Focusing Attention in Self-Supervised Learning for Action Recognition Supplementary Materials

1. Contents

In this supplement, we offer 4 main additional details:

- We provide additional quantitative results in the form of label-wise F1 scores achieved by different DINOv2 checkpoints in various experiments.
- We provide additional ablations, results and details about incorporating temporal information.
- We provide additional qualitative results in the form of attention map visualizations of baseline DINOv2, unguided DINOv2, and segmentation-guided DINOv2 on sheep images of different behaviors.
- We provide further details about model training, evaluation, and inference.

2. Additional Quantitative Results

In the main paper, we give best label-wise F1 scores for unguided DINOv2 evaluated on uncropped data and segmentation-guided DINOv2 evaluated on cropped data. We also compare them to unguided DINOv2 evaluated on cropped data. In Tables 1 and 2, we give the full label-wise F1 scores as the steps of fine-tuning proceed for both cropped settings, using guided and unguided DINOv2.

The results in Tables 1 and 2 are in line with what we expect. The results affirm two observations we already made in the main paper:

- The (unguided and segmentation-guided) fine-tuning provides a consistent performance improvement, with the best F1 score being achieved at or after 100k iterations for all labels in both tables.
- However, the best F1 scores for unguided DINOv2 on cropped images is still comparable to those of unguided DINOv2 evaluated on uncropped images and markedly less than segmentation-guided DINOv2 evaluated on cropped images.

3. Additional Temporal Information Details

We ablate on the window size, as seen in Figure 1.

We find that increasing window size has diminishing returns, with best macro-average F1 score being achieved by a window size of 7.

	Labels						Macro-Average
	Head Down	Lying	Standing	Eating	Head Up	Moving	
0	0.9151	0.8218	0.8130	0.6692	0.2789	0.4668	0.6608
5k	0.9156	0.9401	0.9516	0.7702	0.5119	0.4973	0.7644
10k	0.9229	0.9207	0.9533	0.7793	0.5356	0.5002	0.7687
25k	0.9225	0.9754	0.9658	0.7861	0.5799	0.4306	0.7767
50k	0.9229	0.9810	0.9698	0.7937	0.5186	0.4943	0.7801
100k	0.9231	0.9898	0.9735	0.8127	0.5177	**0.5416**	0.7931
250k	0.9220	0.9926	**0.9782**	0.8195	0.5424	0.5354	0.7984
500k	**0.9246**	0.9904	0.9675	0.8171	0.5963	0.5195	0.8026
1M	0.9215	**0.9937**	0.9742	**0.8232**	**0.6298**	0.5176	0.8100
Guided Best	**0.9246**	**0.9937**	**0.9782**	**0.8232**	**0.6298**	**0.5416**	**0.8152**
Unguided Best	0.9237	0.9811	0.9755	0.7892	0.4954	0.4840	0.7748
% Change	+0.10%	+1.28%	+0.28%	+4.31%	+27.1%	+11.9%	+5.21%

Table 1. F1 scores for DINOv2 fine-tuning with SAM 2 guidance at various iteration checkpoints. Performance with and without guidance is comparable for labels where baseline performance is already high, but SAM 2 guidance shows notable improvements for challenging labels such as *Head Up* and *Moving*.

Iteration No.	Labels					
	Head Down	Lying	Standing	Eating	Head Up	Moving
0	0.9151	0.8218	0.8130	0.6692	0.2789	0.4668
5k	0.9153	0.8970	0.9171	0.7635	0.4311	0.3915
10k	0.9152	0.9188	0.9387	0.7697	0.4133	0.3895
25k	0.9170	0.9546	0.9505	0.7737	0.4049	0.3800
50k	0.9191	0.9671	0.9458	0.7628	0.4272	0.4433
100k	0.9205	0.9808	0.9705	0.7811	0.4680	0.4244
250k	0.9207	0.9768	0.9701	0.7843	0.4643	0.4479
500k	0.9208	0.9810	0.9705	0.7879	0.4937	0.4484
1M	**0.9237**	**0.9811**	**0.9755**	**0.7892**	**0.4954**	**0.4840**

Table 2. F1 score versus DINOv2's self-supervised fine-tuning iteration number, label-wise for unguided DINOv2 evaluated on cropped images.

