

# ➤ ViT-Cow: Self-Supervised Video Pretraining for Few-Shot Dairy Cow Behavior Recognition

J. Allyndrée<sup>1,2</sup>, M. Charles<sup>1</sup>, M. Frere<sup>1</sup>, D. Glaizal<sup>1</sup>, M. Meier<sup>1</sup>, C. Tibi<sup>1</sup>,  
A. Cornuéjols<sup>1</sup>, C. Martin<sup>1</sup>, A. Madouasse<sup>2</sup>

<sup>1</sup> Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA-PS, 22 place de l'Agronomie, 91120 Palaiseau, France

<sup>2</sup> Oniris, INRAE, BIOEPAR, 101 Rte de Gachet, 44300 Nantes, France

Email : [joseph.allyndree@inrae.fr](mailto:joseph.allyndree@inrae.fr)

**29/06/2026**

# Context

GH010262.MP4 - Annotation Coucher\* - BORIS

Project Observations Playback Tools Analysis Help

Ethogram

| Key | Code | Type        | Description             | Color |
|-----|------|-------------|-------------------------|-------|
| 1   | 4    | Point event | Pose le 1er antérieur   |       |
| 2   | 6    | Point event | Ramène ses postérieurs  |       |
| 3   | 3    | Point event | Plie le 1er antérieur   |       |
| 4   | 1    | Point event | Sent le sol             |       |
| 5   | 5    | Point event | Pose 2nd antérieur      |       |
| 6   | 8    | Point event | Replace ses membres     |       |
| 7   | 2    | Point event | Baisse la tête          |       |
| 8   | 7    | Point event | Laisse tomber sa hanche |       |
| 9   | s    | Start       | Début de vidéo          |       |

Player paused  
Current media name: GH010262.MP4 (#1 / 3)  
Media position: 00:00:01.043 / 00:00:13.222  
frame: 25  
Total: 00:00:01.043 / 00:01:32.466

Events for "GH010262.MP4" observation

| Time | Frame index  | Subject | Code | Type  | Modifier | Comment |
|------|--------------|---------|------|-------|----------|---------|
| 1    | 00:00:01.043 | 25      |      | Start |          |         |

Player #1

2025-02-10 12:32:02pm

No focal subject  
Observed behaviors:

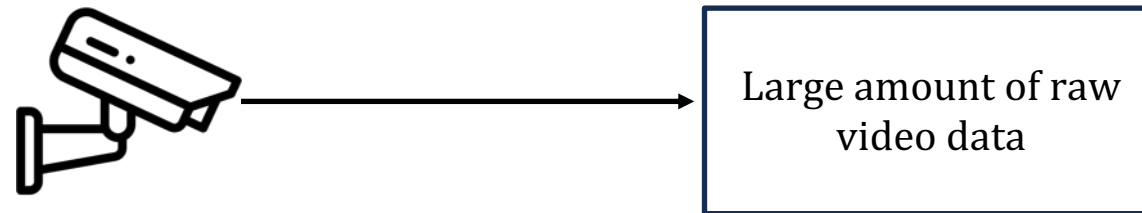


- ➔ Behavior annotation is expensive
- ➔ Labelled videos are scarce
- ➔ Supervised models overfit to farm



# Context

Farm environments usually have CCTV systems



Can we leverage raw cow videos for behavior annotations ?

# Goal of the study



Raw video data

Dataset Creation  
(SAM3.1 + code)



