From Vision to Vocabulary: A Multimodal Approach to Detect and Track Black Cattle Behaviors

Su Myat Noe[1], Thi Thi Zin[2], Pyke Tin[2] and Ikuo Kobayashi[3]
[1]Interdisciplinary Graduate School of Agriculture and Engineering,
[2]Graduate School of Engineering,
[3]Faculty of Agriculture,
University of Miyazaki
1-1 Gakuen Kibanadai-Nishi, Miyazaki 889-2192, Japan
Email: {z319701@student, thithi@cc, ikuokob@cc}.miyazaki-u.ac.jp,
pytetin11@gmail.com

Abstract

This paper investigates the potential of recent image-text foundation models for classifying black cattle mounting behavior without fine-tuning. Our approach begins with the detection and tracking of each individual black cattle using deep learning-based, fine-tuned YOLOv9 and Deep OC SORT tracking. Once completed, we employ zero-shot approaches, explicitly utilizing the multi-modal Large Language and Vision Alignment (LLAVA) and Large Language Model Meta AI (LLaMA) models. These models integrate visual and linguistic information seamlessly, enabling us to leverage pre-trained knowledge to analyze and understand black cattle behavior directly from images and accompanying text descriptions. By utilizing zero-shot learning, we can bypass the resource-intensive process of model fine-tuning, making it a highly efficient approach for behavior classification. Our approach highlights the robustness and flexibility of multimodal foundation models like LLAVA and LLaMA in handling complex tasks in the agricultural domain, demonstrating their potential for broader applications without requiring extensive retraining on specific datasets. Through our experiments, we showcase the accuracy and efficiency of this zero-shot multimodal approach, providing valuable insights into black cattle mounting behavior that can enhance livestock management and monitoring practices.

Keywords: Black cattle detection and tracking, Deep OC SORT, Large Language Model MetaAI (LLaMA), Large Language and Vision Alignment (LLAVA), Large Language Models (LLMs), Vision-Language Models (VLMs), YOLOv9

1 Introduction

Promoting animal welfare relies heavily on accurately understanding animal behavior. One essential aspect of this understanding is the manual labeling of behaviors, which is crucial for analysis. Typically, this involves describing animal postures and behaviors in detail using semantic sentences and assigning behavioral labels to short video clips. This method is inherently complex and time-consuming.

Recent advancements in artificial intelligence have seen significant integration of Large Language Models (LLMs) [1] and Vision-Language Models (VLMs) [2] into automated behavior recognition. These models have shown exceptional ability in

bridging the gap between visual data and textual annotations, facilitating more nuanced and accurate behavioral analysis across various species, including livestock. Vision-language models have been pivotal in advancing behavior recognition, particularly those influenced by architectures such as CLIP (Contrastive Language Image Pre-training) [3]. These models utilize a contrastive learning framework that aligns images with relevant textual descriptions, thereby learning to recognize behaviors from visual cues linked to descriptive labels. The effectiveness of VLMs has been demonstrated in several studies where models trained on generic datasets exhibit remarkable transfer learning capabilities when applied to specific tasks like black cattle behavior recognition in naturalistic settings. In addition to monitoring species like black cattle, our approach extends to broader ecological studies by employing foundation models' zero-shot learning capabilities. Furthermore, our method addresses the significant challenges associated with traditional monitoring methods, which are often labor-intensive and limited by the need for manual data annotation and the high computational and energy costs of training deep learning models. By incorporating zero-shot learning techniques, our framework reduces the dependency on large, annotated datasets and computational resources, making it suitable for applications in resource-constrained environments such as remote wildlife areas and fields like autonomous agriculture and robotic behavior analysis.

This paper introduces a comprehensive approach to black cattle monitoring by employing advanced computer vision techniques that optimize segmentation and behavioral analysis with minimal disturbance to their natural habitats [4]. Our work contributes by implementing a deep learning framework that integrates a fine-tuned YOLOv9 [5] model for segmentation. This integration is designed to tackle the challenge of detecting black cattle in complex backgrounds. Additionally, we combine this framework, the Deep OC SORT model for tracking black cattle [6], and the Large Language and Vision Alignment (LLAVA) and Large Language Model Meta AI (LLaMA) model to classify cattle mounting behavior. This approach enhances the accuracy and robustness of behavior classification in challenging visual contexts, particularly in complex scenes where black cattle may be difficult to distinguish from their surroundings. The complete proposed system is described in following Figure 1.

