

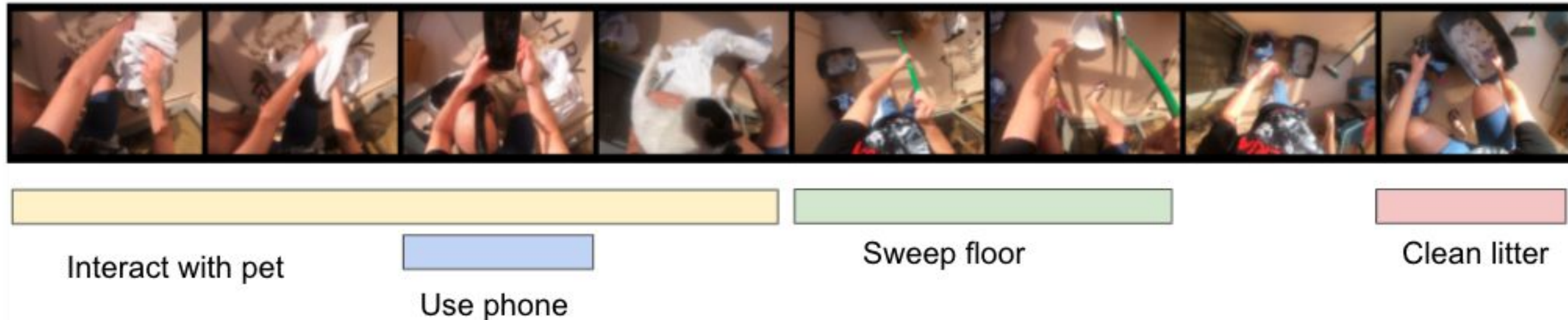
Ego-Only: Egocentric Action Detection without Exocentric Transferring

Huiyu Wang, Mitesh Kumar Singh, Lorenzo Torresani

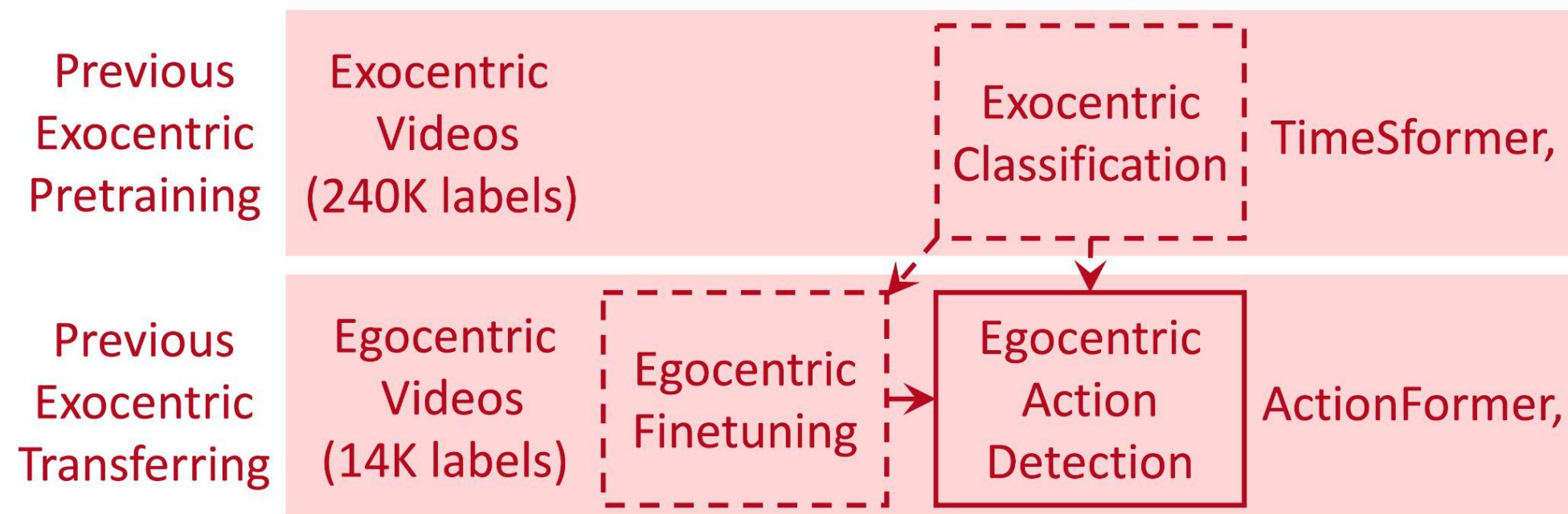
Meta AI

Goal

Detect human actions: (class, start, end)