Using window size 7, we train linear probes with temporal information included. We conduct these on the cropped annotated dataset for the guided DINOv2 checkpoints as well as CLIP as a baseline (Fig. 2). Incorporating temporal information provides a consistent performance boost to all models, from the CLIP and DINOv2 baselines to the checkpoints at various iteration numbers for segmentation-guided DINOv2 self-supervised fine-tuning.

4. Additional Qualitative Results

In the main paper, we display the attention maps of various DINOv2 models on an image of a sheep moving. In this supplement, we briefly elaborate on how these maps are

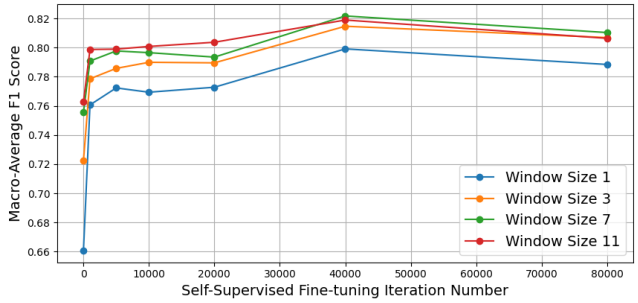


Figure 1. DINOv2 Fine-Tuning w/ SAM 2 Guidance Macro-Average F1 Score vs Temporal Window Size. It seems that the returns are diminishing on increasing window size, and effectively peak at 7 frames. Note that the iteration numbers for the self-supervised fine-tuning on the x-axis are different as these are from an early run to determine relevant parameters.

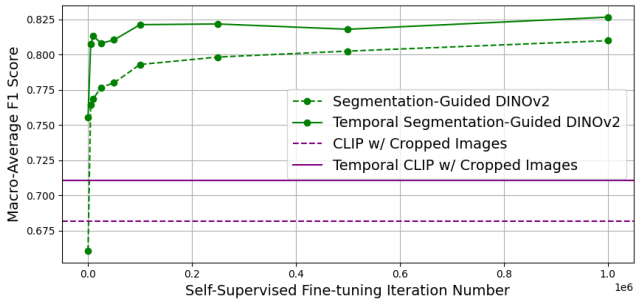


Figure 2. Linear probes on SAM 2 guided DINOv2 fine-tuning with and without temporal information. Temporal information provides consistent performance boosts across baseline CLIP features and various DINOv2 checkpoints, particularly for challenging labels.

visualized and then display attention maps of various DINOv2 models on an image of a sheep sitting, and one of a sheep standing and eating. Additionally, here, we use both the regular version of each image and its cropped version based on SAM-2 guidance, for four sets of attention maps. These can be seen in Figures 3, 4 for cropped images and 4, **??** for uncropped images.

The attention values for 6 heads of the classification token are normalized, then smoothed, and then interpolated to the image size so that it can be overlayed on top of the image. We conducted this process for 3 frames, representing different behaviors: sitting, walking and eating. For the DINOv2 checkpoints - we utilize baseline DINOv2, the best-performing unguided fine-tuned checkpoint, and the best-performing segmentation-guided fine-tuned DINOv2 checkpoint. We display and discuss the results for 3 heads for one of these images in the main paper due to the space constraints. The visualizations for all 6 heads for the images are discussed here and generally display qualitatively similar results.

For the uncropped images, it is noteworthy that while the first vanilla and unguided DINOv2 models are trained or fine-tuned on lots of uncropped images like this, the segmentation-guided DINOv2 is only fine-tuned on cropped versions of images.