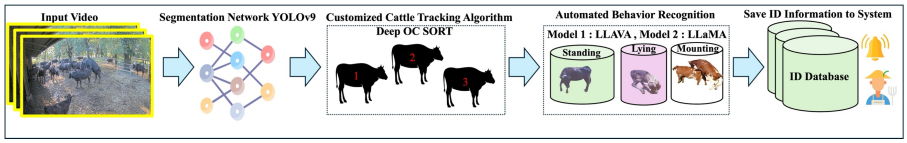


Fig. 1. Proposed Framework: Integrating Finetuned YOLOv9 for detection, Deep OC SORT for tracking, and LLAVA Model for black cattle behavior classification.

The key contributions of this paper are:

(1) Fine-tuned the LLAVA model's parameters to enhance its performance in recognizing and classifying specific cattle behaviors, such as mounting.
(2) Created a query-based system within LLAVA for behavior analysis, allowing for interactive querying of cattle behaviors in video sequences. We designed specific prompts and queries to extract and describe cattle behaviors from video inputs accurately.

(3) Comparing with the LLaMA model will help us determine the visual prompt answer and which model is more appropriate for the precision agriculture field.

2 Methods

2.1 Dataset Pre-processing and Detection Method

We compiled a specialized dataset focusing on black cattle to develop and evaluate our proposed framework. We carefully selected five videos for testing and validation, of which two were to include mounting behavior [7]. To monitor the activities of 55 cattle, a camera was positioned above the pen, capturing footage from a top-corner view at a resolution of 1920 x 1080 pixels and a frame rate of 30 FPS. Video data was continuously collected, but we used footage showing mounting behavior to detect and recognize this activity. This dataset was explicitly curated to fine-tune the YOLOv9 model and carefully validate our system. It is important to note that we refrained from conducting any further training, instead leveraging the fine-tuned YOLOv9 weights parameters for accurate cattle detection without additional training. Our proposed approach, illustrated in Figure 1, employs an efficient detection system using a fine-tuned YOLOv9 model on our custom cattle dataset to ensure precise detection of cattle in video frames. In this stage, YOLOv9 identifies and outlines cattle within the footage by generating accurate bounding boxes around each detected animal.

2.2 Customized Tracking Method with Deep OC SORT

The customized black cattle tracking algorithm with Deep OC-SORT combines the strengths of Deep SORT and OC-SORT to ensure robust tracking, particularly tailored for black cattle. The process begins with initializing the model, leveraging Deep SORT's deep learning-based appearance features and OC-SORT's ability to handle occlusions for consistent tracking. Detection of black cattle in each video frame is achieved using the fine-tuned YOLOv9 model, which generates bounding boxes around each animal. These bounding boxes serve as inputs for the tracking algorithm. Deep appearance features are extracted for each detected cattle using a convolutional neural network (CNN). These features are crucial for distinguishing individual black cattle and maintaining consistent identities across frames. The algorithm utilizes the extracted appearance features, bounding boxes, and motion information (e.g., velocity, trajectory) to match detected black cattle between consecutive frames. This matching process relies on both appearance similarity and spatial proximity. Deep OC-SORT effectively handles occlusions by utilizing overlap consistency to reassign the correct identity to cattle that temporarily disappear and reappear, ensuring accurate tracking even in crowded or complex scenes. After matching, the algorithm updates the tracks for each black cattle, adjusting bounding boxes and appearance features based on new detections. Tracks that remain unmatched for a predefined number of frames are terminated to prevent false positives. The final output is a set of continuous tracks for each detected black cattle, providing detailed information on their movements and interactions over time. This information can be used for further analysis, such as behavior understanding and trajectory prediction, enabling effective monitoring of black cattle in video frames, even under challenging conditions.

