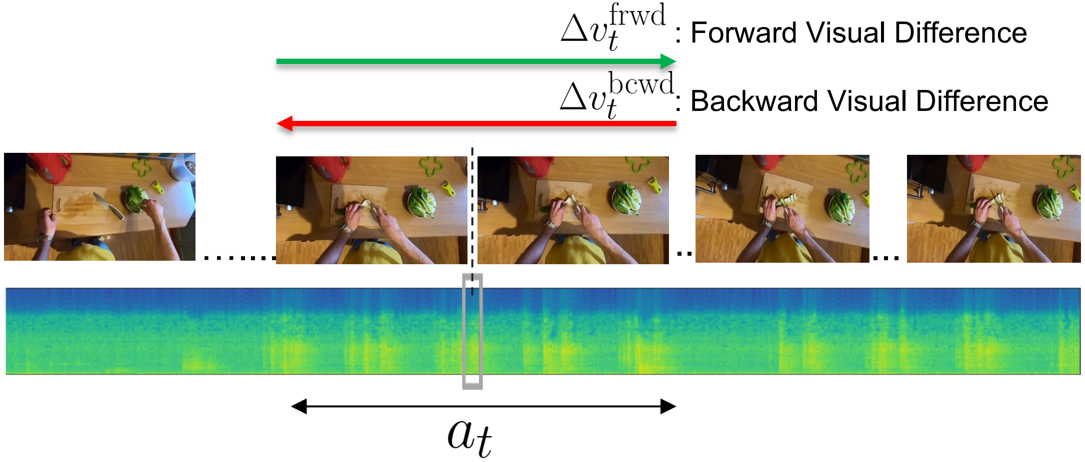


$$\Delta v_t^{frwd} = \phi(v_T^R - v_T^L)$$

$$\Delta v_t^{bcwd} = \phi(v_T^L - v_T^R)$$

Audible State Change Loss ($\mathcal{L}_{AStC}$)

Our proposed objective function tries to *associate the audio with changes in the visual state* during a moment of interaction.



$$\Delta v_t^{frwd} = \phi(v_T^R - v_T^L)$$

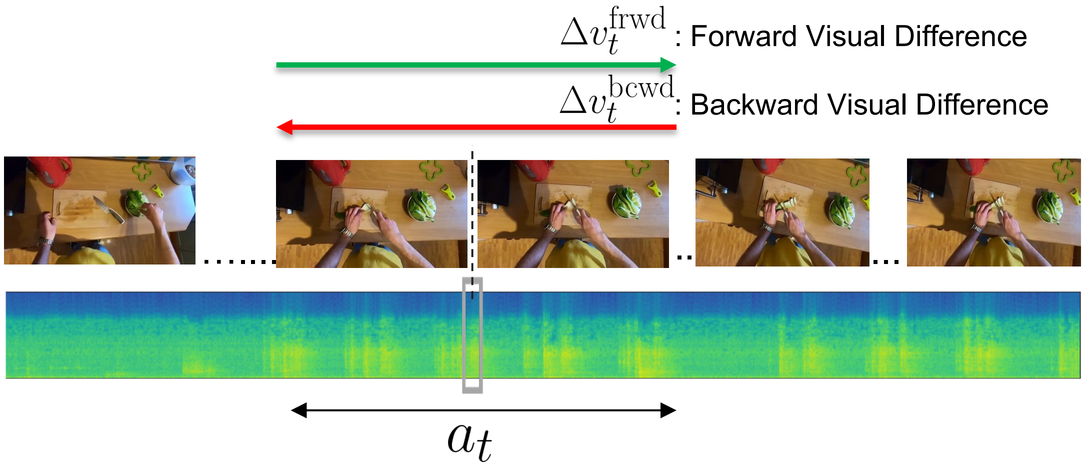
$$\Delta v_t^{bcwd} = \phi(v_T^L - v_T^R)$$

$$< \Delta v_t^{\text{frwd}}, a_t > \uparrow$$

$$\mathcal{L}_{AStC}$$

Audible State Change Loss ($\mathcal{L}_{AStC}$)