Exocentric Transferring

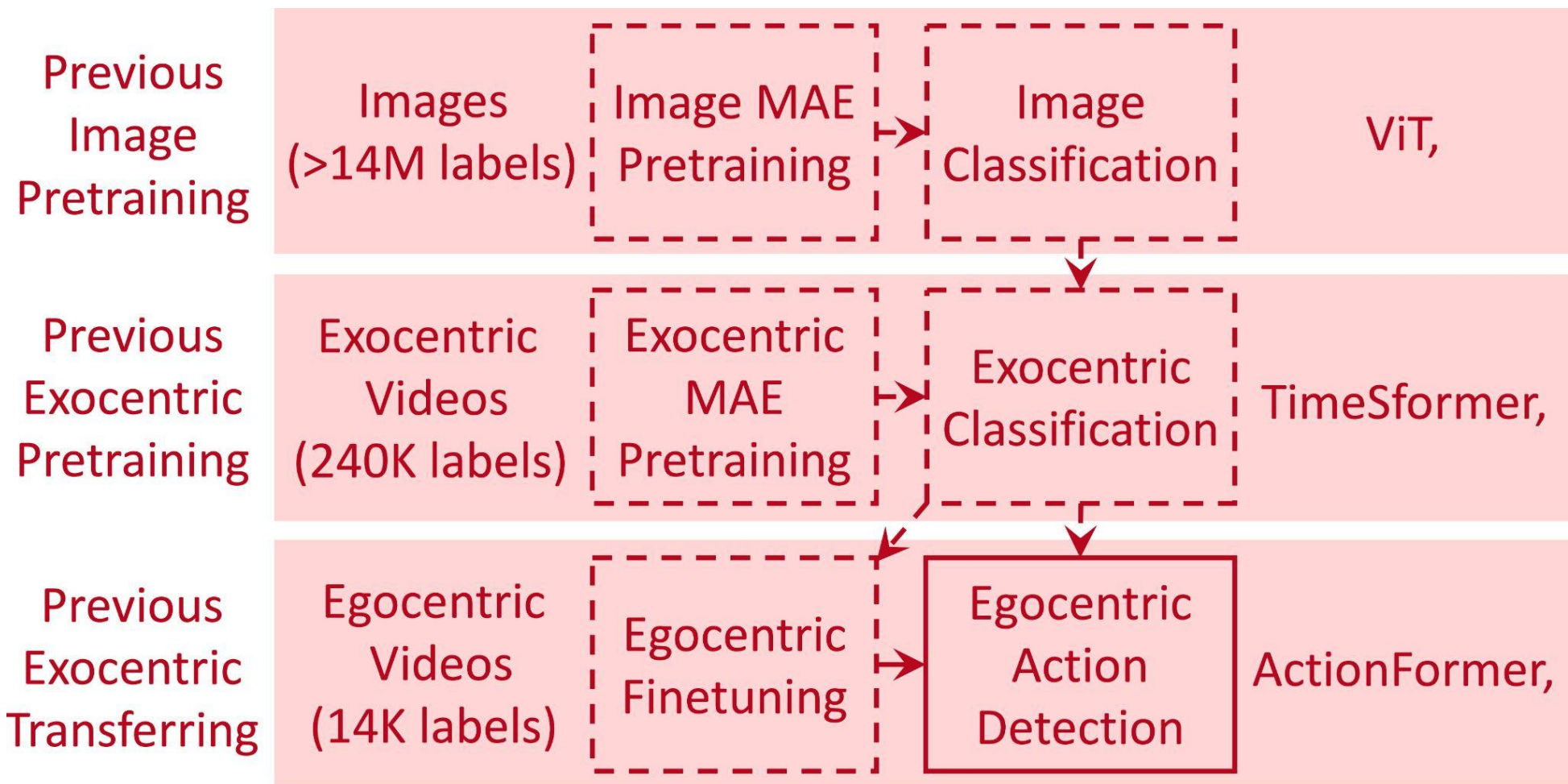


Kay, W., et al. The kinetics human action video dataset. ArXiv 2017.

Bertasius, G., et al. Is space-time attention all you need for video understanding? ICML 2021.

Zhang, C., et al. Actionformer: Localizing moments of actions with transformers. ECCV 2022.