2.3 Zero-shot Vision-Language Models and Visual Prompts

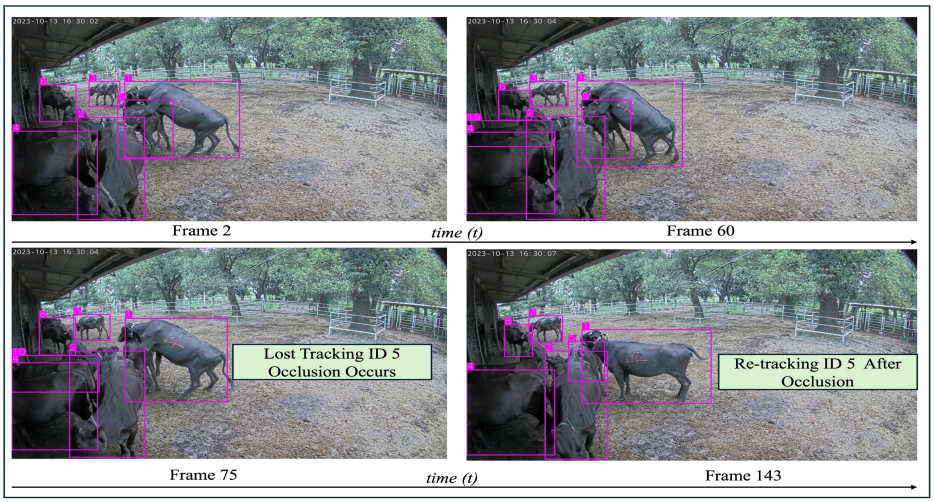
Vision Language Models (VLMs) are multi-model deep learning models that simultaneously process image and text data. Typically trained on extensive datasets comprising billions of samples, these models are highly effective in zero-shot tasks due to their comprehensive training. We select the Large Language and Vision Alignment (LLAVA) model, which excels at the image captioning task by integrating visual and linguistic information to generate precise and contextually relevant descriptions. The process begins with input preprocessing, where an image is acquired and standardized through resizing, normalization, and sometimes augmentation. This preprocessed image is input into a vision encoder, such as a Convolutional Neural Network (CNN) or Vision Transformer (ViT) [8], which extracts high-dimensional feature representations capturing essential visual details like objects, textures, colors, and spatial relationships. The output is a feature map that encapsulates the core visual information of the image.

Next, a cross-modal attention mechanism aligns these visual features with the language model, ensuring the visual information is contextually integrated into the text generation process. The visual features are transformed into embeddings compatible with the language model, often combined with positional encodings to retain spatial context. The language model, typically a transformer-based architecture like GPT or BERT, processes these embeddings to generate coherent and contextually appropriate textual descriptions. This involves predicting a sequence of words that accurately describe the image content. Training the LLAVA model involves using large-scale datasets, such as MS COCO, which provide paired images and text descriptions. During inference, the trained LLAVA model generates captions for new input images, like the training phase, but without ground truth supervision. We engage in a Visual Question-answering task for the multimodal Large Language Models (LLMs). We prompt the model with the question: "What are they doing?" and retrieve scores for each predicted response.

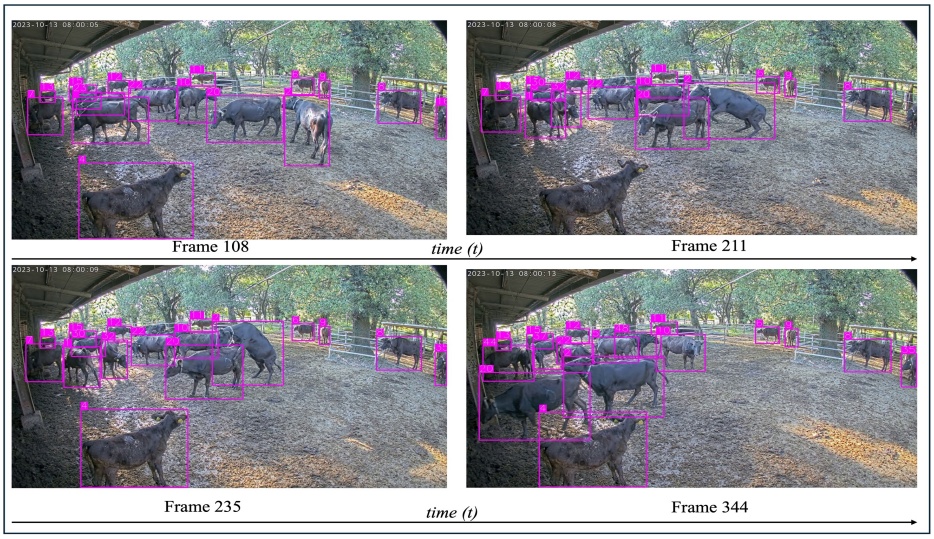
3 Experimental Results

The findings from this study underscore the efficiency of YOLOv9 and Deep OC SORT in recognizing black cattle behavior. Additionally, we integrated the performance of the LLAVA and LLaMA models in zero-shot behavior analysis to validate our comprehensive approach further. YOLOv9 demonstrated outstanding performance in detecting cattle across diverse environmental settings, achieving an exceptional mAP50 score of 0.9856. This high level of object detection accuracy is crucial for the precise localization of cattle within video frames, ensuring reliable monitoring.

