Preliminary Use of Vision Language Model Driven Extraction of Mouse Behavior Towards Understanding Fear Expression

Paimon Goulart*
University of California, Riverside
pgoul002@ucr.edu

Jordan Steinhauser*
University of California, Riverside
jstei007@ucr.edu

Kylene Shuler
University of California, Riverside
kshul004@ucr.edu

Edward Korzus
University of California, Riverside
edkorzus@ucr.edu

Jia Chen
University of California, Riverside
jiac@ucr.edu

Evangelos E. Papalexakis
University of California, Riverside
epapalex@cs.ucr.edu

Abstract

Integration of diverse data will be a pivotal step towards improving scientific explorations in many disciplines. This work establishes a vision-language model (VLM) that encodes videos with text input in order to classify various behaviors of a mouse existing in and engaging with their environment. Importantly, this model produces a behavioral vector over time for each subject and for each session the subject undergoes. The output is a valuable dataset that few programs are able to produce with as high accuracy and with minimal user input. Specifically, we use the open-source Qwen2.5-VL model and enhance its performance through prompts, in-context learning (ICL) with labeled examples, and frame-level preprocessing. We found that each of these methods contributes to improved classification, and that combining them results in strong F1 scores across all behaviors, including rare classes like freezing and fleeing, without any model fine-tuning. Overall, this model will support interdisciplinary researchers studying mouse behavior by enabling them to integrate diverse behavioral features, measured across multiple time points and environments, into a comprehensive dataset that can address complex research questions.

1 Introduction

Understanding how mice express fear and how neuronal networks encode threat and safety cues requires precise behavioral annotations synchronized with neural activity. Existing tools often lack the flexibility or scalability needed to capture the nuanced repertoire of behaviors that emerge during fear and safety learning. In this work, we propose the use of vision-language models (VLMs) as a scalable and generalizable solution to automatically annotate behavior from video, offering a powerful advancement for fields like behavioral neuroscience.

In our experimental setup, mice were conditioned to associate a specific environment with threat and then tested in either the threatening or a safe context. Behavior during these sessions was recorded,

*Equal contribution.

Preprint.

and the goal was to annotate the mouse's actions at every second of video, producing a rich behavior vector that can be integrated with other neural datasets (e.g., calcium imaging, Neuropixels). Moving beyond the traditional binary measure of fear as freezing, our framework allows for the classification of a wide range of behaviors, from exploratory sniffing to active fleeing, enabling more nuanced interpretation of safety learning and threat perception.

While existing tools such as DeepLabCut and MoSeq have laid the foundation for video based behavior analysis, they come with notable limitations [1, 2]. DeepLabCut requires extensive user input and may fail to capture subtle exploratory behaviors. MoSeq minimizes user input but is limited to clustering high-frequency stereotyped motifs. In contrast, our VLM approach supports direct behavior labeling via natural language prompts and offers both flexibility and interpretability, with minimal setup or retraining.

Our model takes as input a video of mouse behavior and, given a user defined prompt, returns per-second annotations of behavior and environment. These annotations can then be used independently or integrated with other experiment modalities to ask more complex scientific questions about cognition, emotion, and learning. Our code is available at `https://github.com/Pie115/VLM-Labeling`.

2 Methods

2.1 Data Description

The goal of our model is to extract key indicators of mouse behavior during fear expression and safety learning. Mice were trained to associate a specific environment with threat, and then tested in either the same threatening environment or a similar but safe one. The model annotates these video recordings to identify behaviors that reflect fear or safety, depending on the environment.

Fear-related behaviors are defensive and typically reflect a high perception of threat. These include freezing, fleeing, and periphery bias. Freezing refers to the complete absence of movement other than breathing and is a canonical rodent fear response used to avoid detection [3–11]. Fleeing involves sudden darting or escape behavior in response to a perceived threat [3, 5, 12]. Periphery bias refers to the tendency of mice to remain near the chamber walls; reduced center exploration is a common proxy for anxiety or fear [13].

