

Supplementary Information

A modular open-source computer vision pipeline enables individual-level behavior analysis of group-housed pigs

Haiyu Yang¹, Enhong Liu¹, Jennifer Sun¹, Sumit Sharma¹, Meike van Leerdam¹, Sebastien Franceschini², Puchun Niu¹, and Miel Hostens¹

¹Department of Animal Science, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY 14853, USA

²University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium

S1 Considerations and Trials for Deciding Model

In order to identify the best model, we conducted some evaluations of several open-source models. Two distinct video datasets were used: a proprietary recording of our own dairy cows and the publicly available Edinburgh Pig Behavior Video¹.

S1.1 Object Detection and Localization

Aside from the OWLv2 model, we also experimented with the YOLOv12 model. YOLOv12 represents a significant advancement in the YOLO family by integrating an innovative area attention (A2) mechanism that overcomes the computational limitations of traditional attention approaches, which typically scale quadratically with input size and become impractical for high-resolution agricultural imagery². By dividing feature maps into distinct areas and computing attention locally, A2 reduces complexity to linear time, enabling efficient processing of large images while preserving global context. This attention-centric, real-time object detection framework outperforms previous YOLO versions and other detectors across all model scales (N, S, M, L, X). YOLOv12-L achieves a state-of-the-art 53.7% mAP, 0.4% higher than YOLOv11-L, with superior object localization and background suppression as illustrated in Figure S1.

In our implementation, YOLOv12 performed zero-shot detection on initial video frames within approximately 15 seconds, generating bounding boxes for all detected objects; however, filtering is often necessary to exclude non-target items (like water troughs or feeding equipment commonly present in environments such as dairy barns).

We also experimented with other zero-shot object detection and localization models such as OWLv2³, but on the cattle footage its performance was not as good as YOLO's, as shown in Figure S2. OWLv2 produced more false predictions and required additional adjustments of the confidence threshold to isolate useful detections.

However, the relative performance of the two models reverses when applied to the Edinburgh Pig Behavior Video from Bergamini et al.¹.

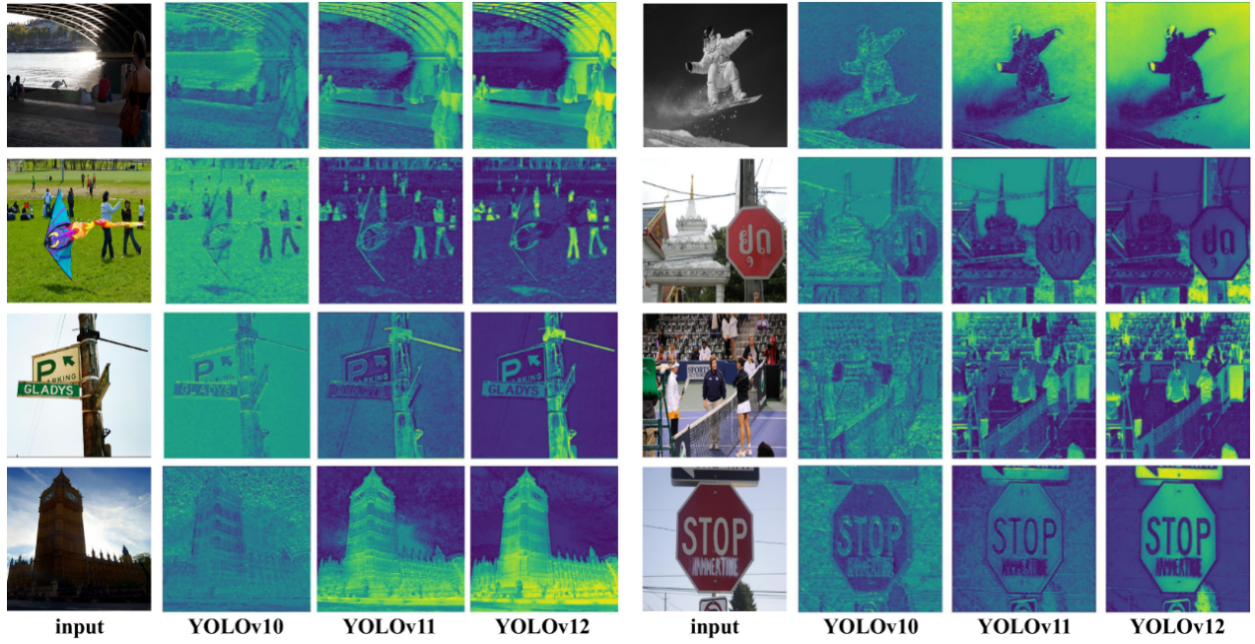


Figure S1. Comparative visualization of attention heat maps across YOLO versions, demonstrating YOLOv12's superior object localization and background suppression². YOLOv12 exhibits a better ability to extract object contours from the same input images.