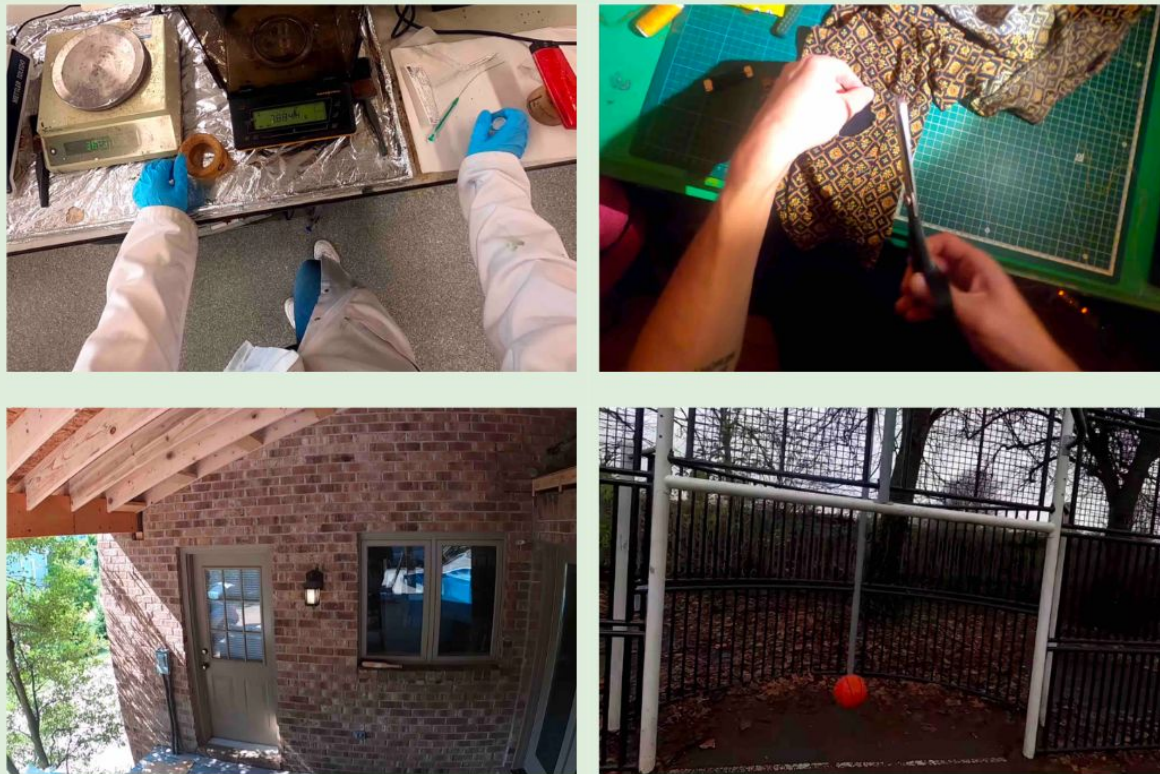
Exocentric Transferring



Deng, J., et al. Imagenet: A large-scale hierarchical image database. CVPR 2009.
Dosovitskiy, A., et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021.
He, K., et al. Masked autoencoders are scalable vision learners. CVPR 2022.

Challenging to Transfer

Egocentric Videos
(length: 480 seconds)



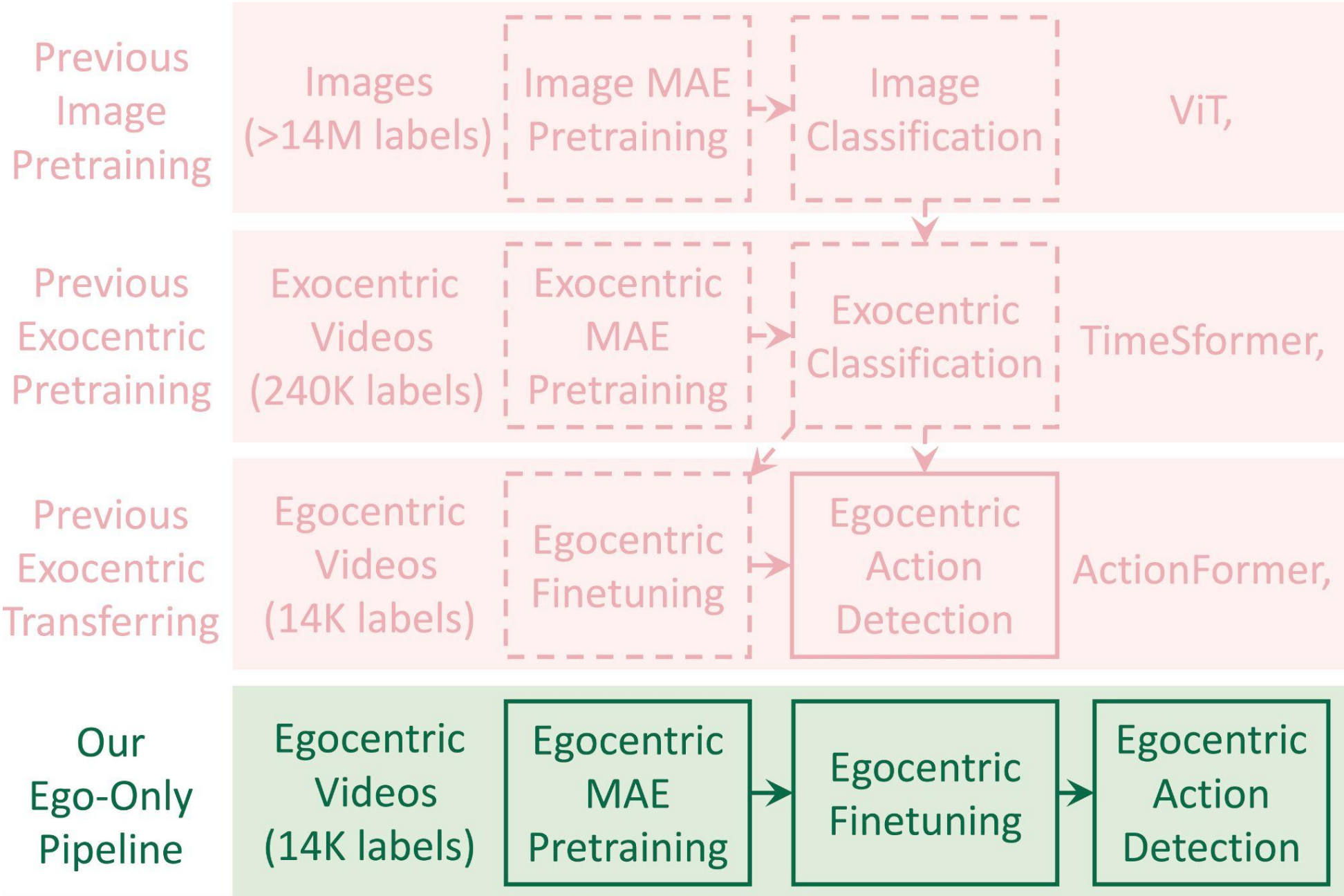
Exocentric Videos
(length: 10 seconds)