In Figure 2, the experimental results depict the tracking accuracy of mounting behavior. Mounting in cattle involves one animal mounting another, typically lasting only 2 to 10 seconds. During this process, it is crucial to detect the behavior, especially at its peak accurately.



(a) IDs 1 and 5 are engaged in mounting behavior. Although tracking is momentarily lost for ID 1 and 5 in frame 75 due to occlusion, it is re-established correctly by frame 143.



(b) IDs 2 and 20 are involved in mounting behavior, and it is observed that they maintain their respective IDs without switching even after the behavior is completed.

Fig. 2. (a, b). The experimental results of the detection achieved by the fine-tuned YOLOv9 model and the tracking performed by Deep OC SORT demonstrate the accurate monitoring of mounting behavior.

Employing the Multiple Object Tracking Accuracy (MOTA) metrics in our research facilitated a comprehensive evaluation of our tracking approach's effectiveness

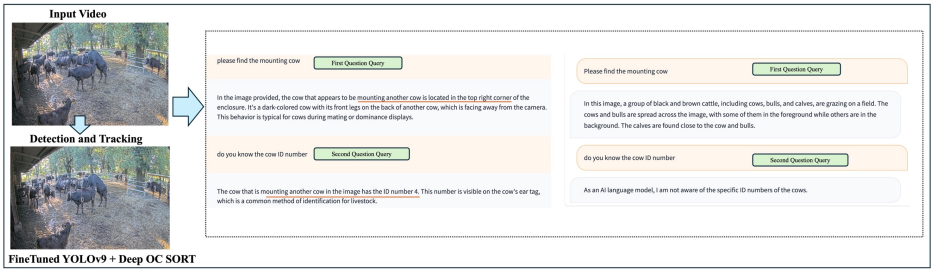
on the black cattle dataset, as detailed in Table 1. Each video represents multiple cameras operating simultaneously for 30 minutes at 30 frames per second. The total inference time for all videos is 25 minutes, which runs parallel with multi-processing, with an average accuracy of 94.79%, calculated using equation (1) of the MOTA calculation, which considers False Negative (FN), False Positives (FP), ID switches (IDSW), and Ground truth detection (gDet) information.

$$MOTA = 1 - \frac{(|FN|+|FP|+|IDSW|}{|gDet|} \quad (1)$$

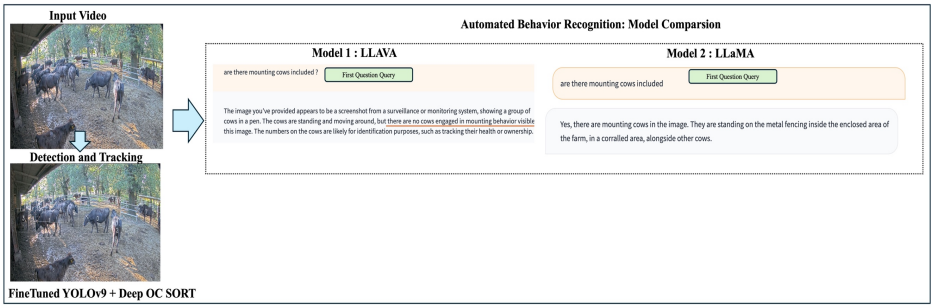
Table 1. Comparison of tracking accuracy

Video Sequences	Video 1	Video 2	Video 3	Video 4	Video 5
Accuracy (%)	92.49	95.73	96.92	96.23	92. 57
Average Accuracy (%)	94.79				

We compared two models for the behavior action recognition results: LLAVA and LLaMA. The findings show that LLAVA (Model 1) can accurately identify cattle behavior more precisely than LLaMA. For detailed experimental results, please refer to the following.

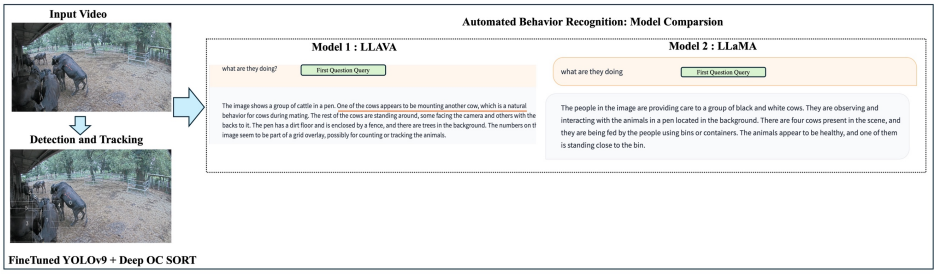


(a) Input the mounting behavior; the first query aims to find the mounting behavior in the ranch. LLAVA correctly answers the first query: "Mounting another cow is located in the top corner." The second query follows, asking for the ID number, and LLAVA accurately responds with "ID number 4," which is the correct answer. Despite extensive testing with the same questions, Model 2: LLaMA, consistently fails to provide accurate answers.

