Focus of Attention in Self-Supervised Learning for Action Recognition

Vansh Tibrewal[1], Bart Thomson[2], Michael Hugelshofer[2], Henning Richter[2], Pietro Perona[1], Neehar Kondapaneni[1*], Markus Marks[1*†]

[1]California Institute of Technology, Pasadena, CA, USA
[2]University of Zurich, Zurich, Switzerland
†corresponding author: marks@caltech.edu
*equal contribution

CVPR Nashville JUNE 11-15, 2025

Background

Motivation

- Developing **accurate quantitative models** of animal behavior is **important**.
- **Manual** animal action recognition is too **time** and **labor-intensive** for large datasets.
  - **Supervised approaches struggle** in the low-data regime.
- Existing approaches:
  - Focus on specific behaviors or specific animals.
  - Use **pose-estimation techniques** which **lose rich background** in natural settings.
  - Require a lot of annotated data.

Our Contribution

- We collect a **novel dataset** of videos of **sheep** that **underwent surgery** and annotated behaviors relevant to their wellbeing.
- We use **SAM-2 to guide DINOv2 fine-tuning** on those videos and demonstrate that this guidance can significantly **improve feature extraction**.
- We train a classifier on the fine-tuned DINOv2 embeddings of the video and use it to **identify behavioral differences between experimental groups**.

