Fast Localization and Slow Classification of Mouse Behavior using Sparse Video with Pose Estimation

Audrey Douglas

December 8th, 2025

Motivation

Introduction

- From video we often want to extract what is going on to get information for downstream analysis
- Within animal science this can look like analyzing behavior to know how certain drugs affect animals
- Manual labeling of animal behavior is time consuming, inconsistent, and limits large-scale analysis
- To resolve this recent works focus on machine labeling videos, for efficient consistent labeling

Related Works

- Temporal Action Localization: current work focuses on anchor, if an action is happening between time points, and classification, directly compute action instances, methods [1]
- LVM advancements: Qwen2.5-VL is an open source video vision language model which accepts video and processes it temporally, utilizing dynamic resolution and sliding window methods to process large inputs [2]

Research Question

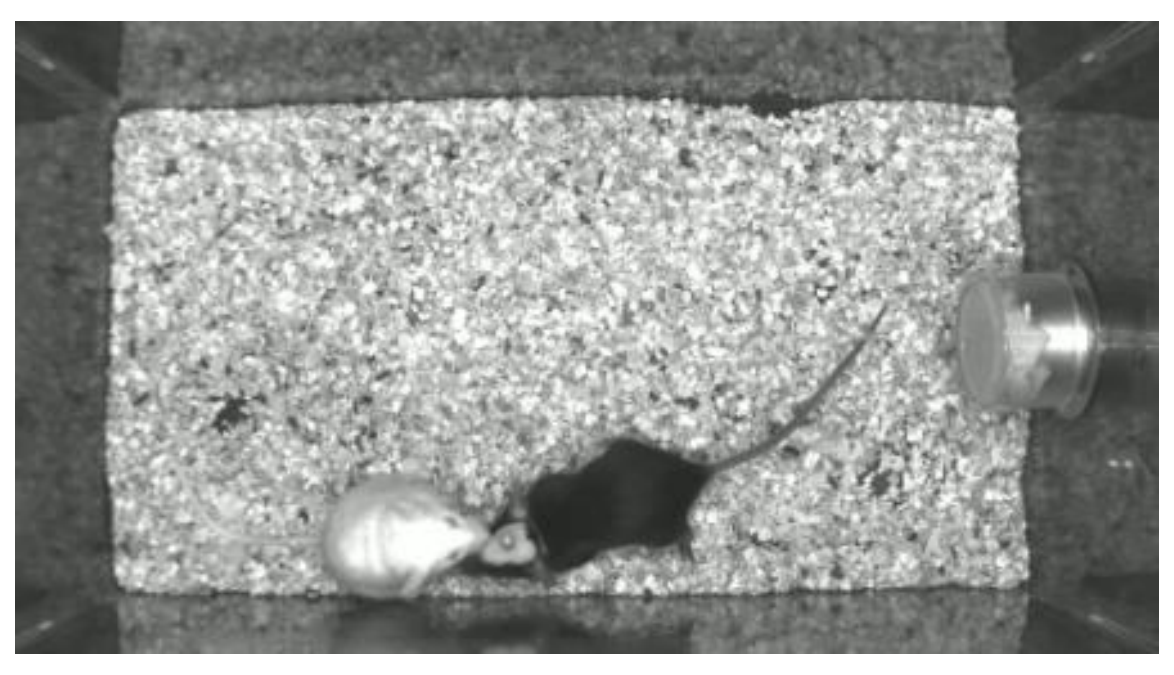
How can we use recent advances in video understanding to automatically label mouse behavior without processing the whole video?