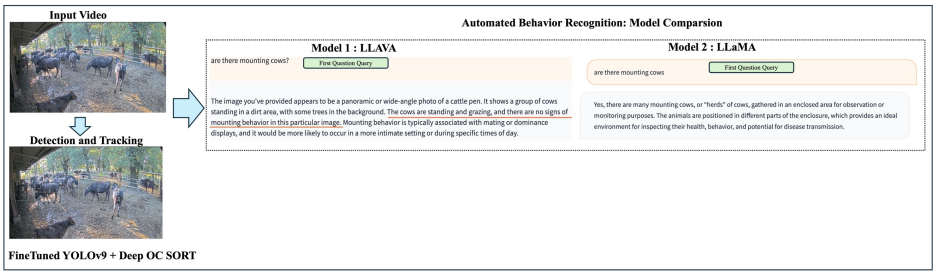


(b) Input the video of a standing cow only, and the first query asks whether any mounting behavior is included. LLAVA correctly responds, "No mounting behavior is included in this video," which is the accurate answer. Despite extensive testing with the

same questions, Model 2: LLaMA consistently fails to provide accurate answers. Although the input images do not include mounting cows, it incorrectly identifies them as present.



(c) The input video depicts mounting behavior among cows, and the first query asks, "What are they doing?" LLAVA correctly answers that one of the cows appears to be mounting another cow in the video. Additionally, it accurately describes the background situation of the cattle ranch. Model 2: LLAMA incorrectly identifies people's presence in images where none exist, providing a generic response instead.



(d) The input video depicts the cattle's standing and grazing behavior. The first query asks, "Are mounting cows included in that scene?" LLAVA correctly responds that the cows are standing and grazing, with no mounting behavior observed. n Model 2: LLAMA, it wrongly classifies behavior even when no mounting cows are included in the images.

Fig. 3. (a,b,c,d) Results of the LLAVA and LLaMA automated behavior recognition system showing accurate identification and description of complex cattle behaviors. The system advanced in vision and language modalities, demonstrating significant improvements over traditional methods.

4 Conclusion and Future Work

In our current research phase, we have developed a comprehensive framework for cattle behavior analysis and tracking, leveraging a customized combination of YOLOv9 and Deep OC SORT for accurate monitoring. This framework integrates object detection and tracking with behavior recognition through LLAVA and LLaMA, enabling real-time analysis of cattle behaviors. We introduced a novel cattle dataset tailored

for this purpose, achieving high detection accuracy mAP: 0.9856 % with our fine-tuned YOLOv9 model and an average accuracy of 94.79% for the five videos.

Our future efforts will enhance the framework's efficiency and adaptability to various environmental conditions. We plan to explore model compression techniques for deployment on mobile and edge devices and investigate adaptive algorithms to improve robustness and long-term effectiveness. These advancements broaden the framework's utility in diverse real-world agricultural settings, contributing to precision agriculture and ecological research initiatives. Ultimately, our framework seeks to promote environmentally friendly farming practices and enhance animal welfare.

References

[1] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N. and Mian, A., 2023. A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

[2] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. and Ring, R., 2022. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35, pp.23716-23736.

[3] Zhang, H., Li, X. and Bing, L., 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

[4] Myat Noe, S., Zin, T.T., Tin, P. and Kobayashi, I., 2023. Comparing state-of-the-art deep learning algorithms for the automated detection and tracking of black cattle. Sensors, 23(1), p.532.

[5] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. arXiv preprint arXiv:2402.13616(2024).

[6] Maggiolino, G., Ahmad, A., Cao, J. and Kitani, K., 2023, October. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In 2023 IEEE International Conference on Image Processing (ICIP) (pp. 3025-3029). IEEE.

[7] Noe, S.M., Zin, T.T., Tin, P. and Kobayashi, I., 2022. Automatic detection and tracking of mounting behavior in cattle using a deep learning-based instance segmentation model. Int. J. Innov. Comput. Inf. Control, 18(1), pp.211-220.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual Instruction Tuning." arXiv preprint arXiv:2304.08485 (2023).