- No actor in view
- Object interaction
- Domain shift
- Class granularity
- Long-tail
- Long-form
- Localization

Ego-Only: Egocentric Action Detection without Exocentric Transferring

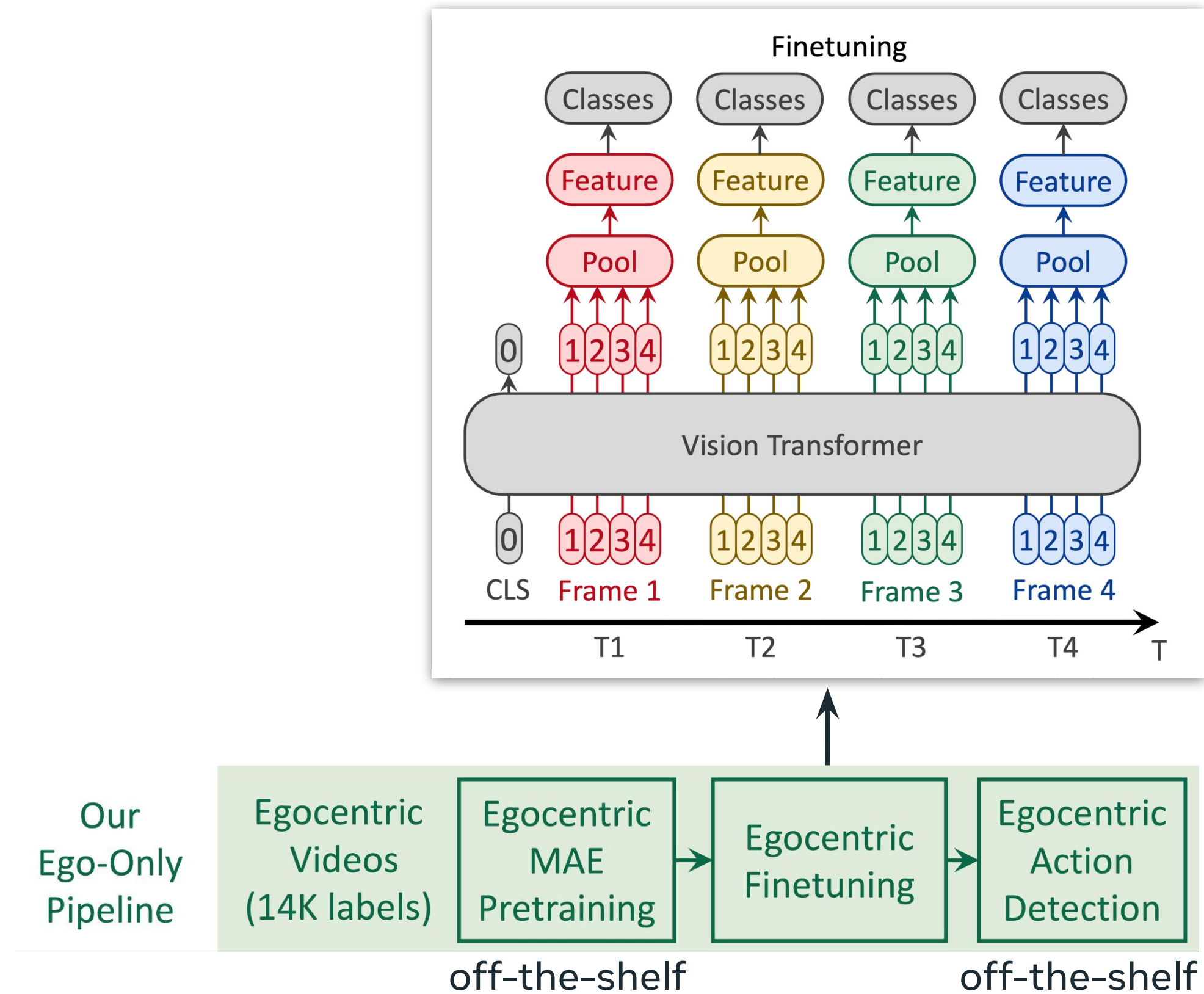
Ego-Only



better
with sufficient supervision

Ego-Only: Egocentric Action Detection without Exocentric Transferring

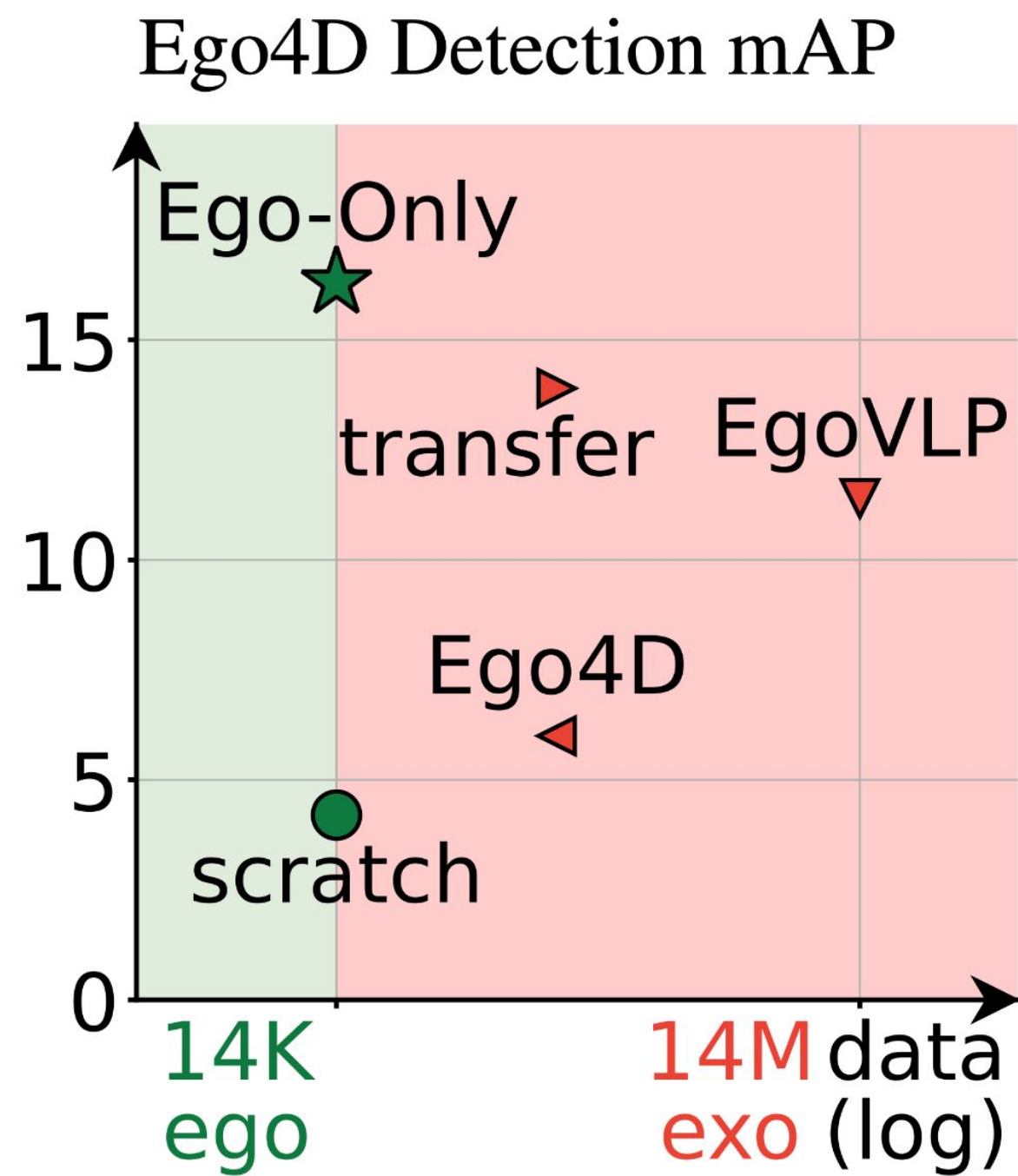