Figure S2. Bounding boxes overlaid with predictions from YOLOv12 (left) showing successful detections of cows within the frame, and OWLv2³ detection output (right) demonstrating false positive generation at default confidence settings.



Figure S3. Detection results on overhead pig housing footage showing YOLOv12’s underdetection (left) and OWLv2’s complete coverage (right) at 0.5 confidence threshold. YOLOv12 detected three of the eight pigs visible in the frame, demonstrating substantial underdetection. In contrast, OWLv2 achieved complete coverage by identifying all eight target pigs and even two additional, highly occluded pigs from an adjacent pen.

S1.2 Segmentation and Tracking

We also experimented with other video segmentation and tracking models such as XMem⁴, but its performance was not as strong and may struggle in deployment within complex barn environments.



Figure S4. Comparative segmentation performance on occluded cattle showing SAMURAI’s precise instance separation (middle) versus XMem’s mask merging (right). The original frame is shown on the left.

With the SAM2 video predictor (Figure S4, middle), the cows are clearly separated, whereas the XMem⁴ video predictor generated a single mask covering three overlapping cows, failing to produce precise mask predictions in challenging scenes.

S1.3 Choice of Individual Modules

For the object detection and localization component, initial experiments with YOLOv12 demonstrated promising efficiency and accuracy characteristics (see Figure S3); however, the model failed to generate reliable predictions when applied to pig detection tasks. Consequently, we adopted OWLv2 as our primary object detection and localization model. OWLv2 exhibited robust performance in our pig detection experiments, demonstrating its suitability for this specific agricultural application.

The selection of the segmentation and tracking module was based on comparative evaluation between SAMURAI and XMem models. SAMURAI demonstrated superior segmentation capabilities, producing more precise object boundaries and maintaining clearer separation between individual animals, particularly in challenging scenarios involving overlapping subjects (Figure S4). The model’s ability to preserve distinct contours even during close animal interactions proved critical

for accurate individual tracking. These performance advantages led to the selection of SAMURAI as the segmentation and tracking component within our pipeline.

For the feature extraction backbone, we evaluated both DINOv2 and CLIP models to determine their suitability for visual behavior analysis. As the current application does not require cross-modal text-image understanding, and given that DINOv2 demonstrated marginally superior performance in visual feature discrimination (Figure S5), we selected DINOv2 as our feature extraction model. The t-SNE visualizations revealed more distinct behavioral clusters with DINOv2 embeddings, suggesting enhanced discriminative power for behavior classification tasks.

S2 Validation on CBVD-5

S2.1 CBVD-5 Dataset Description

The CBVD-5 dataset from Li et al.⁵ served as a collection of dairy cow behavior keyframes with corresponding annotations, enabling us to evaluate the cross-species generalizability of our framework. The dataset originates from continuous video surveillance of dairy cows’ daily activities on a commercial farm, with cameras operating 24 hours per day over a five-day period. This extensive monitoring period ensured comprehensive coverage of various behaviors across different times of day and environmental conditions.

From continuous video footage, keyframes were systematically selected for behavior annotation, resulting in 27,501 valid labeled data points. Each annotation contained precise spatial information about individual cow locations paired with one of five behavioral categories: standing, lying down, feeding, drinking, and rumination, capturing the primary activities essential for monitoring dairy cow welfare and productivity.

Given that the dataset provided sparse annotations across multiple cows without persistent individual identifiers, it presents different evaluation opportunities compared to the pig dataset. Consequently, we do not utilize this dataset for benchmarking our object detection, localization, tracking, and segmentation modules, which require consistent individual identification across frames. Instead, the primary evaluation focus centers on assessing the feature extraction capabilities of our framework and the classifier’s ability to differentiate between behavioral patterns based on these extracted features.

For our evaluation, we focused on the standing and lying-down categories, which provide a basic test of the feature extraction capability of our pipeline. This decision was motivated by two factors: the subtle nature of rumination behavior requires exceptionally high-quality video for accurate assessment, and the feeding and drinking classes exhibited significant sample imbalance that could bias the evaluation results. After filtering to these two primary behaviors, we obtained 30,391 individual cropped samples suitable for analysis, providing a substantial dataset for evaluating behavioral classification performance in dairy cattle.

S2.2 Dataset Preparation

We applied our automated cropping module to extract individual animal images from the CBVD-5 dataset annotations using the bounding boxes provided along with the pictures. The existing ground truth bounding boxes were utilized directly.