Framework

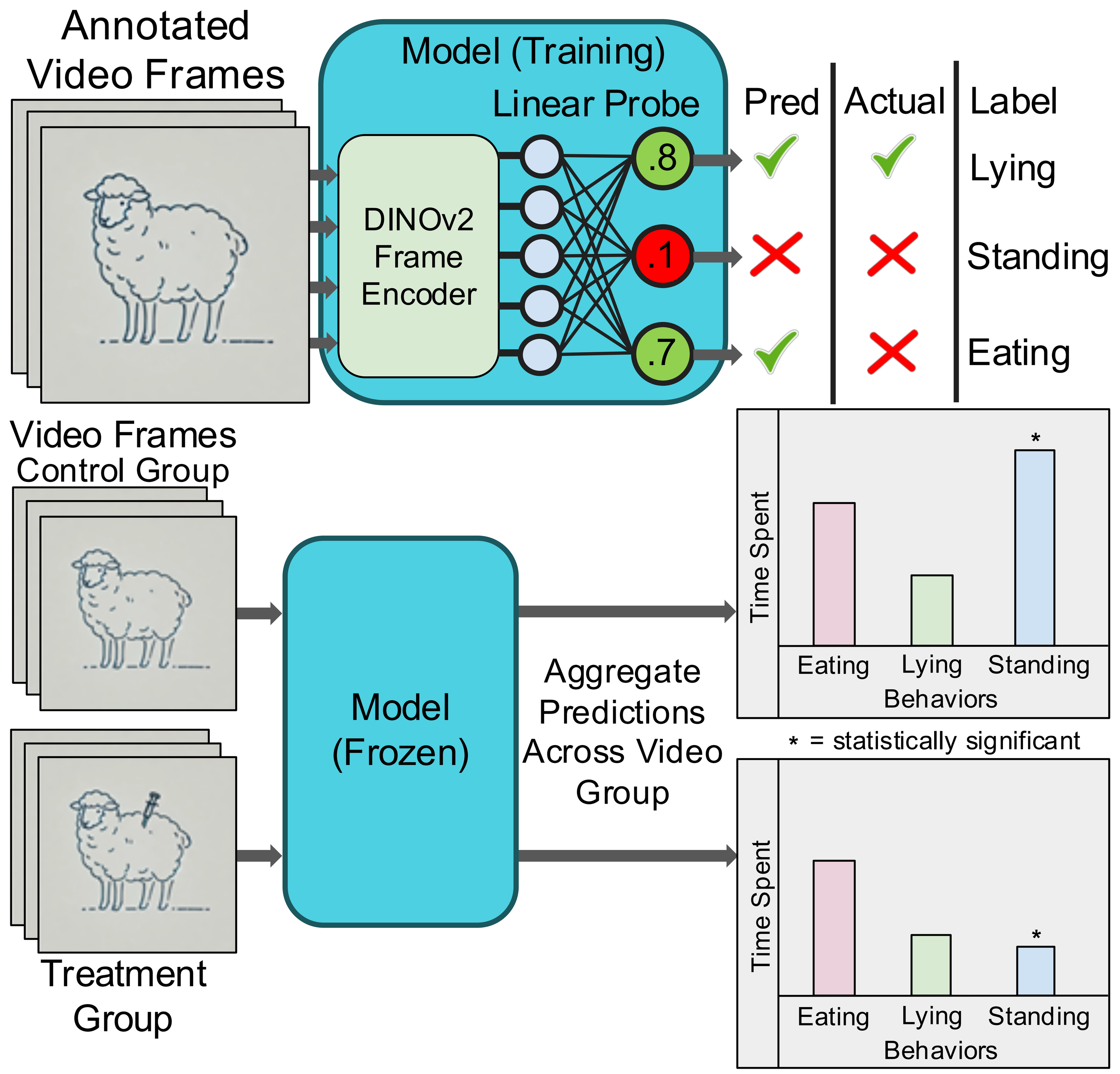


Fig 1. **Training + Inference Phase.** Linear probe trained to classify sheep behaviors with frozen feature extractor. Trained linear probe is used to predict behaviors of different groups of sheep.