In Figure 4 for the sitting image, the visualizations show that:

1. Vanilla DINOv2 is activated at the sheep but also sporadically all over the background.
2. Unguided DINOv2 is activated more at the sheep, but also sometimes at the cage or background objects (heads 2, 3, 5, 6).
3. Segmentation-guided DINOv2 is activated as much at the sheep as unguided, but less so on the background.

In Figure **??** for the standing/eating image, the visualizations show that:

1. Vanilla DINOv2 is activated at the sheep but also at the other edges of the image.
2. Unguided DINOv2 is activated at the sheep, but also heavily on the background.
3. Segmentation-guided DINOv2 is activated at the sheep more than unguided and less on the background than unguided.

Similarly, for the cropped images, the segmentation-guided DINOv2 attention maps are even more focused on the sheep, and the vanilla and unguided DINOv2 attention maps are even less so. In Figure 3 for the sitting image, the visualizations show that:

1. Vanilla and unguided DINOv2 are activated sporadically, and largely on the background.
2. Segmentation-guided DINOv2 is activated much less sporadically and largely on different parts of the sheep.

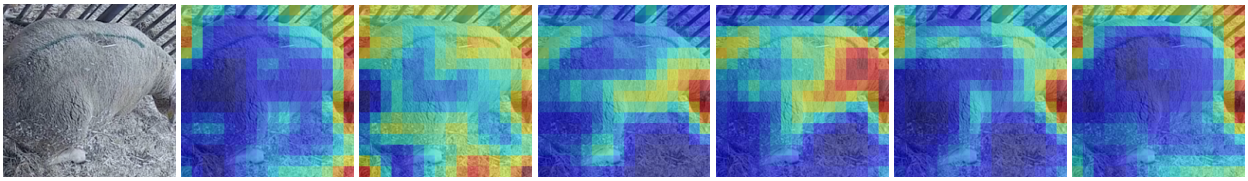
In Figure 4 for the standing/eating image, the visualizations show that:

1. Vanilla DINOv2 is activated sporadically around the sheep.
2. Unguided DINOv2 is activated almost entirely on the background.
3. Segmentation-guided DINOv2 is activated more at the sheep than all prior models.

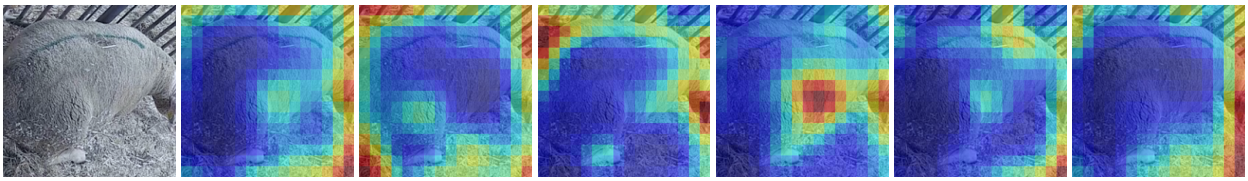
These results support the findings in the main paper that the features learned by segmentation-guided DINOv2 are robust and sheep-centric, and generally better than vanilla or unguided DINOv2. This is especially impressive for uncropped images because of the fact that segmentation-guided DINOv2 is the only model not fine-tuned or trained on uncropped images and yet maintains "good" attention maps out-of-distribution.

5. Additional Training, Evaluation, and Inference Details

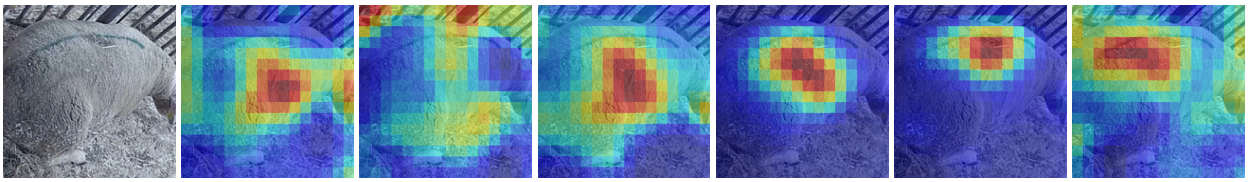
We modify the original code for DINOv2 for our self-supervised fine-tuning. We use the same loss functions and



(a) DINOv2 Baseline Attention Maps for all 6 heads.