In contrast, safety-related behaviors suggest a perception of low threat. These include grooming, a stereotyped self-cleaning behavior seen in relaxed states [5, 14], and exploratory behavior, in which the animal actively investigates its surroundings through nose-poking, rearing, and sniffing [6, 12, 15]. Rearing refers to when the mouse lifts its forepaws and stands on its hindlimbs, while sniffing involves head movement and an alert upright posture indicative of curiosity. Center exploration is also considered a safety signal, as mice avoid the center of an open arena under threat.

In total, our dataset consists of 3240 seconds of annotated mouse behavior across multiple video sessions. After merging grooming and exploring into a single category representing safety behavior, the label distribution is as follows: freezing (410/3240, 12.7%), fleeing (21/3240, 0.6%), and grooming/exploring (2809/3240, 86.7%).

2.2 Vision Language Model

In order to consistently and accurately annotate the behavior data of mice, we opt towards a vision language model (VLM). Specifically, we use Qwen2.5-VL as it is an open source model that achieves state-of-the-art performance [16–18]. Although this model is pre-trained on a broad mixture of text and vision data, we discovered it is not well tuned for specific scientific or animal behavior tasks. However, despite this limitation, we find that it is still possible to repurpose through careful prompt engineering and strategic input formatting. Notably, we do not fine-tune the model weights at any stage; all task-specific behavior recognition is achieved through prompting and in-context learning alone.

In our setup, Qwen2.5-VL receives either a full video or individual frames as its visual input, paired with a textual prompt instructing the model to identify the behavior observed.

2.3 Video Preprocessing

To obtain a per-second behavior label vector aligned with each video, we encountered several practical challenges. Initially, we found it extremely difficult for the model to accurately estimate the number of seconds, let alone generate a behavior prediction for each one, when given the full video as input. In general, VLMs are known to struggle with temporal reasoning, especially in tasks such as second-by-second and timelapse interpretation [19]. While recent work has begun to address these limitations, models like Qwen2.5-VL still exhibit a noticeable lack of temporal understanding in long video inputs [18].

To address this, we split each video into individual one-second segments and process them separately. Rather than asking the model to reason over the full duration of the video and identify the behavior occurring at each second, we provide the model with a single second of video at a time and ask it to classify the dominant behavior within that interval. This approach reduces the burden of temporal reasoning and allows the model to focus solely on visual understanding. Importantly, by restricting the model to output only one label at a time, we also reduce the likelihood of hallucinations or cascading errors that often arise in more open ended generations [20]. As a result, our method yields behavior predictions that are not only more accurate but also directly aligned with the desired temporal resolution, one label per second of video.

Another limitation we identified stems from how Qwen2.5-VL processes videos. Since the model samples only a subset of frames from the input video, it is entirely possible for it to miss minor actions or subtle movements that are critical for accurately identifying mouse behavior [18, 21]. To address this, we preemptively split each video into its individual frames and feed each frame into the model independently. By doing so, we effectively force the model to observe the full content of the video, which we hypothesize helps it better distinguish between behaviors, especially those triggered by brief or subtle frame-to-frame transitions.

2.4 In-Context Learning

In-Context Learning (ICL) is a technique that enables language models to perform classification, regression, and other tasks without the need for fine-tuning or task-specific training [22, 23]. While ICL has traditionally been explored in text only settings, recent work has extended its use to visual and multimodal tasks as well [24–26]. Despite this progress, the application of ICL within VLMs, especially for structured video understanding in scientific domains, remains relatively underexplored.

Given these gaps, we hypothesize that ICL can be particularly effective in helping VLMs reason about complex and domain specific behaviors, such as those observed in laboratory mice, by providing a few labeled visual examples as context. In our setup, we explore the use of ICL to improve behavior labeling by supplying the model with a small number of annotated image label pairs (few-shot demonstrations) directly in the prompt. The goal is to improve recognition of subtle or ambiguous behaviors without requiring additional training or domain adaptation.

2.5 Labeling Pipeline

By combining all these methods we come up with the following labeling pipeline shown in 1. First, each input video is split into one-second segments. Each segment is then decomposed into its individual frames, which are fed into the VLM as a single visual unit. This input ensures that the model observes all visual content within the second, rather than relying on its internal frame sampling, which could miss important frames.