Ego-Only



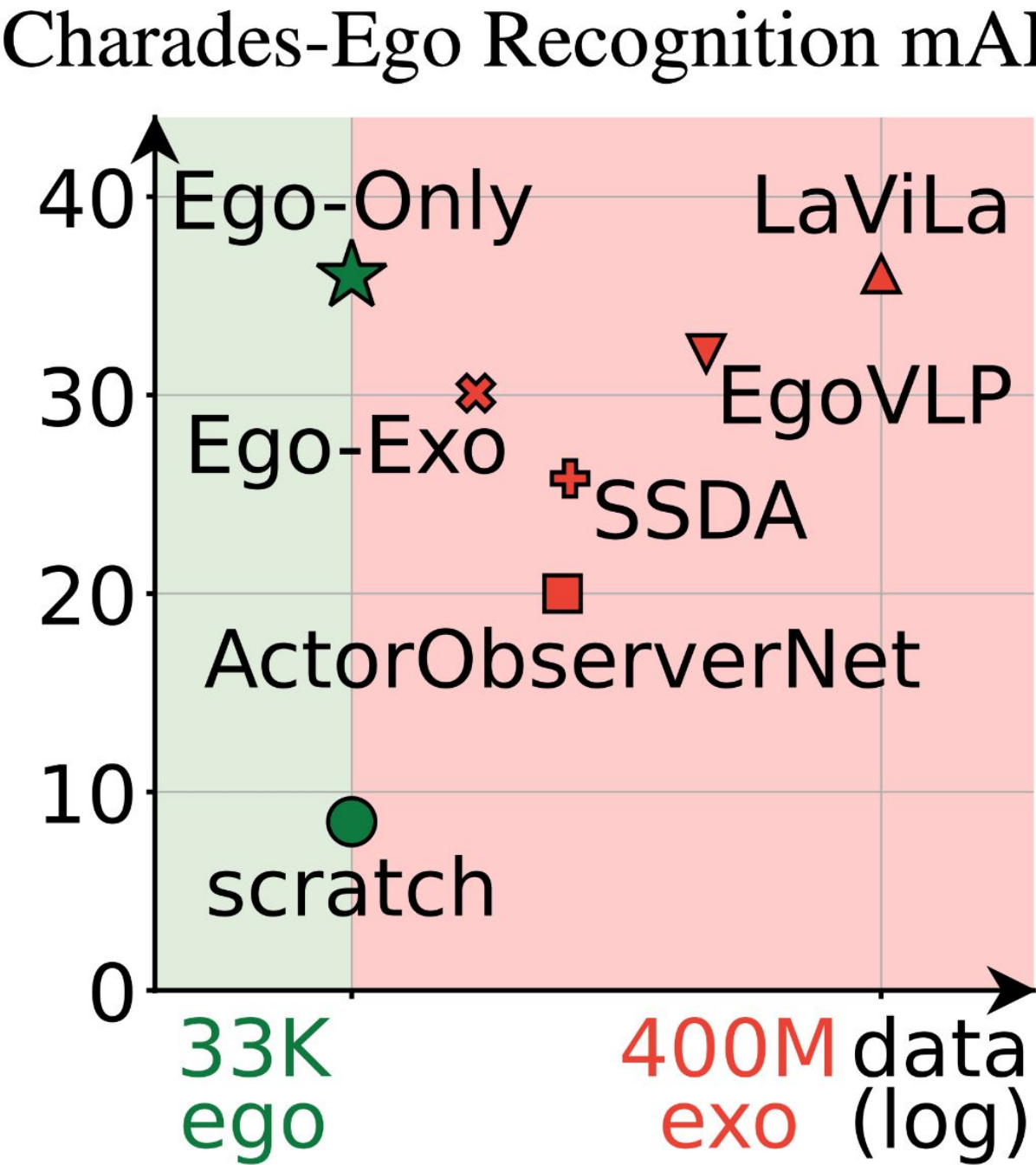
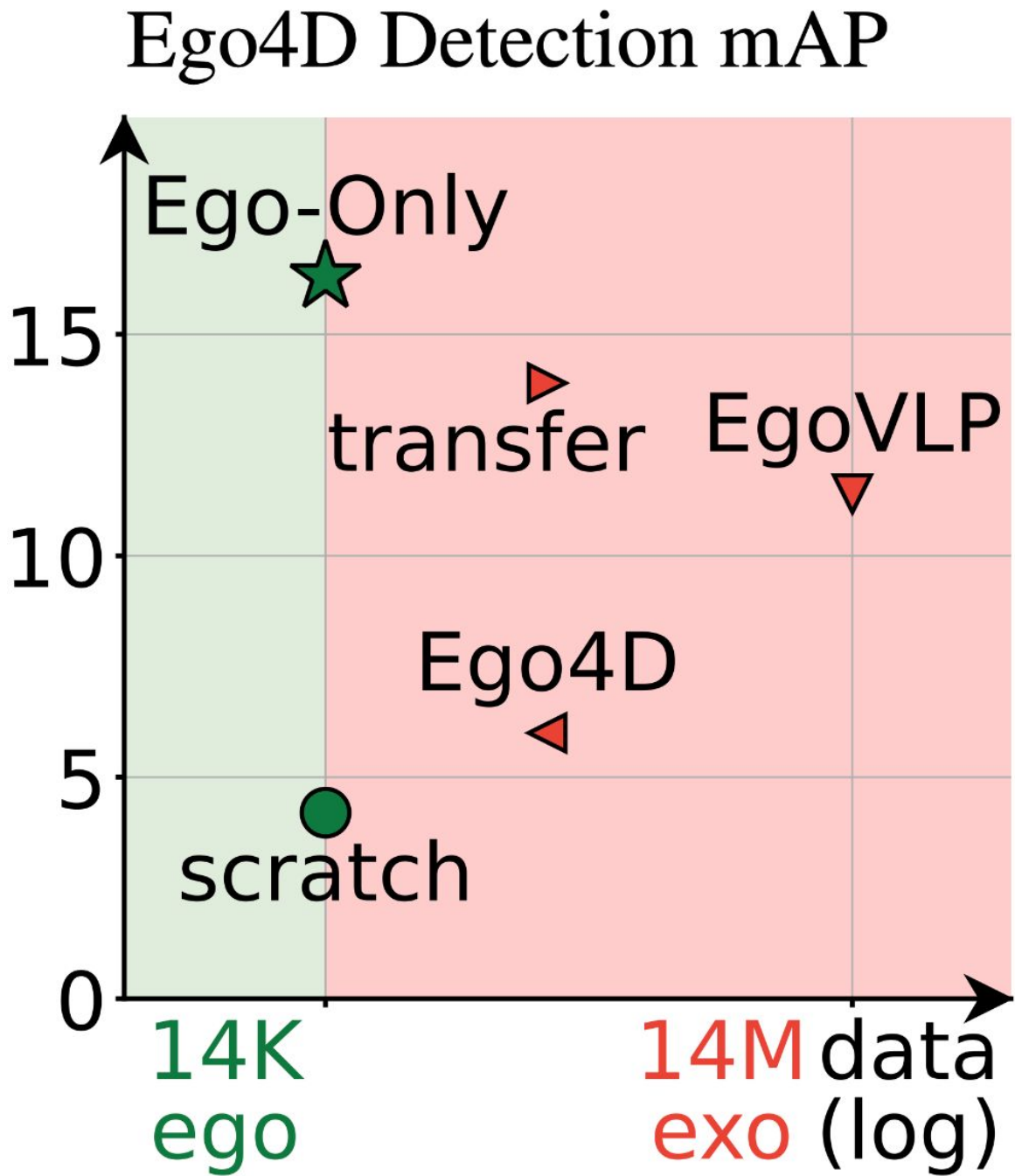
Feichtenhofer, C., et al. Masked autoencoders as spatiotemporal learners. NeurIPS 2022.