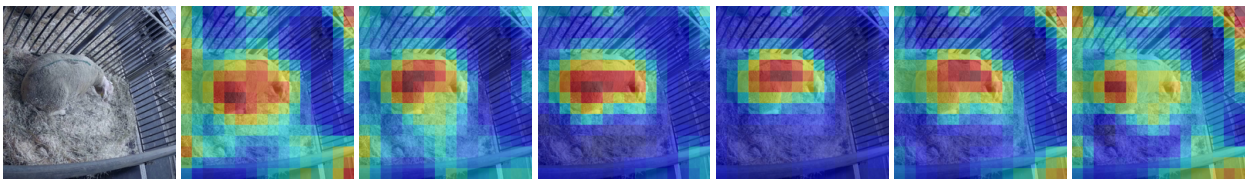


(b) Unguided fine-tuned DINOv2 Attention Maps for all 6 heads.

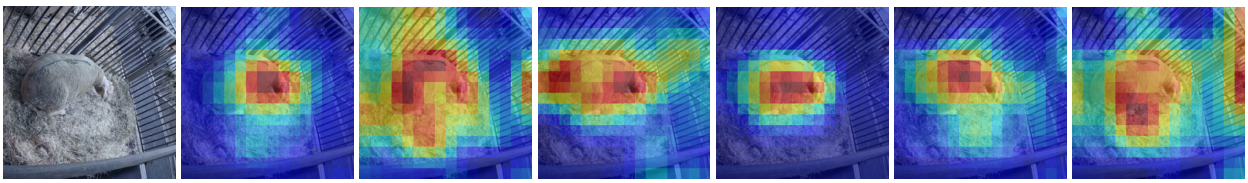


(c) Segmentation-guided fine-tuned DINOv2 Attention Maps for all 6 heads.

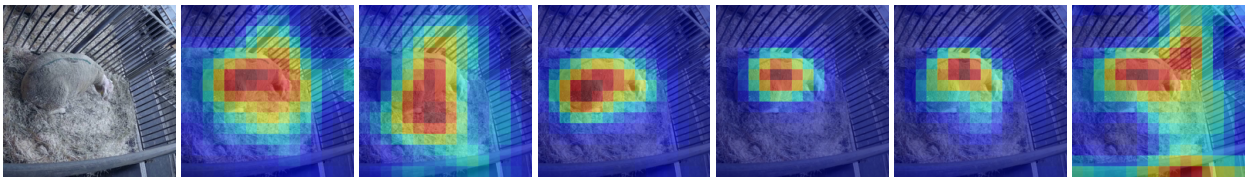
Figure 3. **Attention Map Visualizations for Cropped Sitting image.** (a) attention maps for 6 ViT heads of a pre-trained DINOv2 model, (b) attention maps for 6 ViT heads of a DINOv2 model fine-tuned on the sheep data, (c) attention maps for 6 ViT heads of a pre-trained DINOv2 model fine-tuned on the sheep data using segmentation guidance.



(a) DINOv2 Baseline Attention Maps for all 6 heads.



(b) Unguided fine-tuned DINOv2 Attention Maps for all 6 heads.



(c) Segmentation-guided fine-tuned DINOv2 Attention Maps for all 6 heads.