For each input video, we prepend a small number of labeled examples drawn from a held out set of videos. These ICL examples are formatted in the same way; being one-second segments decomposed into individual frames and paired with their known behavior labels. All examples and the target video segment are then fed into the model using one of two prompts; a simple prompt or a complex prompt. The simple prompt asks the model to choose one of the behavior labels, while the complex prompt includes behavior definitions to provide additional guidance. An example of the simple and complex prompt is shown in Figures 2 and 3. The simple prompt follows the same format, but instead omits the behavior definitions and is asked to choose one of the labels directly.

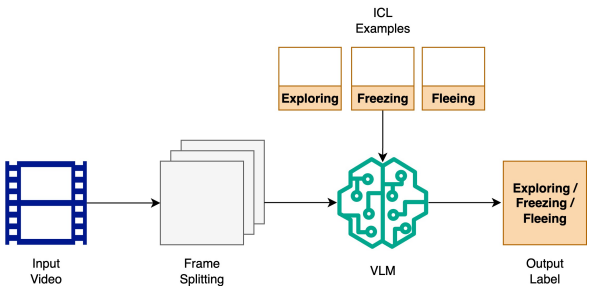


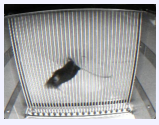
Figure 1: Overview of our video labeling pipeline.

Finally, the model returns a label for the given video segment. This process is repeated for each second of video, resulting in a complete per-second behavior vector aligned with the original video length.

Simple Prompt with In-Context Learning

Task:
Label the mouse's behavior seen per second for this n-second video segment such that we have a label for each second of the video. For example: `[behavior_1, behavior_2, ..., behavior_n]`. Annotation key: `Freezing`, `Fleeing`, `Exploring/Grooming`. Only output the vector! Given the following examples, label the last video to the best of your ability.

Examples:

Video/Frames 1:  → `[Exploring/Grooming]`

Video/Frames 2:  → `[Freezing]`

Video/Frames 3:  → `[Fleeing]`

Target (Video/Frames): **Input** → `[ ]`

Figure 2: Simple ICL prompt: the model receives a few labeled examples and a target video segment, and is asked to choose one behavior label.

3 Experiments

To evaluate the contribution of each component in our labeling pipeline, we conducted a series of experiments assessing their individual and cumulative effects on classification performance. We began with the off-the-shelf Qwen2.5-VL model using both the simple and complex prompts without any ICL or preprocessing. We then incrementally incorporated ICL, followed by the full frame-splitting strategy. This allows us to isolate the impact of each method and quantify their combined effect on performance.

Complex Prompt with In-Context Learning

Task:
Label the mouse's behavior in this n-second clip. Return exactly n label(s) in a Python list `[l1, ..., ln]`.
Allowed: `Freezing`, `Fleeing`, `Exploring/Grooming`. `Freezing` = absolutely no visible movement across the whole second (no head/ear/whisker/tail or body motion). `Exploring/Grooming` = any visible movement that isn't fast fleeing; includes slow stepping in place, head/whisker/ear/tail motion, sniffing, re-orienting, or brief rearing with little displacement. `Fleeing` = fast, sustained locomotion, large displacement, motion blur, or dashing.
If unsure between fleeing and exploring, choose fleeing if movement is more rapid. Rules: if any movement is seen in any frames, do NOT output `Freezing`. Output only the list.

Examples:

Video/Frames 1:  → `[Exploring/Grooming]`

Video/Frames 2:  → `[Freezing]`

Video/Frames 3:  → `[Fleeing]`

Target (Video/Frames): **Input** → `[ ]`

Figure 3: Complex ICL prompt: the model is given definitions for each behavior alongside example demonstrations. Note that non-ICL uses the same prompt without examples.

We evaluated each configuration on all 3240 seconds of data using per-class F1 scores to account for the heavy class imbalance.

We see the results of these configurations in Figure 4. Across all configurations, we observe a clear progression in model performance as each component of the pipeline is added. Starting with the off-the-shelf model, we find that the simple prompt achieves moderate F1 scores for the dominant class (grooming/exploring), but performs poorly on freezing and fails entirely to capture fleeing behavior. The complex prompt improves Grooming/Exploring and slightly boosts freezing recognition, but still misses fleeing.