Zhang, C., et al. Actionformer: Localizing moments of actions with transformers. ECCV 2022.

Results

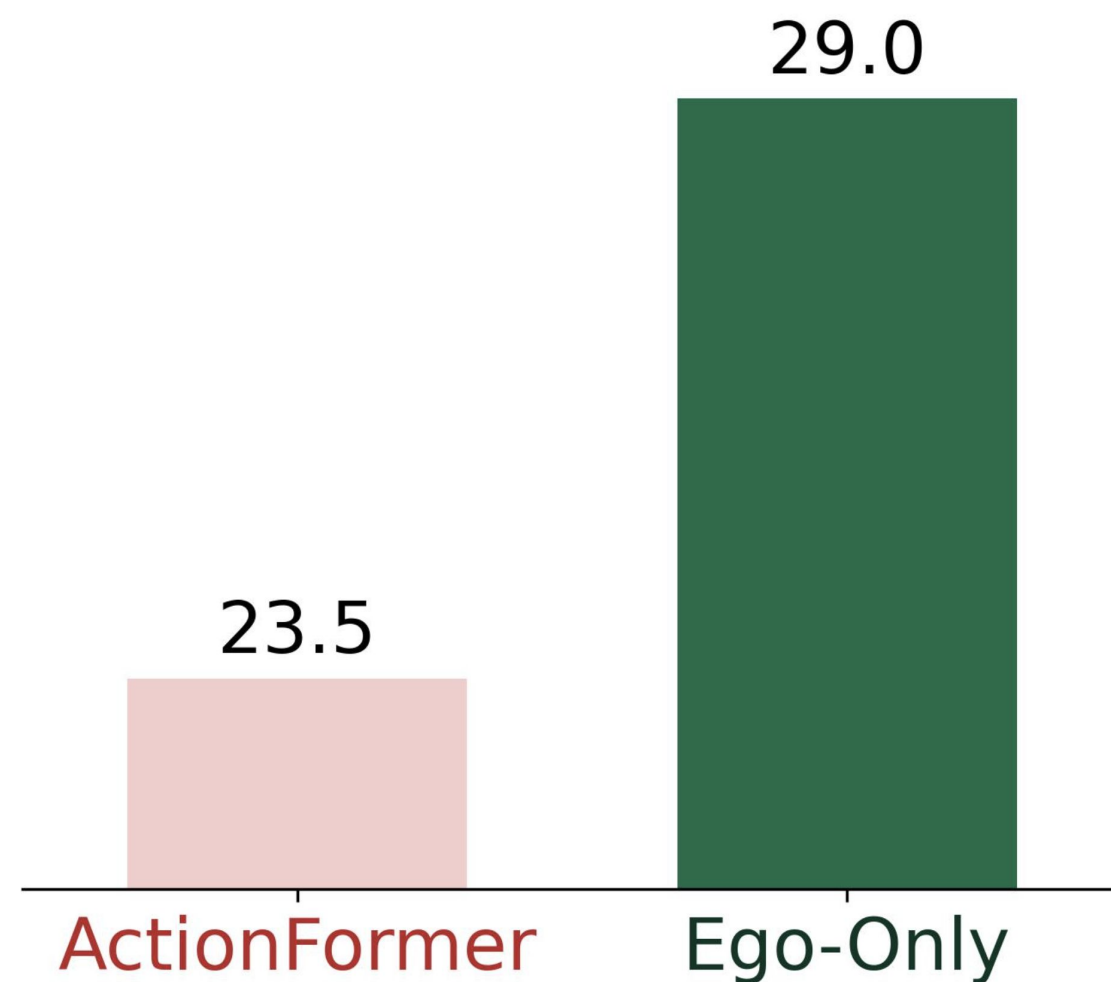


Results



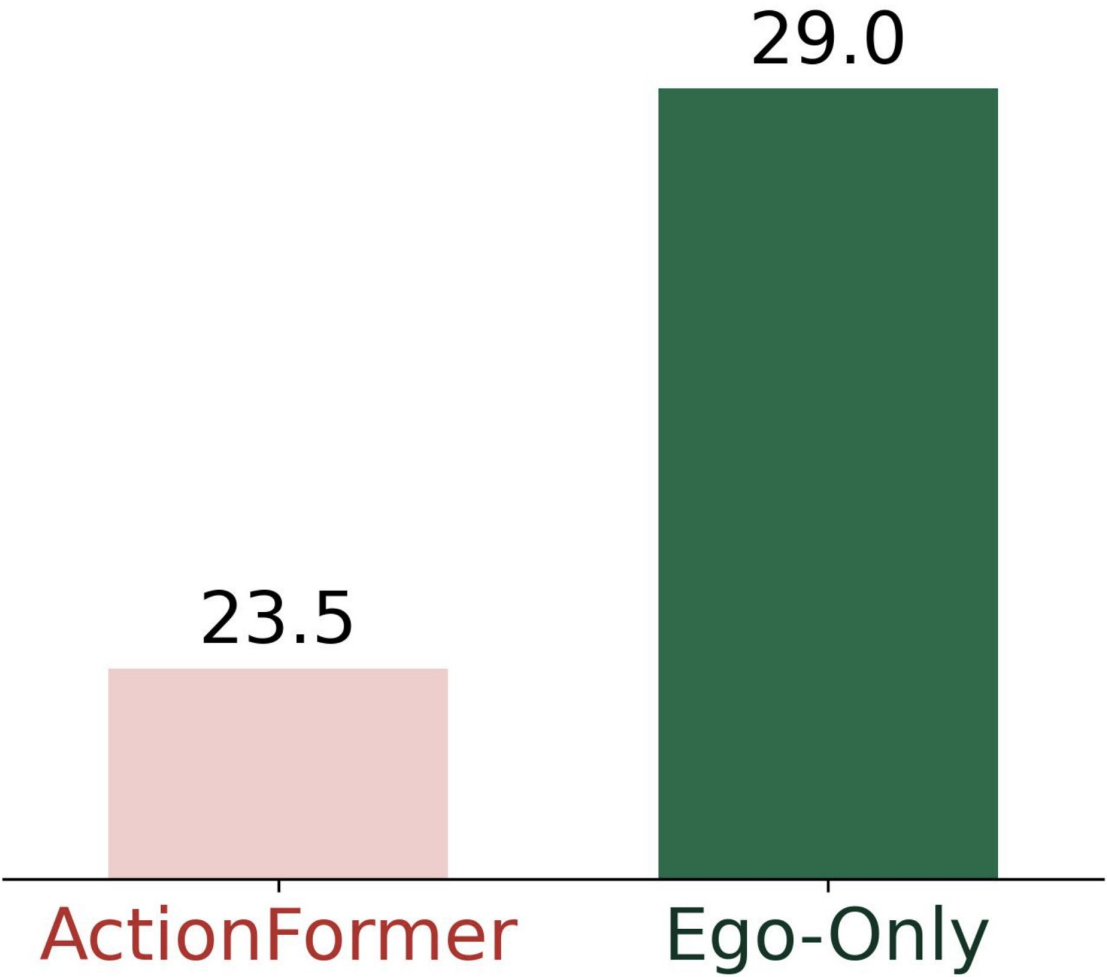
EPIC-Kitchens-100

Action Detection mAP



EPIC-Kitchens-100

Action Detection mAP



Action Recognition Top-1

method	variant	verb	noun
ViViT	ViViT-L/16x2, IN-21k+K400	66.4	56.8
MoViNet	MoViNet-A6, 120 frames	72.2	57.3
MTV	MTV-B, WTS-60M, 280p	69.9	63.9
MTCN	MFormer-HR, IN-21k+K400+VGG-Sound	70.7	62.1
Omnivore	Swin-B, IN21k+IN-1k+K400+SUN	69.5	61.7