Figure 4. **Attention Map Visualizations for Uncropped Sitting image.** (a) attention maps for 6 ViT heads of a pre-trained DINOv2 model, (b) attention maps for 6 ViT heads of a DINOv2 model fine-tuned on the sheep data, (c) attention maps for 6 ViT heads of a pre-trained DINOv2 model fine-tuned on the sheep data using segmentation guidance.

weighting as the original DINOv2, with a batch size of 64 on a single RTX 4090 GPU. We initialize from the publicly available ViT-S DINOv2 checkpoints for the student and teacher networks, and randomly initialize the DINO-heads since the weights are not publicly available. We utilize bicubic interpolation to adapt the positional encoding from the final model resolution of 518x518 pixels to the original training resolution of 224x224 pixels, in line with prior work on self-supervised fine-tuning of DINOv2 for domain adaptation [1]. We use the original hyperparameters from training with a base learning rate of 2×10^{-4}. We temporally subsample to 1 in every 10 frames out of the 13 million frames (26 million shots) in the full unannotated dataset to reduce redundancy in the dataset without losing much information. We use an official epoch length of 400 iterations for 2700 epochs, thus covering the entire dataset about 25 times.

For SAM-2 cropping, in the main paper we mentioned that we manually identify sheep locations in the first frame, and then propagate them forward through time. Since SAM-2 loads all frames in a video into memory at once, we cannot load the whole video at once. Instead, we process videos in batches of 505 frames. The last 5 frames in each batch are the first 5 in the next batch, allowing us to use the detected masks as the initial prompts for the next batch, in point and box form. This framework reliably handles situations where there is no sheep detected for most circumstances. The only circumstance in which this framework requires manual intervention is if all 5 of the last frames in a batch have no sheep detected, since then there is no way to prompt the start of the next batch. In this case, one must manually intervene to remove these images or provide a starting prompt for the next batch. However, in our dataset, we find this to be extremely rare. In other datasets where the animals are in-frame less reliably, we might explore more automated ways of handling this exception. This could be as simple as extending the start of the batch backwards in time until a frame where there is an animal, or utilizing SAM-2 automatic mask generation with heuristic tools to identify the animals in a zero-shot manner.

For the linear probe, during training, we apply:

1. A random grayscale transformation with a 0.5 probability, since our dataset includes frames at night which are in grayscale.
2. A random resized crop transformation to the model resolution.
3. A random horizontal flip.

We use SGD as our optimizer with 0.9 momentum and 0 weight decay, and a Cosine Annealing LR scheduler with a 0 minimum learning rate, in line with the setup used for linear evaluation in the original DINOv2 repository. We train for 12 epochs, where each epoch stops after covering 0.25 of the dataset, for 3 effective epochs. This is done since the linear probe can converge very quickly for some labels. We sweep a range of values for learning rate from 1×10^{-5} to 1×10^{-1}, as well as vary whether we use the last or 3rd last layer for feature extraction for DINOv2. The learning rates mentioned are the base learning rates for a batch size of 256, which we scale to our actual batch size of 12. Empirically, we find that the 3rd last layer of DINOv2 has better features for linear probing for this task for vanilla DINOv2, unguided DINOv2, and segmentation-guided DINOv2.

For evaluating the performance of different feature extractors, we take the best label-wise F1 scores achieved, across any run, to represent that feature extractor, since a linear probe in a multi-label classification setting means every label is independent of other labels, and as such each of these runs is equivalent to 10 different runs for each of the labels. Accordingly, at inference time for identifying statistical differences between groups of sheep, we take the best checkpoints from different runs for each of the labels and combine them into one linear probe which we then apply to data.

6. Dataset Details

We split the 12 videos into 9 training and 3 testing videos. The annotations have 10 labels - *Head Down, Lying, Standing, Eating, Head Up, Moving, Regurgitating, Turning Left, Turning Right, Drinking* - in order of frequency. In addition, we also have unannotated videos of 14 sheep over 4 days, 5 hours a day, for a total of 280 hours of footage. Each frame is split into 2 different camera views of the sheep, for a total of 26 million shots. We treat both shots (hereafter referred to as "frames") as separate data points with the same labels. We made NSD publicly available.

References

[1] Benedikt Roth, Valentin Koch, Sophia J. Wagner, Julia A. Schnabel, Carsten Marr, and Tingying Peng. Low-resource finetuning of foundation models beats state-of-the-art in histopathology, 2024. 4