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

Audrey Douglas

Problem

We want to understand and process multi-agent behavior

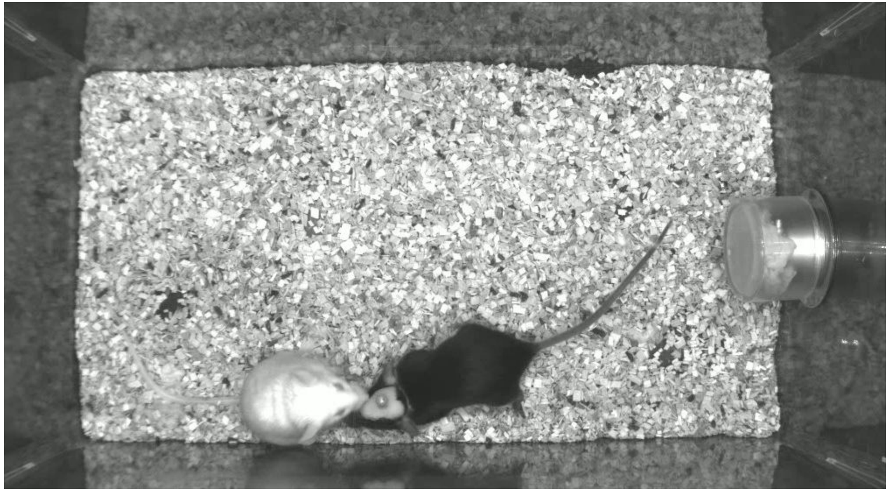
- Manual labeling behavior
    - Time consuming
    - Inconsistent
- Machine labeling behavior
    - Inaccurate
    - Not generalizable

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

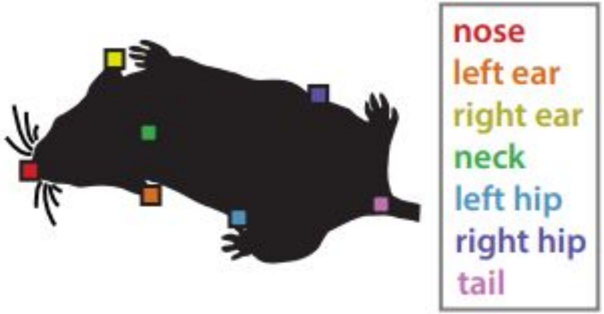
Dataset - Input

Caltech Mouse Social Interactions (CalMS21) Dataset

Videos



Pose-Detection



Dataset - Labels

Per frame labeled video by expert annotator

Supervised

- 70 videos
- 3 behaviors (attack, investigation, mount, other)

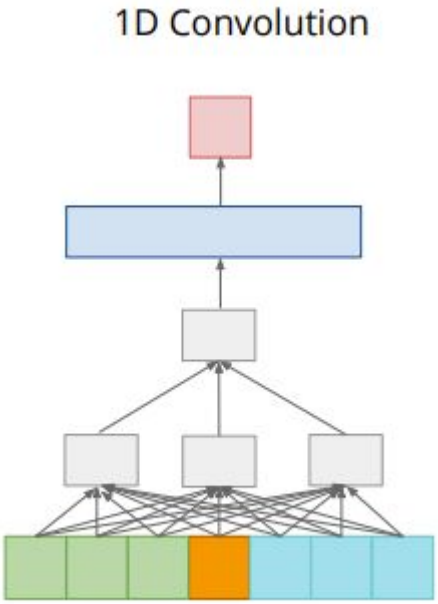
Few-Shot Task

- 7 behaviors, ~3 videos each

Temporal Action Localization - Stage 1

- Trained with supervised task
- 1D convolutional architectures, centered at current frame
- Predicts 0 for action and 1 for no action

- Action proposals considered to start where 1/2 of the next 300 frames are actions

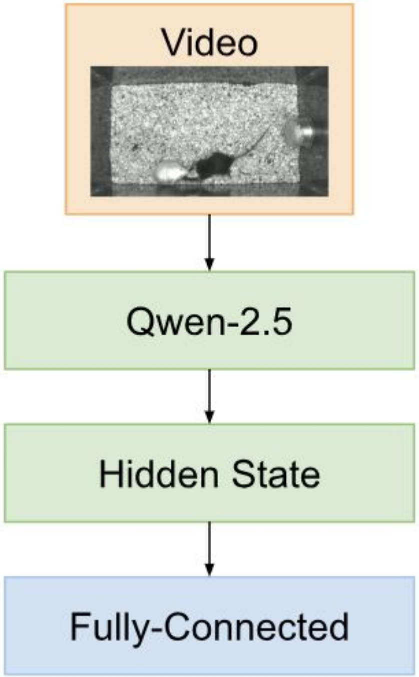


Temporal Action Localization - Stage 2

- Action proposals and the 5 seconds after fed as video into Qwen-2.5
- Embeddings Extracted

To predict per video actions

- Neural network trained on supervised task
- KNN performed for few shot task



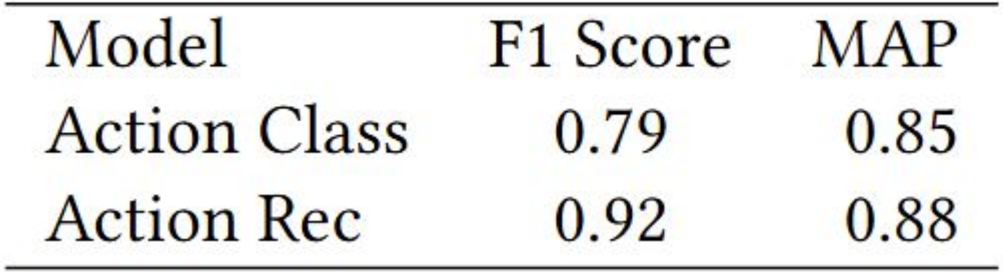
Video

Qwen-2.5

Hidden State

Fully-Connected

Results - Action Localization



Model	F1 Score	MAP
Action Class	0.79	0.85
Action Rec	0.92	0.88

824 action clips found for stage 2

Results - Embeddings

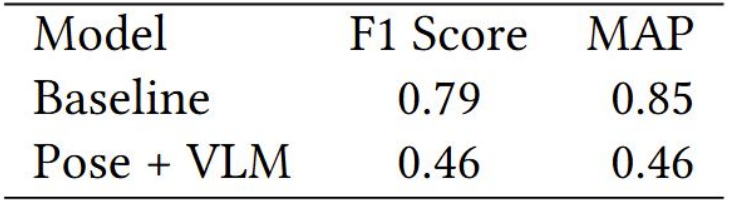
Accuracy: 0.47

MAP: 0.34

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.

Results - Pipeline

Supervised Task



Model	F1 Score	MAP
Baseline	0.79	0.85
Pose + VLM	0.46	0.46

Few Shot Learning

- 0.01 MAP score
- For sniff-face, identified 1/27 clips

Conclusion

- Action localization for known actions is easy but not generalizable
- VLMs have made advancements in amount of data they can process
- VLMs are too language focused
- Having a rich embeddings space is important