Label : Lying



## Traditional Approach

Behavior  
annotations on a  
subset of subvideos

Small amount of  
labelled  
subvideos

YOLO based model for  
behavior classification

Large amount of  
unannotated  
subvideos

Pre-training VideoMAE  
on pixel reconstruction  
pretext task

VideoMAE fine-tuned

## Self-Supervised Learning

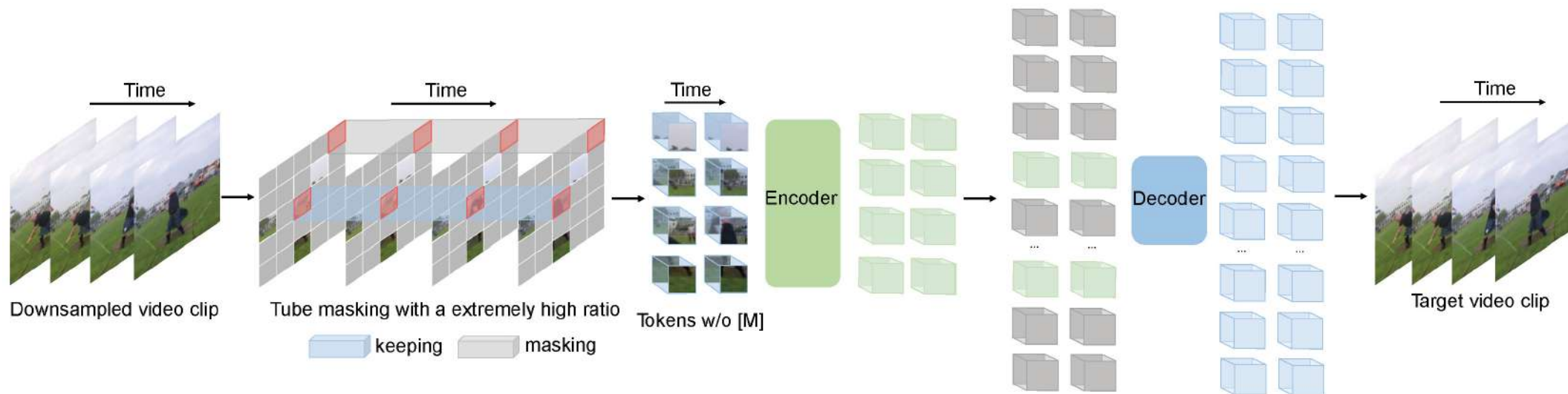


Carion, N. et al., 2026. SAM 3: Segment Anything with Concepts.

Tong, Z., Song, Y., Wang, J., Wang, L., 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training.

# Self-Supervised Learning for CV tasks

## VideoMAE architecture – Tube masking as a pretext task for pixel reconstruction



Tong, Z., Song, Y., Wang, J., Wang, L., 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training.

# Data presentation



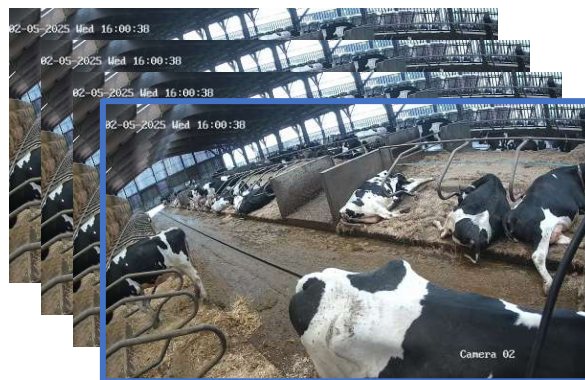
x 5 dairy farms  
(total of 376  
days of videos)

# Dataset construction

1 hour video (12fps)



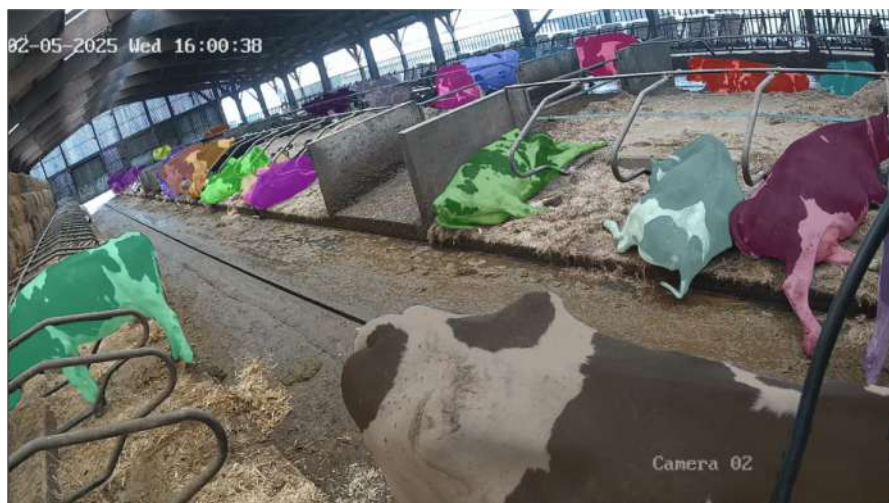
20 random clips of 64 frames



subsample clips (3fps)