Our proposed objective function tries to *associate the audio with changes in the visual state* during a moment of interaction.



$$\Delta v_t^{frwd} = \phi(v_T^R - v_T^L)$$

$$\Delta v_t^{bcwd} = \phi(v_T^L - v_T^R)$$

$$< \Delta v_t^{\text{frwd}}, a_t > \uparrow \quad < \Delta v_t^{\text{bcwd}}, a_t > \downarrow$$

$$\mathcal{L}_{AStC}$$

Datasets (1)

EPIC-Kitchens-100 : Consists of 100 hours of activities in the kitchen

Damen, Dima, et al. "Rescaling egocentric vision." *arXiv preprint arXiv:2006.13256* (2020).

Datasets (2)

Ego4D : Contains 3,670 hours of ego-centric video covering daily activities in the home, workplace, social settings, etc.

Grauman, Kristen, et al. "Ego4d: Around the world in 3,000 hours of egocentric video." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

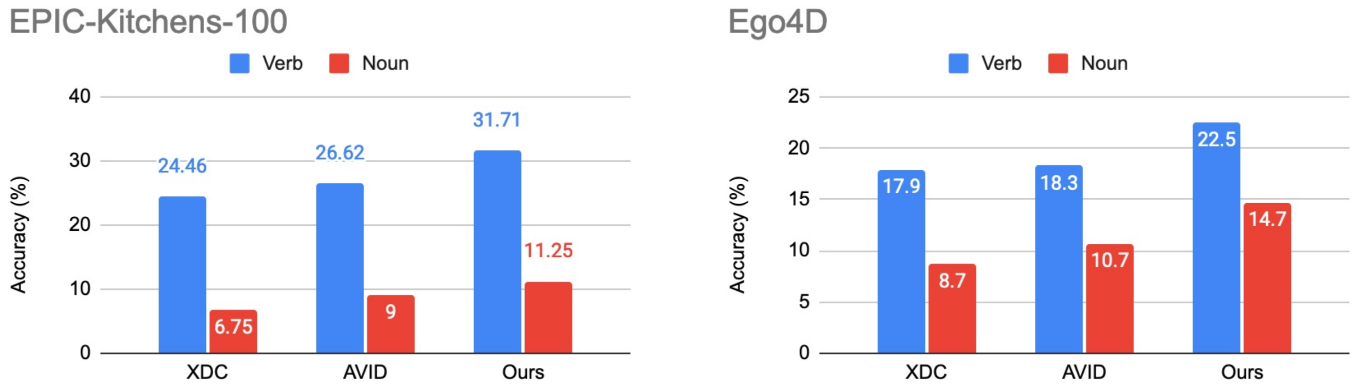
Quantitative Results on Action Recognition

Predict the verb and noun of an action

Damen, Dima, et al. "Rescaling egocentric vision." *arXiv preprint arXiv:2006.13256* (2020).

Discussion of Results (1)

Compared to the baselines – XDC and AVID, our method performs better by significant margins.