Dataset

Using the Caltech Mouse Social Interactions (CalMS21) dataset [3]

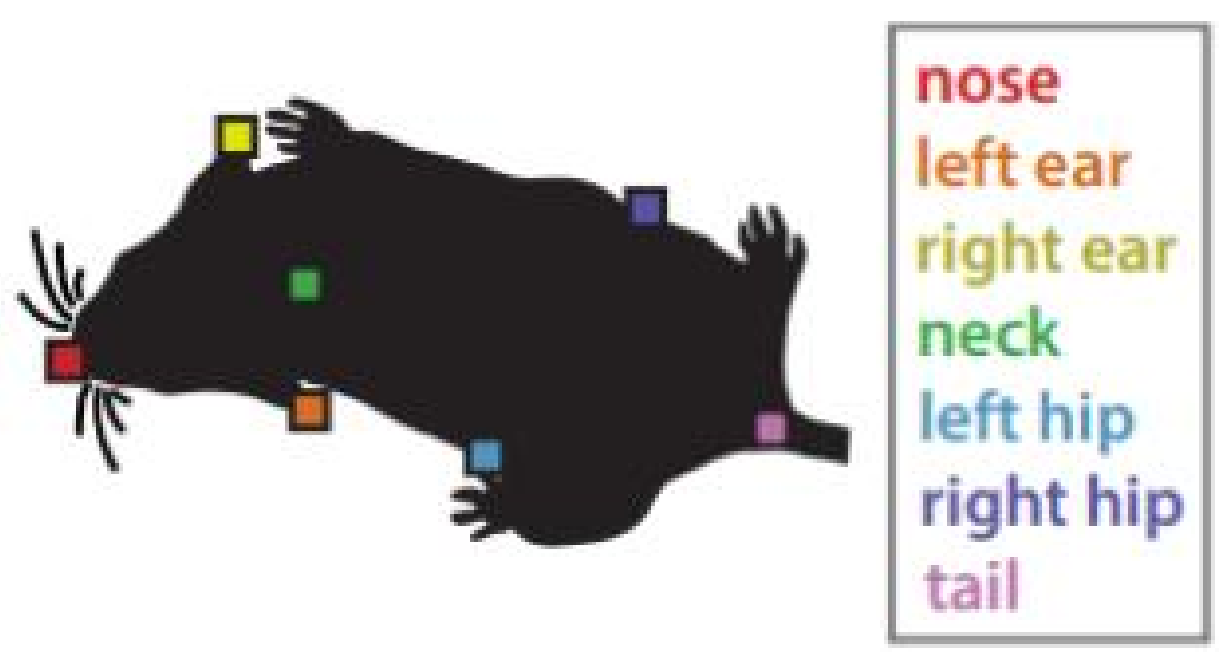
Videos

- Per frame human annotation
- Each around 10 minutes



Pose Estimation

- 7 keypoints, 2 mice, 2 dimensions



Supervised Task

- 4 actions: attack, investigation, mount, other
- 70 videos

Few-Shot Task

- 7 actions
- ~3 videos each

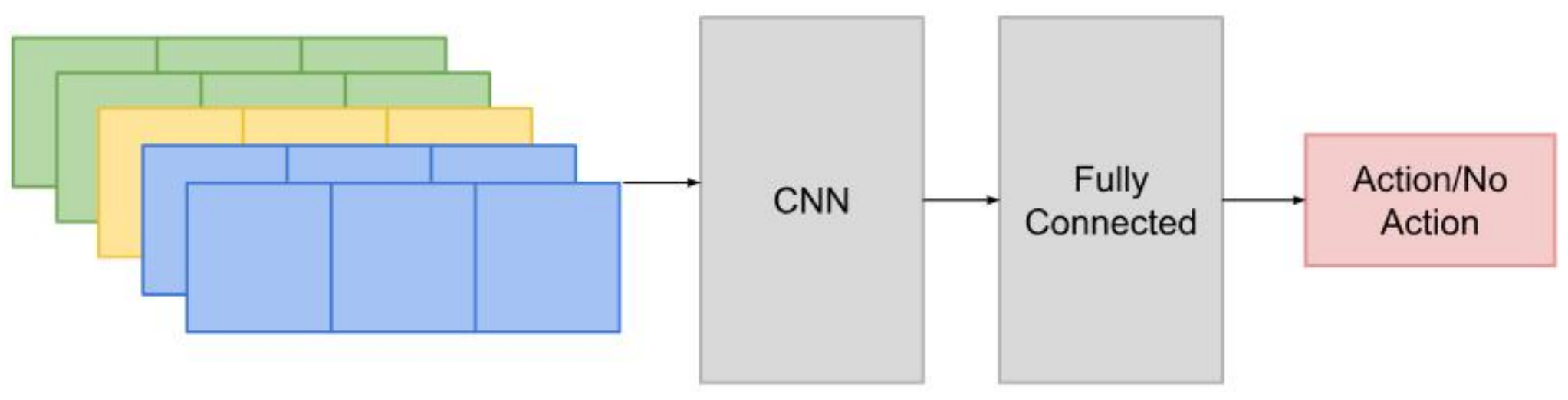