S2.3 CLIP Implementation

CLIP⁶ learns joint representations of images and text by training on hundreds of millions of image–text pairs scraped from the internet. The model employs two parallel encoders—a Vision Transformer for images and a Transformer for text—that are trained simultaneously to produce embeddings in a shared multimodal space.

The training objective uses a contrastive loss that maximizes the cosine similarity between matching image–text pairs while minimizing similarity between non-matching pairs. Given a batch of N image–text pairs, CLIP computes pairwise cosine similarities:

$$S_{ij} = \frac{I_i \cdot T_j}{\|I_i\| \|T_j\|} \quad (\text{S1})$$

where I_i and T_j denote the image and text embeddings, respectively.

These similarities are scaled with a temperature parameter τ :

$$L_{ij} = \tau \cdot S_{ij} \quad (\text{S2})$$

CLIP optimizes the model to maximize the similarity scores for matching pairs while minimizing scores for non-matching pairs. The image-to-text loss is:

$$\mathcal{L}_i = -\log \frac{\exp(L_{ii})}{\sum_{j=1}^N \exp(L_{ij})} \quad (\text{S3})$$

The text-to-image loss is:

$$\mathcal{L}_t = -\log \frac{\exp(L_{ii})}{\sum_{j=1}^N \exp(L_{ji})} \quad (\text{S4})$$

The total loss is the average of both directions:

$$\mathcal{L} = \frac{1}{2} (\mathcal{L}_i + \mathcal{L}_t) \quad (\text{S5})$$

This bidirectional loss ensures that both the image and text encoders learn complementary representations. CLIP can perform zero-shot classification by comparing an input image to text descriptions of different classes. The model computes the similarity between the image embedding and text embeddings for phrases like “a photo of a [class]” and predicts the class with the highest similarity score.

Computing the image embedding:

$$I_{\text{test}} = I(\text{test image}) \quad (\text{S6})$$

Computing text embeddings for each class:

$$T_c = T(\text{“a photo of a \{class\}”}) \quad (\text{S7})$$

Predicting the class with maximum similarity:

$$\hat{y} = \arg \max_c \frac{I_{\text{test}} \cdot T_c}{\|I_{\text{test}}\| \|T_c\|} \quad (\text{S8})$$

This ability to generalize to new tasks without additional training makes CLIP particularly versatile for various computer vision applications. To get more versatile features, CLIP-ViT-Large-Patch14 was selected for its strong zero-shot accuracy.

The implementation followed a similar pipeline to DINOv2 but includes additional capabilities for text-based queries, enabling zero-shot behavior classification without fine-tuning. However, in our experiments, we used these models primarily for feature extraction rather than zero-shot classification, as we trained supervised classifiers on the extracted embeddings.

S2.4 Feature Extraction Comparison on Cows

We compared DINOv2 and CLIP embeddings using t-SNE visualization to evaluate their discriminative power for behavior classification.