The introduction of ICL provides notable gains, especially for the rare classes. With ICL, the model begins to label fleeing behavior and achieves consistent improvements in freezing detection across both prompt styles. This suggests that visual demonstrations help guide the model toward rare or ambiguous behaviors that are hard to detect from the raw video alone.

Finally, incorporating full frame-splitting further amplifies these improvements. The combination of ICL and frame-wise processing (top right plot) yields the best overall results, with strong F1 scores across all behaviors. Notably, ICL + Complex Prompt with frame-wise input recovers both freezing and fleeing with measurable accuracy.

These findings confirm that prompt design, ICL, and preprocessing each contribute meaningfully, with their combination yielding the strongest results.

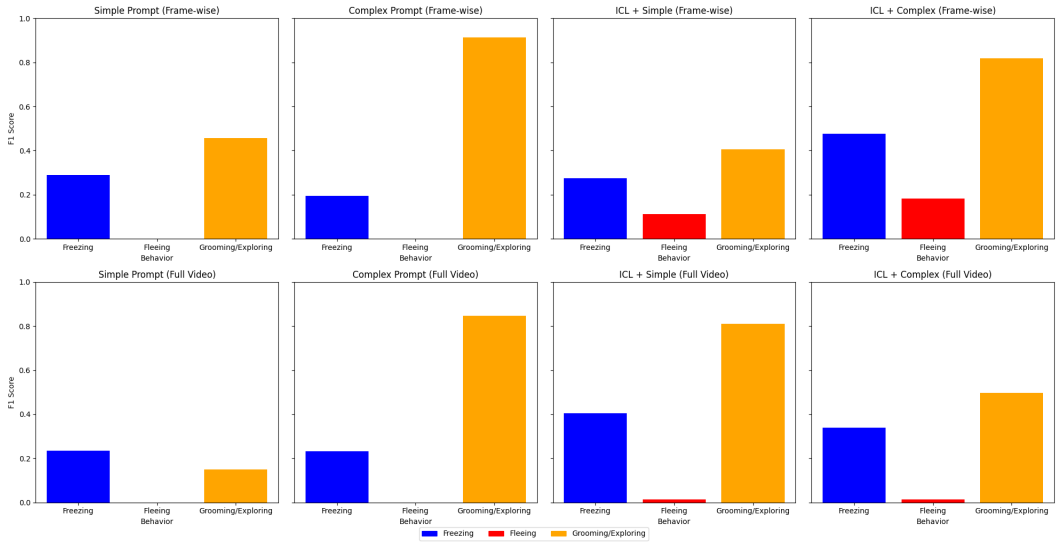


Figure 4: Per class F1 scores across all pipeline configurations. We compare the effects of using simple vs. complex prompts, with and without ICL, as well as the impact of full video input vs. frame-wise input. ICL and frame splitting significantly improve performance, especially for rare behaviors.

4 Conclusions

We present a VLM pipeline for annotating mouse behavior with minimal supervision, enabling second-by-second labeling of fear and safety related actions. Through experimentation, we show that prompt design, ICL, and frame-level preprocessing each contribute significantly to performance, with their combination yielding the most accurate results, especially for rare but scientifically critical behaviors like freezing and fleeing.

This work demonstrates the feasibility of using VLMs to generate behavioral annotation vectors that can be integrated with neural recordings or other experimental data, ultimately leading to insights into threat and safety learning, among other animal behaviors and cognitions.

Future work will explore incorporating temporal context across video segments, labeling of longer segments, richer environmental annotations, and human-in-the-loop correction mechanisms to further enhance label quality.

References

[1] Tanmay Nath, Alexander Mathis, An Chi Chen, Anish Patel, Matthias Bethge, and Mackenzie W. Mathis. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, 14:2152–2176, 2019. doi: 10.1038/s41596-019-0176-0. URL `https://doi.org/10.1038/s41596-019-0176-0`.

