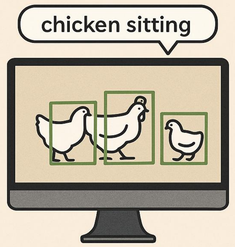
Zero-Shot Detection and Action Recognition of Broiler Chickens Using Vision-Language Models

Ahmad Amirivojdan
Department of Biosystems Engineering and Soil Science

Motivation

Traditional monitoring relies on human observation, which is limited in scale and consistency.



Vision-Language Models offer a scalable approach by interpreting farm scenes using text prompts like “chicken sitting” or “chicken walking”.

In this project, we explore the **zero-shot** capabilities of vision-language models for **detecting chickens** and **recognizing their actions** in farm images, aiming to evaluate their practicality for **automated poultry monitoring** without task-specific training.

Visual Question Answering

Prompt

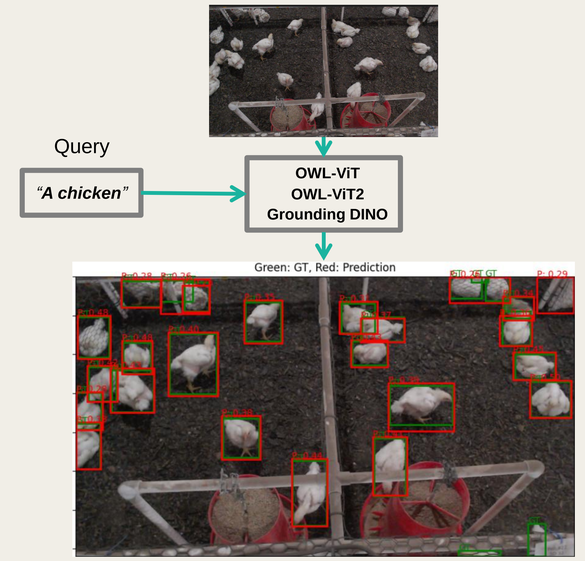
You are a chicken behavior expert. You will be given an image of a chicken and a description of its actions. Look closely at the chicken's posture and its interaction with objects in the scene to determine the correct action. In only one word, what is the chicken doing in the image? Choose from: **drinking**, **feeding**, **preening**, **wing-flapping** or if the image is not clear, say "**unknown**".

Models

(Gemma, Llama, etc.)

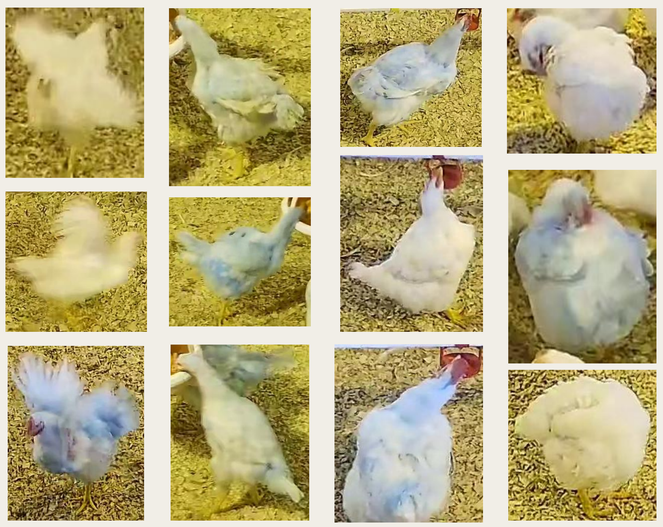
- **Drinking**
- **Feeding**
- **Wing Flapping**
- **Preening**

Object Detection



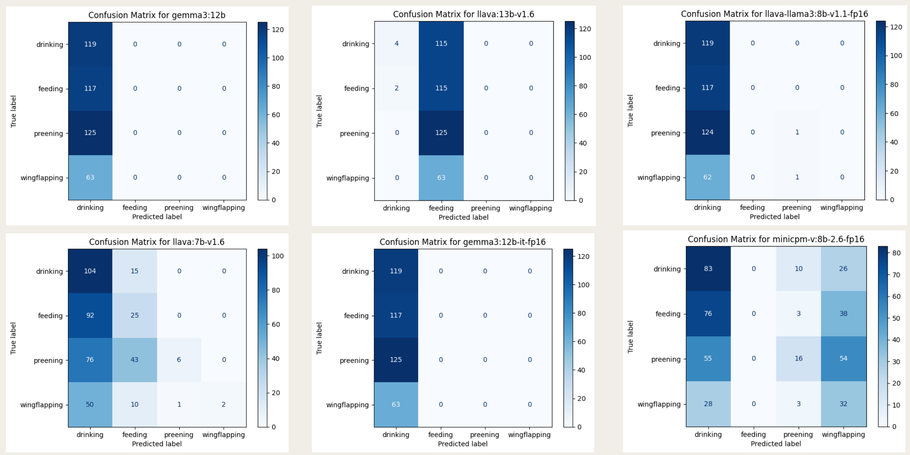
Data Preparation

- 103 images were collected and annotated in COCO format from the JRTU poultry facility.
- A total of 424 cropped instances were labeled across four action classes: 119 drinking, 117 feeding, 125 preening, and 63 wingflapping.



Results

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
gemma3:4b-it-fp16	0.28	0.24	0.25	0.14
minicpm-v:8b-2.6-fp16	0.31	0.26	0.33	0.24
llava-llama3:8b-v1.1-fp16	0.28	0.2	0.25	0.11
gemma3:12b-it-fp16	0.28	0.07	0.25	0.11
llava:7b-v1.6	0.32	0.61	0.29	0.22
llava:13b-v1.6	0.28	0.24	0.25	0.12
gemma3:12b	0.28	0.07	0.25	0.11



Model	AP@[.5:.95]	AP@.5	AP@.75
OWL-ViT	0.24	0.49	0.22
OWL-ViT 2	**0.64**	**0.93**	**0.74**
Grounding Dino	0.50	0.82	0.65

- On object detection task Owl-ViT2 shows the highest AP

Conclusion and Future Works

- Object detection with OWL-ViT 2 is practical and reliable for identifying chickens.
- Action recognition via VQA remains limited; most models show low accuracy and F1.
- Use higher-resolution images to improve detection and action understanding.
- Apply finetuning on domain-specific chicken behavior datasets.
- Explore one-shot or few-shot learning to boost VQA performance with minimal labels.

References

- Minderer, Matthias, et al. "Simple open-vocabulary object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.
- Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

THE UNIVERSITY OF TENNESSEE KNOXVILLE