SAM-2 Guidance

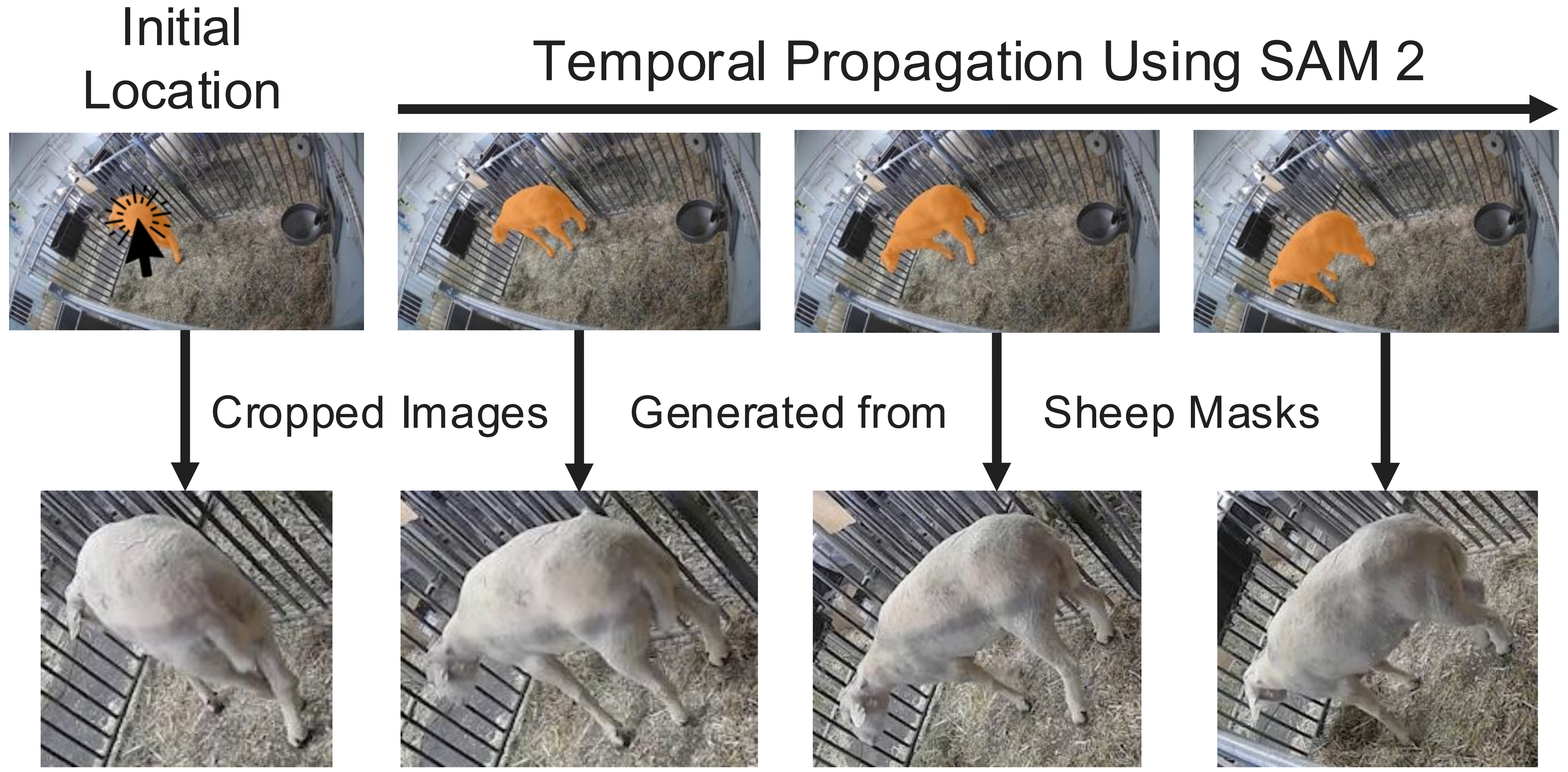


Fig 2. We use SAM-2 to crop around the sheep and then fine-tune DINOv2 on the cropped version of the dataset.

Application

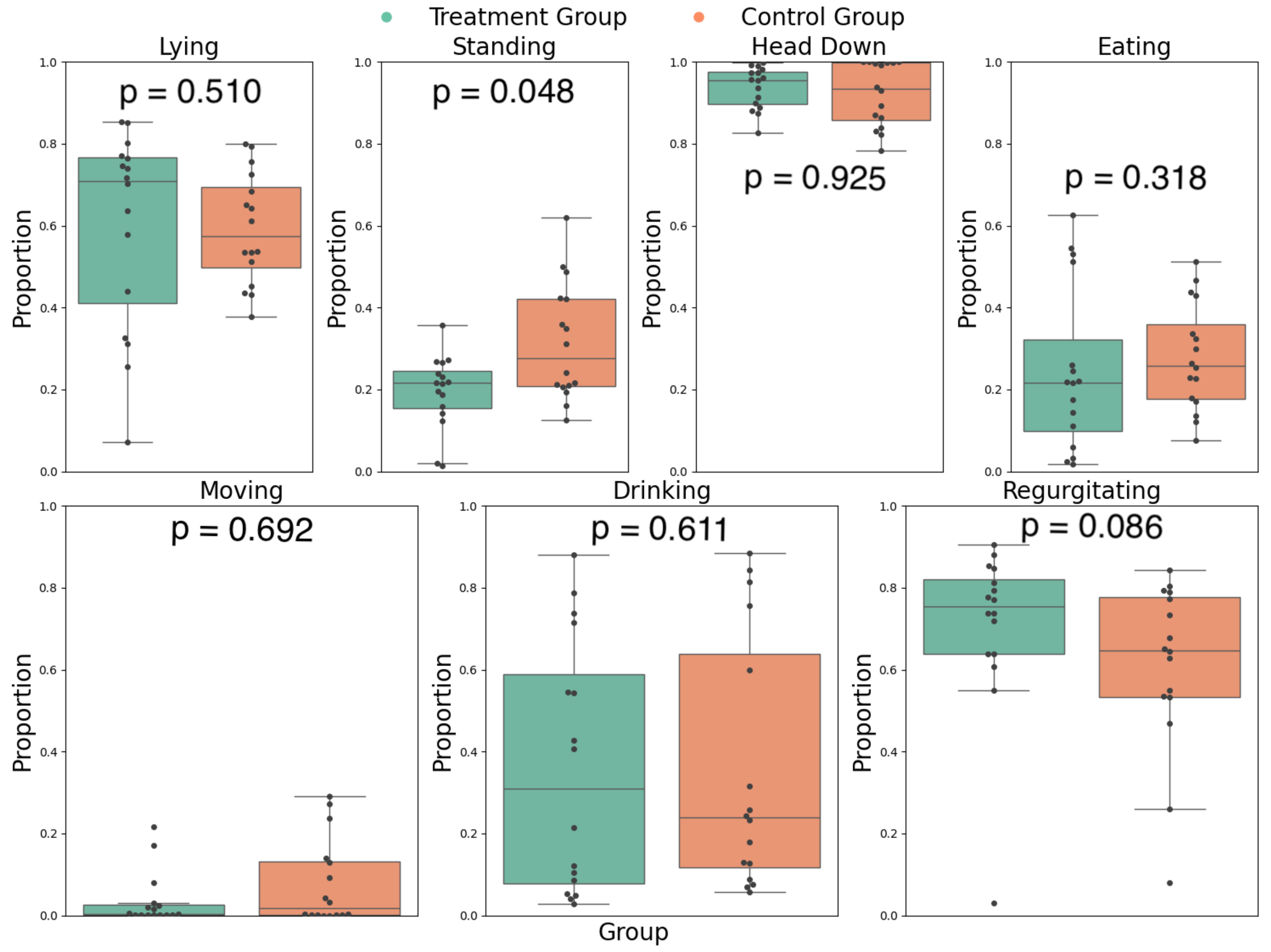


Fig 3. Distribution of proportions of labels detected in videos of the 2 groups of sheep. There's a statistically significant difference in the proportion of "Standing", with the control group standing more, which is expected neuroscientifically.