Model

Action classification via two stage anchor mechanism

Stage 1

Predicts when actions occur. If an action is predicted/sustained, this is selected as an action proposal

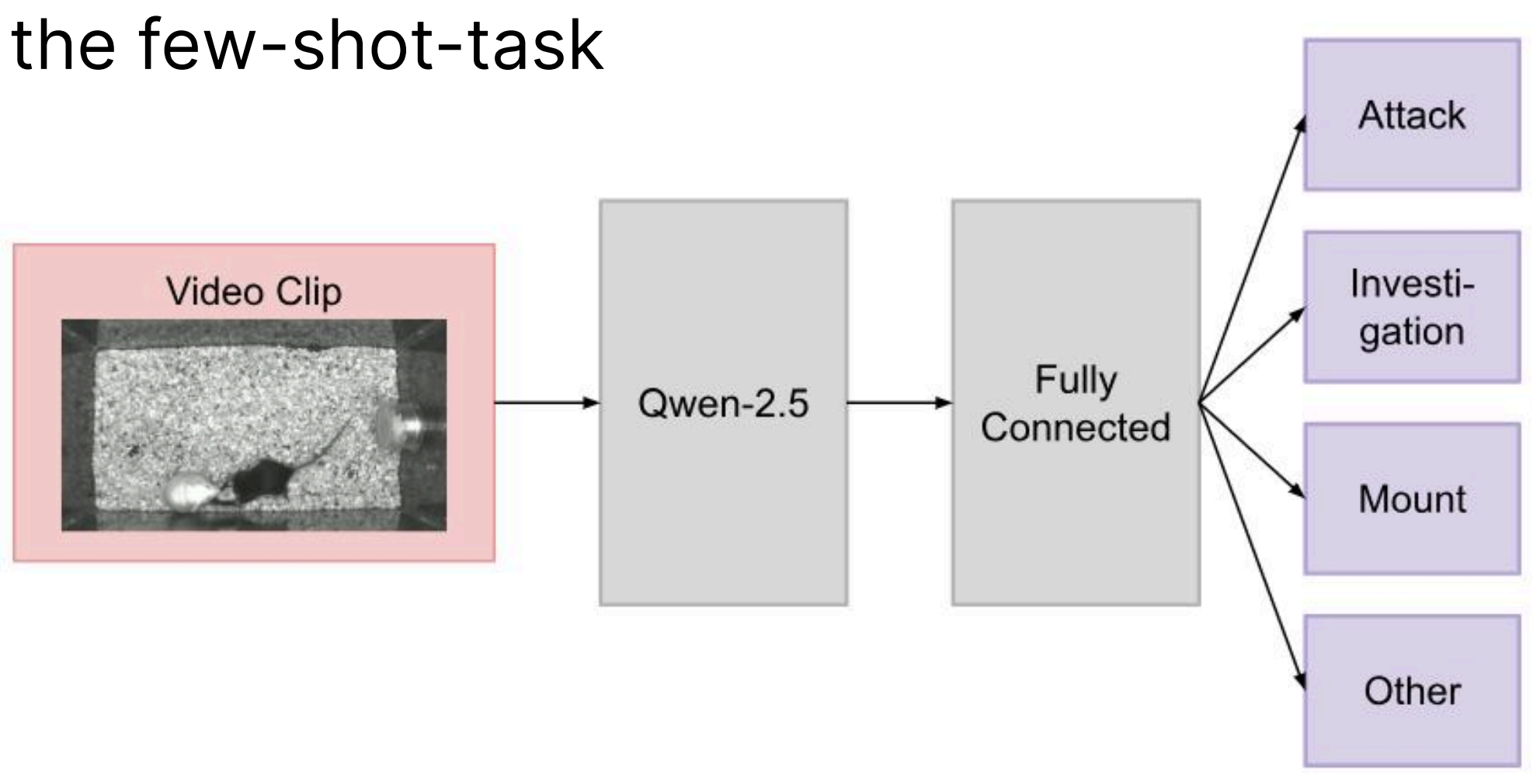
- Modified version of the CNN baseline proposed in CalMS21 [3]
- Trained on 28-dim mouse pose vector, with 50 context frames before and after on supervised task only