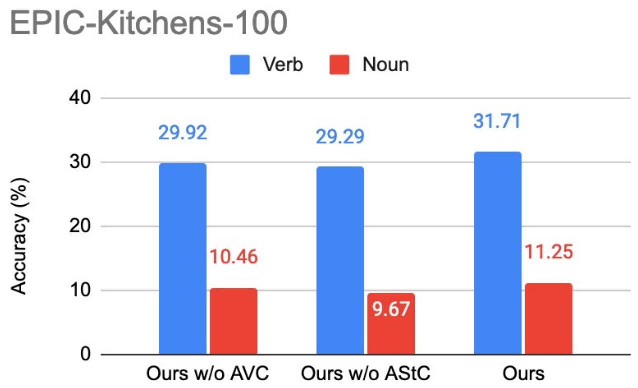


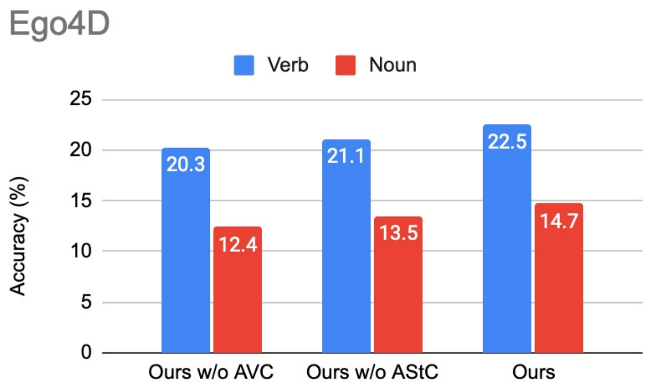
Action Recognition (Top-1 Accuracy) on EPIC-Kitchens-100 (left) and Ego4D (right)
Higher is better.

[XDC]. Alwassel, Humam, et al. "Self-supervised learning by cross-modal audio-video clustering." NeurIPS 2020
[AVID]. Morgado, Pedro, et. al. "Audio-visual instance discrimination with cross-modal agreement.", CVPR. 2021.

Discussion of Results (2)

Ablation Study (1) - Our method w/o AVC and w/o AStC: Each term enhances the representations obtained through large-scale audio-visual pre-training.

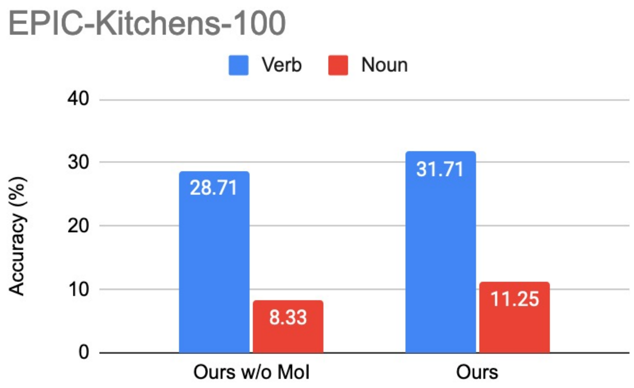


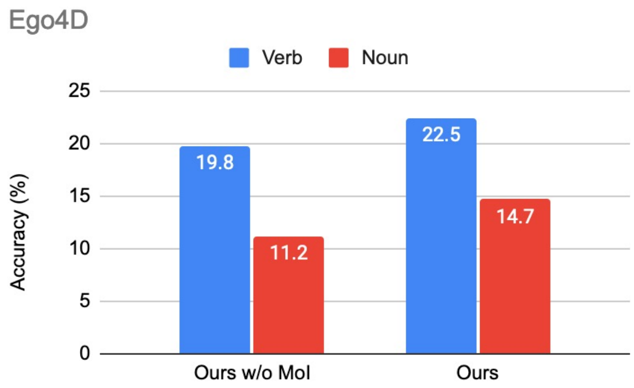


Action Recognition (Top-1 Accuracy) on EPIC-Kitchens-100 (left) and Ego4D (right)
Higher is better.

Discussion of Results (3)

Ablation Study (2) - Our method without moments of interaction (MoI): Detecting moments of interaction helps representation learning.





Action Recognition (Top-1 Accuracy) on EPIC-Kitchens-100 (left) and Ego4D (right)
Higher is better.

Main Takeaways/Conclusion

- We propose an audio-driven self-supervised method for learning representations of egocentric video of daily activities.

- For better representations of daily activities, learning should **focus on moments of interaction** (MoI).
    - Simple spectrogram-based MoI detector works, but there is room for improvement.

- For better representations of daily activities, models should **learn from the changes in the environment** caused by agents interacting with the world.
    - Audio can be informative of both the objects in an environment and changes in their state.