MeMViT	MeMViT, 32×3, K600, 105.6 sec	71.4	60.3
LaViLa	TSF-L, WebImageText+Ego4D	72.0	62.9
Ego-Only	ViT-L, 32 frames, 3.2 sec	73.3	59.4

MAE Reconstruction

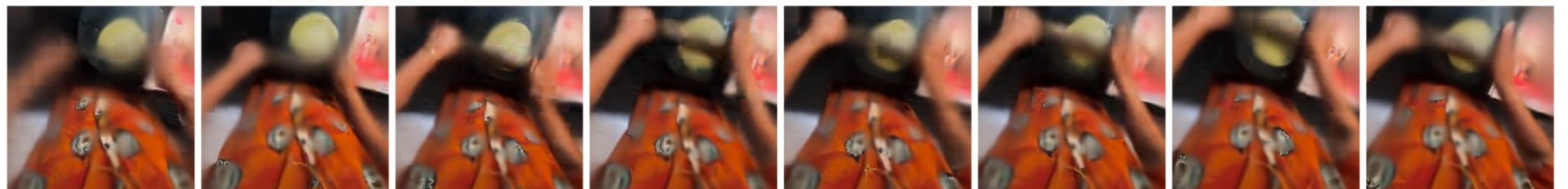
Original



Masked



Prediction



MAE Reconstruction

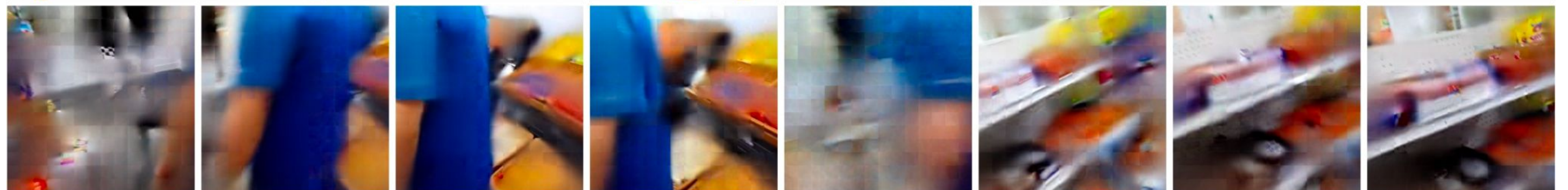
Original



Masked



Prediction



Summary

1. Ego-Only can be trained without exocentric transferring
2. Ego-Only is an order of magnitude more label-efficient
3. Ego-Only improves results over the state-of-the-art