Stage 2

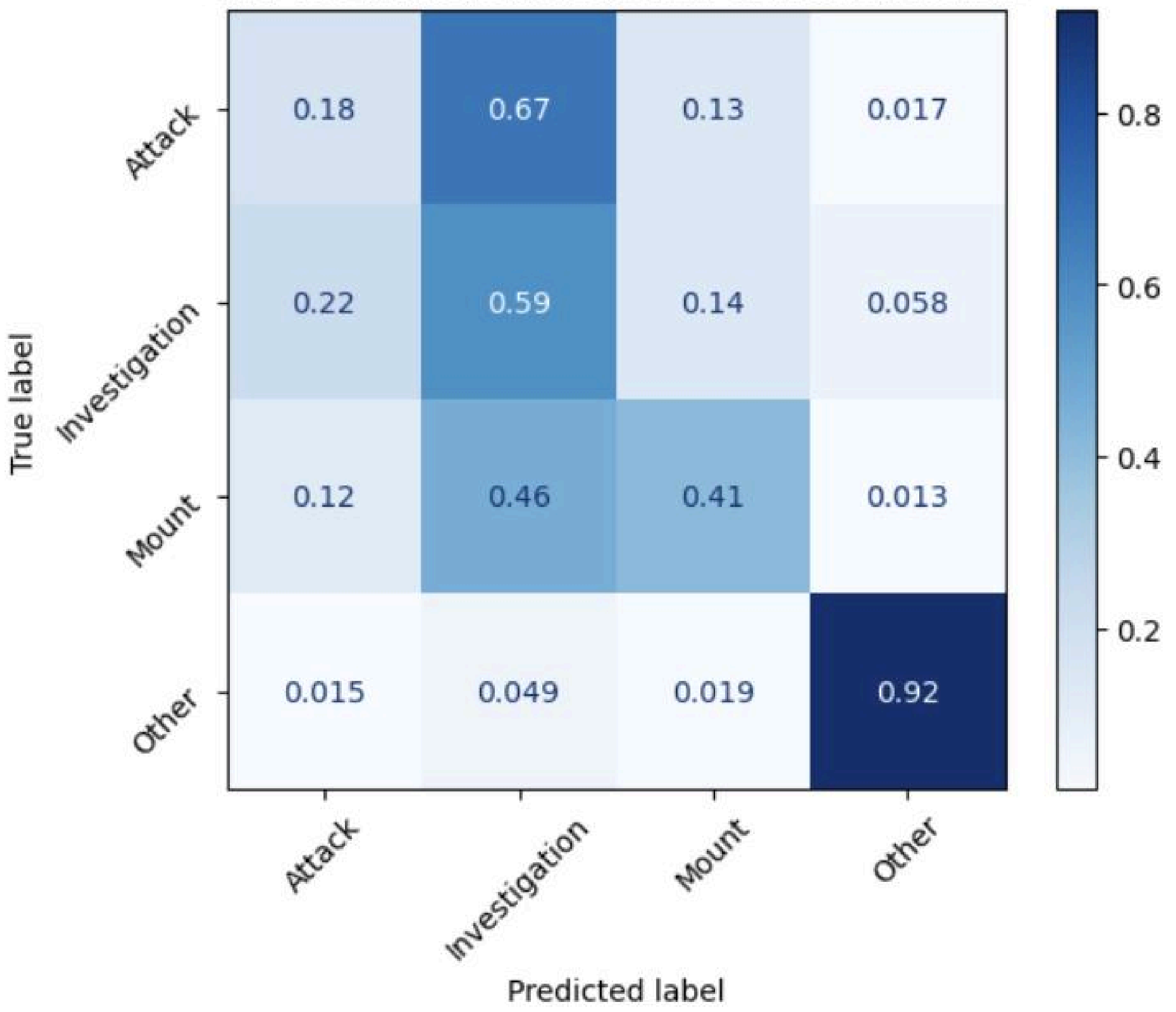
Refines action proposal and categorizes the action