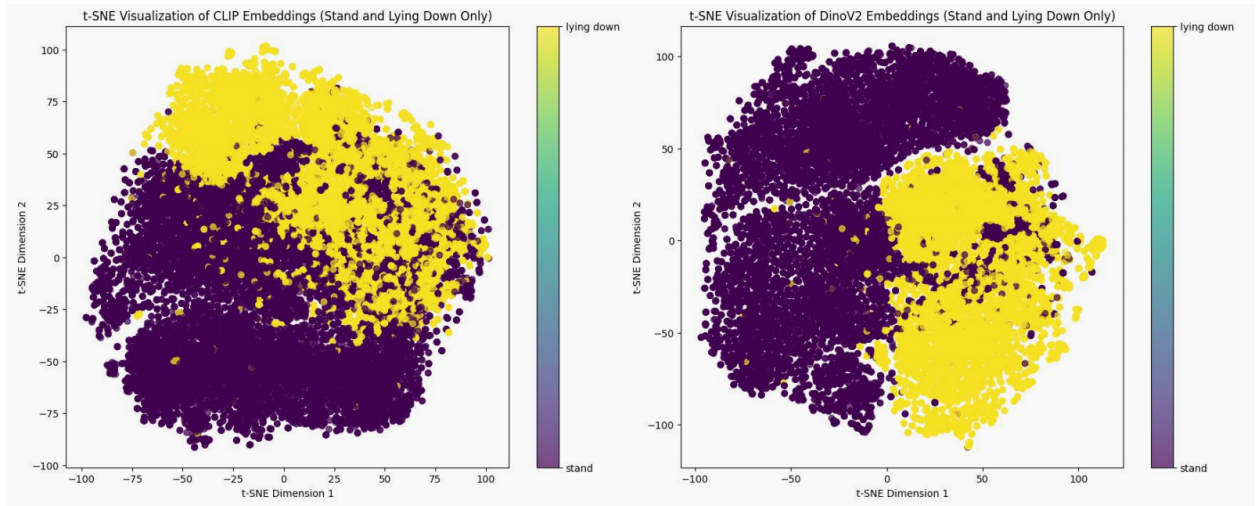


Figure S5. t-SNE visualization of CLIP embeddings (left) and DINOv2 embeddings (right) for cattle behaviors (standing vs. lying down). DINOv2 shows distinct cluster separation, while CLIP embeddings reveal overlapping behavioral clusters.

The visualizations reveal that DINOv2 produces more distinct clusters for different behaviors, while CLIP embeddings show moderate separation with overlapping regions.

S2.5 Classification of Cow Behaviors

We evaluated MLP classifiers on the extracted features from both DINOv2 and CLIP models to assess their discriminative power for behavior classification, with training typically converging within 3–5 epochs.

Table S1. Binary classification performance of MLP models using DINOv2 and CLIP features for standing versus lying behavior detection of dairy cows.

Model	Feature Type	Accuracy	Precision	Recall	F1-Score
MLP	DINOv2	98.3%	0.982	0.982	0.982
MLP	CLIP	98.2%	0.981	0.983	0.982

Both feature extraction methods achieved high performance, with nearly identical results (Table S1). The MLP classifier using DINOv2 features achieved 98.3% accuracy with a test loss of 0.0645, while the CLIP-based classifier achieved 98.2% accuracy with a slightly lower test loss of 0.0565.

Detailed performance metrics by class reveal consistent performance across both behaviors (Table S2).

The high performance of both feature extraction methods indicates that our framework successfully captures discriminative features for behavior classification. The marginal difference between DINOv2 and CLIP suggests that both approaches effectively encode behavioral information,

Table S2. Class-specific performance metrics for standing and lying behavior classification using DINOv2 and CLIP features.

Feature Extractor	Behavior	Precision	Recall	F1-Score	Support
DINOv2	Standing	0.987	0.984	0.985	3022
	Lying	0.977	0.980	0.978	2042
CLIP	Standing	0.990	0.980	0.985	3022
	Lying	0.971	0.985	0.978	2042

with DINOv2 showing a slight edge in overall accuracy. The balanced performance across classes demonstrates the robustness of our approach for identifying both standing and lying behaviors with minimal bias toward either category.

S3 Glossary

A2 Area Attention

AI Artificial Intelligence

CBVD-5

Cow Behavior Video Dataset (5 categories)

CLS

Class token (in transformers)

CLIP

Contrastive Language-Image Pre-training

CNN

Convolutional Neural Network

CPU

Central Processing Unit

DINOv2

Self-Distillation with NO labels, version 2

FPS

Frames Per Second

GB Gigabyte

GPU

Graphics Processing Unit

I/O Input/Output

IoU Intersection over Union

LSTM

Long Short-Term Memory

mAP

mean Average Precision

MLP

Multi-Layer Perceptron

NVMe
Non-Volatile Memory Express

OWLv2
Open-World Localization Vision model, version 2

RAM
Random Access Memory

ReLU
Rectified Linear Unit

RGB
Red, Green, Blue (color model)

SAM
Segment Anything Model

SAM 2
Segment Anything Model, version 2

SAMURAI
Segment Anything Model Upgraded for Real-time Adaptation and Inference

SSD
Solid State Drive

t-SNE
t-distributed Stochastic Neighbor Embedding

V100
NVIDIA Volta 100 GPU

ViT Vision Transformer

XMem
Extended Memory (video segmentation model)

YOLO
You Only Look Once

References

- [1] L. Bergamini, S. Pini, A. Simoni, R. Vezzani, S. Calderara, R. B. D’Eath, and R. B. Fisher. Extracting accurate long-term behavior changes from a large pig dataset. In *16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pages 524–533. SciTePress, 2021.
- [2] Y. Tian, Q. Ye, and D. Doermann. YOLOv12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.
- [3] M. Minderer, A. Gritsenko, and N. Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.
- [4] H. K. Cheng and A. G. Schwing. XMem: Long-term video object segmentation with an Atkinson–Shiffrin memory model. In *European Conference on Computer Vision (ECCV)*, pages 640–658. Springer Nature Switzerland, 2022.

- [5] K. Li, D. Fan, H. Wu, and A. Zhao. A new dataset for video-based cow behavior recognition. *Scientific Reports*, 14(1):18702, 2024.
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.