segment and track with SAM3.1



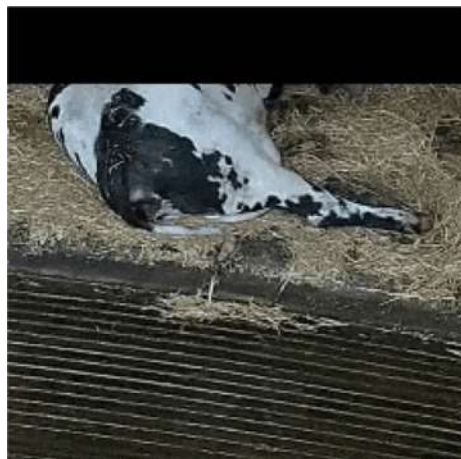
isolate and extract subvideos



standardize to 224x224



## Instances of subvideos



Total of 71,284 subvideos

# Ethogram and labelled subvideos

## Ethogram of the classes of the behaviors developed in the study

| Behavior      | Description                                                                             |
|---------------|-----------------------------------------------------------------------------------------|
| Standing      | Cow is standing up with the four legs fully extended on the floor                       |
| Lying         | Cow is lying down with the body in contact with the floor and the four limbs beneath it |
| Walking       | Cow is moving with at least one leg lifted off the ground.                              |
| Feeding       | Cow is head in the feeding rack, either head down ingesting food or head up ruminating  |
| Drinking      | Cow is head in the water trough, head down ingesting water                              |
| Scratching    | Cow is actively scratching its head on a part of the barn                               |
| Self-grooming | Cow is licking itself, with the head in contact with its own body                       |
| Other         | Any other exhibited behavior not included in the present ethogram                       |

Table 6: Definition of the cow postures and behaviors used in the study. Posture behaviors are indicated in blue, they can happen simultaneously with other behaviors.

## Labelled subvideos

| Class         | Support | Prop. (%) |
|---------------|---------|-----------|
| Standing      | 1116    | 29.2      |
| Lying         | 1564    | 40.9      |
| Walking       | 169     | 4.4       |
| Feeding       | 846     | 22.1      |
| Drinking      | 29      | 0.8       |
| Self-grooming | 30      | 0.8       |
| Scratching    | 52      | 1.4       |
| Other         | 17      | 0.4       |
| Total         | 3823    | 100       |

# Results

| Class           | Support | YOLOv8 (baseline) |       |       | $\Delta$ VideoMAE (pretrained) |            |             |
|-----------------|---------|-------------------|-------|-------|--------------------------------|------------|-------------|
|                 |         | P                 | R     | f1    | $\Delta P$                     | $\Delta R$ | $\Delta f1$ |
| Drinking        | 29      | 0.222             | 0.138 | 0.170 | −0.222                         | −0.138     | −0.170      |
| Feeding         | 846     | 0.827             | 0.881 | 0.853 | −0.038                         | +0.035     | −0.005      |
| Lying           | 1564    | 0.949             | 0.930 | 0.940 | +0.010                         | −0.011     | −0.002      |
| Other           | 17      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       |
| Scratching      | 52      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       |
| Self-grooming   | 30      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       |
| Standing        | 1116    | 0.710             | 0.780 | 0.743 | +0.031                         | +0.040     | +0.036      |
| Walking         | 169     | 0.393             | 0.260 | 0.313 | +0.496                         | +0.308     | +0.380      |
| <b>Macro</b>    | —       | 0.388             | 0.374 | 0.377 | +0.034                         | +0.029     | +0.030      |
| <b>Micro</b>    | —       | 0.816             | 0.816 | 0.816 | +0.027                         | +0.027     | +0.027      |
| <b>Weighted</b> | —       | 0.797             | 0.816 | 0.805 | +0.026                         | +0.027     | +0.024      |