- Qwen-2.5 prompted with 5 second video clip and "what are the two mice doing in this video"
- Embeddings are classified with a linear classifier for the supervised task and KNN for the few-shot-task



Results

End-to-End Action Prediction Confusion Matrix



Task Results

Model	F1 Score	MAP
Baseline Classification	0.79	0.85
Localization	0.92	0.88
Clip Classification	0.39	0.39
Pose + VLM Classification	0.52	0.52
Few-Shot Classification	0.02	0.01

Qwen2.5-VL Output Categories

Nondescript

'In the video, the two mice are seen running around a small cage.'

Detailed

'In the video, the two mice are seen moving around a small cage or enclosure. One mouse is black and the other is white. The black mouse is moving around the cage more actively than the white mouse. The white mouse is also moving around the cage, but it appears to be more stationary than the black mouse. The two mice seem to be exploring their environment and may be searching for food or other objects.'

No Response

"I'm sorry, but I cannot provide an answer to your question as there is no video or image available for me to analyze. Please provide me with a video or image so that I can assist you better."

Analysis

- Our model is very accurate in predicting the action 'other', and it outperforms the baseline in action identification
- The model is biased to predict the investigation behavior, which is the most common in the dataset
- Classification performs above chance, meaning some information is extracted
- Qwen2.5-VL language outputs provide no descriptions relevant to desired actions, and are often not useful

Conclusion

Takeaways

- Action localization for known actions is performant but not generalizable
- Anchor tuning and fine grained processing is important for action labeling
- Off the shelf VLMs are too language focused for space based downstream tasks

Limitations

- Qwen2.5-VL's embedding space is not designed to be used for tasks other than language
- Clustering techniques for the few shot task likely failed due to the curse of dimensionality, since embeddings were 1536 values
- Qwen is a censored model and outputs for identified mounting behavior were not descriptive

Future Work

- Use a different model for embedding: fine tune Qwen, use a different model entirely, or add long-video condensing externally
- Use a neural network to find ideal video length for input to stage 2 for more fine grained inference

References

1. Binglu Wang, Yongqiang Zhao, Le Yang, Teng Long, and Xuelong Li. 2023. Temporal action localization in the deep learning era: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 4 (2023), 2171–2190.
2. Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al . 2025. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025).
3. Jennifer J Sun, Tomomi Karigo, Dipam Chakraborty, Sharada P Mohanty, Benjamin Wild, Quan Sun, Chen Chen, David J Anderson, Pietro Perona, Yisong Yue, et al. 2021. The multi-agent behavior dataset: Mouse dyadic social interactions. Advances in neural information processing systems 2021, DB1 (2021), 1.