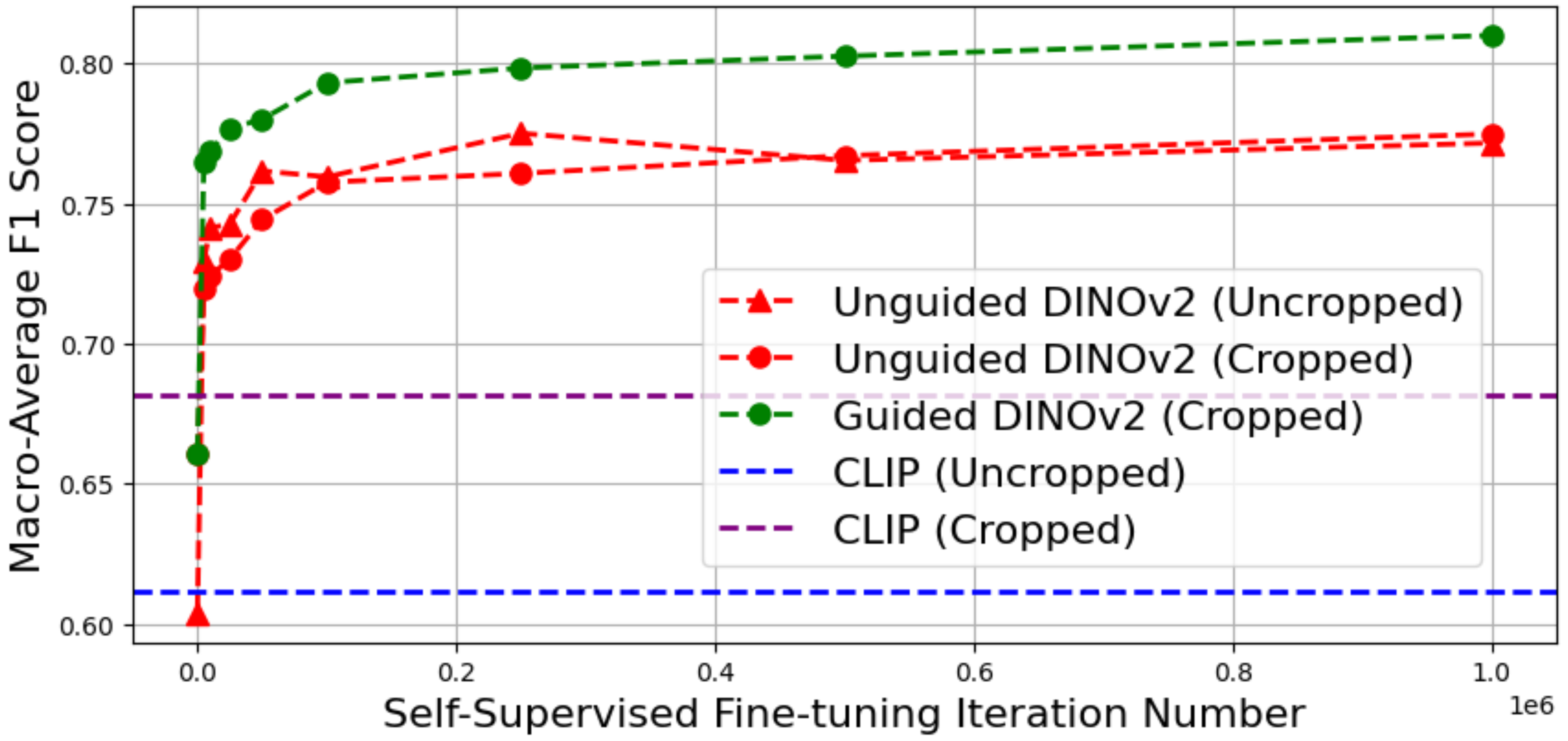
Quantitative Improvements

	Labels						Macro-Average
	Head Down	Lying	Standing	Eating	Head Up	Moving	
SAM 2 Guidance							
✗	0.9237	0.9811	0.9755	0.7892	0.4954	0.4840	0.7748
✓	**0.9246**	**0.9937**	**0.9782**	**0.8232**	**0.6298**	**0.5416**	**0.8152**
% Change	+0.10%	+1.28%	+0.28%	+4.31%	+27.1%	+11.9%	+5.21%
Temporal Information							
✗	0.9246	0.9937	**0.9782**	0.8232	0.6298	0.5416	0.8152
✓	**0.9331**	**0.9943**	0.9754	**0.8375**	**0.6711**	**0.5732**	**0.8308**
% Change	+0.92%	+0.06%	-0.29%	+1.74%	+6.56%	+5.83%	+1.91%

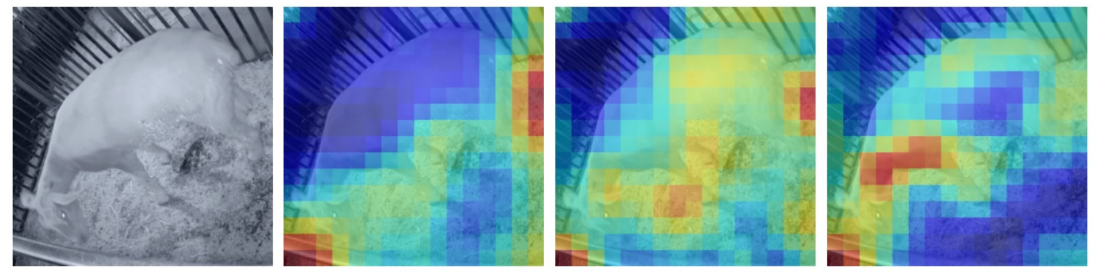
SAM 2 guidance provides a significant, consistent downstream performance boost over regular DINOv2 fine-tuning, across all labels.

Incorporating temporal information further improves performance on dynamic behaviors such as moving.

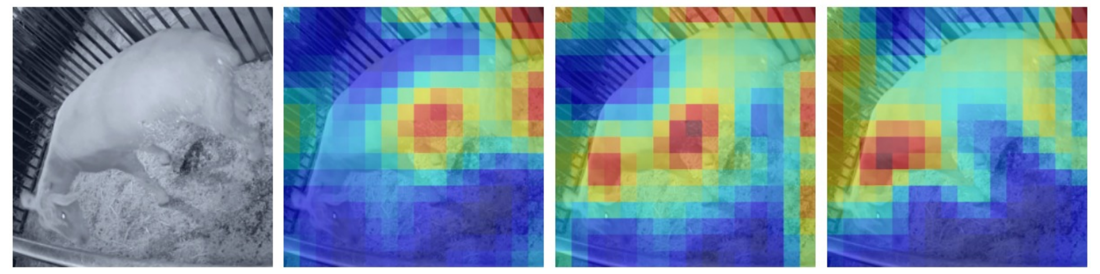
Baseline vs Regular DINOv2 vs Guided DINOv2



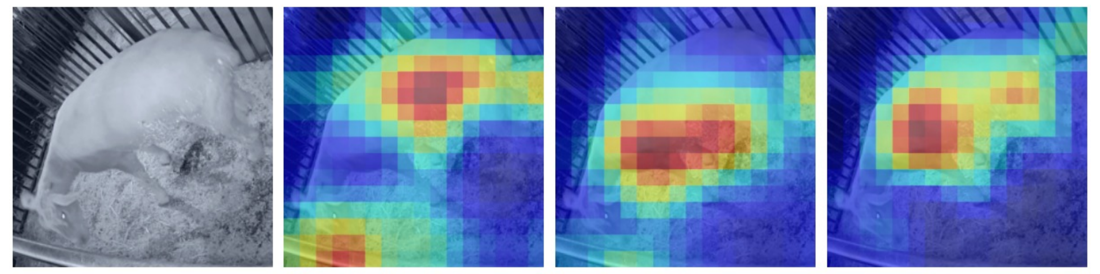
Qualitative Improvement



(a) DINOv2 Baseline Attention Maps for 3 heads.



(b) Unguided fine-tuned DINOv2 Attention Maps for 3 heads.



(c) Segmentation-guided fine-tuned DINOv2 Attention Maps for 3 heads.