# Results

| Class           | Support | YOLOv8 (baseline) |       |       | $\Delta$ VideoMAE (pretrained) |            |             |
|-----------------|---------|-------------------|-------|-------|--------------------------------|------------|-------------|
|                 |         | P                 | R     | f1    | $\Delta P$                     | $\Delta R$ | $\Delta f1$ |
| Drinking        | 29      | 0.222             | 0.138 | 0.170 | -0.222                         | -0.138     | -0.170      |
| Feeding         | 846     | 0.827             | 0.881 | 0.853 | -0.038                         | +0.035     | -0.005      |
| Lying           | 1564    | 0.949             | 0.930 | 0.940 | +0.010                         | -0.011     | -0.002      |
| Other           | 17      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       |
| Scratching      | 52      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       |
| Self-grooming   | 30      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       |
| Standing        | 1116    | 0.710             | 0.780 | 0.743 | +0.031                         | +0.040     | +0.036      |
| Walking         | 169     | 0.393             | 0.260 | 0.313 | +0.496                         | +0.308     | +0.380      |
| <b>Macro</b>    | —       | 0.388             | 0.374 | 0.377 | +0.034                         | +0.029     | +0.030      |
| <b>Micro</b>    | —       | 0.816             | 0.816 | 0.816 | +0.027                         | +0.027     | +0.027      |
| <b>Weighted</b> | —       | 0.797             | 0.816 | 0.805 | +0.026                         | +0.027     | +0.024      |

→ Better classification on highly temporal behaviors (Standing, Walking)

# Results

| Class           | Support | YOLOv8 (baseline) |       |       | $\Delta$ VideoMAE (pretrained) |            |             | Class         | Support | Prop. (%) |
|-----------------|---------|-------------------|-------|-------|--------------------------------|------------|-------------|---------------|---------|-----------|
|                 |         | P                 | R     | f1    | $\Delta P$                     | $\Delta R$ | $\Delta f1$ |               |         |           |
| Drinking        | 29      | 0.222             | 0.138 | 0.170 | -0.222                         | -0.138     | -0.170      | Standing      | 1116    | 29.2      |
| Feeding         | 846     | 0.827             | 0.881 | 0.853 | -0.038                         | +0.035     | -0.005      | Lying         | 1564    | 40.9      |
| Lying           | 1564    | 0.949             | 0.930 | 0.940 | +0.010                         | -0.011     | -0.002      | Walking       | 169     | 4.4       |
| Other           | 17      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       | Feeding       | 846     | 22.1      |
| Scratching      | 52      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       | Drinking      | 29      | 0.8       |
| Self-grooming   | 30      | 0.000             | 0.000 | 0.000 | 0.000                          | 0.000      | 0.000       | Self-grooming | 30      | 0.8       |
| Standing        | 1116    | 0.710             | 0.780 | 0.743 | +0.031                         | +0.040     | +0.036      | Scratching    | 52      | 1.4       |
| Walking         | 169     | 0.393             | 0.260 | 0.313 | +0.496                         | +0.308     | +0.380      | Other         | 17      | 0.4       |
| <b>Macro</b>    | —       | 0.388             | 0.374 | 0.377 | +0.034                         | +0.029     | +0.030      |               |         |           |
| <b>Micro</b>    | —       | 0.816             | 0.816 | 0.816 | +0.027                         | +0.027     | +0.027      |               |         |           |
| <b>Weighted</b> | —       | 0.797             | 0.816 | 0.805 | +0.026                         | +0.027     | +0.024      |               |         |           |

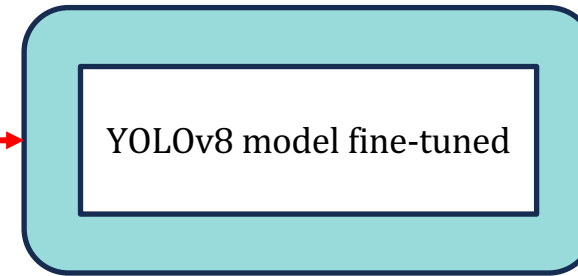
➔ Poor performances on low support classes

# Take home message

## Self-Supervised Learning



## Traditional Approach



- ➔ Very similar results with both models (SSL does slightly better)
- ➔ Big difference on the Walking behavior : high movements behaviors are better predicted
- ➔ Still need a big amount of annotations to achieve good classification

## Acknowledgements

This work was supported by a state grant managed by the French National Research Agency under the France 2030 program, reference ANR-22-PEAE-0008.

The authors would also like to especially thanks the farmers that allowed data collection on their farms.

---

Another talk I will be giving tomorrow (Session 7) : “Evaluating standing and lying behaviors measured with collar-mounted accelerometers for lameness detection in dairy cows”



Research page



INRAE

2<sup>nd</sup> EAAP Conference on Artificial Intelligence 4 Animal Science Ghent - Belgium  
20260629 – ViTCow – Presented by Joseph Allyndrée

AgroParisTech

Oniris  
VetAgroBio Nantes  
ÉCOLE NATIONALE

p. 14

Any questions ?