[2] Sheng Lin, William F. Gillis, Caleb Weinreb, Alexandra Zeine, Sean C. Jones, Elizabeth M. Robinson, Jeffrey Markowitz, and Sandeep R. Datta. Characterizing the structure of mouse behavior using motion sequencing. *Nature Protocols*, 19(11):3242–3291, Nov 2024. doi: 10.1038/s41596-024-01015-w. URL `https://doi.org/10.1038/s41596-024-01015-w`.

[3] D. Blanchard. Mouse defensive behaviors: Pharmacological and behavioral assays for anxiety and panic. *Neuroscience & Biobehavioral Reviews*, 25(3):205–218, May 2001. doi: 10.1016/S0149-7634(01)00009-4.

[4] Makoto Takemoto and Wen-Jie Song. Cue-dependent safety and fear learning in a discriminative auditory fear conditioning paradigm in the mouse. *Learning & Memory*, 26(8):284–290, Jul 2019. doi: 10.1101/lm.049577.119.

[5] Robert Gerlai. Contextual learning and cue association in fear conditioning in mice: A strain comparison and a lesion study. *Behavioural Brain Research*, 95(2):191–203, Oct 1998. doi: 10.1016/S0166-4328(97)00144-7.

[6] T. Rao Laxmi, Oliver Stork, and Hans-Christian Pape. Generalisation of conditioned fear and its behavioural expression in mice. *Behavioural Brain Research*, 145(1–2):89–98, Oct 2003. doi: 10.1016/S0166-4328(03)00101-3.

[7] Jacob W. Clark, Sean P. A. Drummond, Daniel Hoyer, and Laura H. Jacobson. Sex differences in mouse models of fear inhibition: Fear extinction, safety learning, and fear–safety discrimination. *British Journal of Pharmacology*, 176(21):4149–4158, Apr 2019. doi: 10.1111/bph.14600.

[8] Karin Roelofs and Peter Dayan. Freezing revisited: Coordinated autonomic and central optimization of threat coping. *Nature Reviews Neuroscience*, 23(9):568–580, Jun 2022. doi: 10.1038/s41583-022-00608-2.

[9] Lindsay C. Laughlin, Danielle M. Moloney, Shanna B. Samels, Robert M. Sears, and Christopher K. Cain. Reducing shock imminence eliminates poor avoidance in rats. *Learning & Memory*, 27(7):270–274, Jun 2020. doi: 10.1101/lm.051557.120.

[10] Jeremy M. Trott, Ann N. Hoffman, Irina Zhuravka, and Michael S. Fanselow. Conditional and unconditional components of aversively motivated freezing, flight and darting in mice. *eLife*, 11, May 2022. doi: 10.7554/eLife.75663.

[11] Nancy J. Smith, Sara Y. Markowitz, Ann N. Hoffman, and Michael S. Fanselow. Adaptation of threat responses within the negative valence framework. *Frontiers in Systems Neuroscience*, 16, May 2022. doi: 10.3389/fnsys.2022.886771.

[12] G. Griebel, C. Belzung, R. Misslin, and E. Vogel. The free-exploratory paradigm. *Behavioural Pharmacology*, 4(6), Dec 1993. doi: 10.1097/00008877-199312000-00009.

[13] Russell D. Romeo, Astrid Mueller, Helene M. Sisti, Sonoko Ogawa, Bruce S. McEwen, and Wayne G. Brake. Anxiety and fear behaviors in adult male and female C57BL/6 mice are modulated by maternal separation. *Hormones and Behavior*, 43(5):561–567, May 2003. doi: 10.1016/S0018-506X(03)00063-1.

[14] Aijaz Ahmad Naik, Xiaoyu Ma, Maxime Munyeshyaka, Ellen Leibenluft, and Zheng Li. A new behavioral paradigm for frustrative nonreward in juvenile mice. *Biological Psychiatry Global Open Science*, 4(1):31–38, Jan 2024. doi: 10.1016/j.bpsgos.2023.09.007.

[15] Jamie N. Justice, Christy S. Carter, Hannah J. Beck, Rachel A. Gioscia-Ryan, Matthew McQueen, Roger M. Enoka, and Douglas R. Seals. Battery of behavioral tests in mice that models age-associated changes in human motor function. *AGE*, 36(2):583–595, Oct 2013. doi: 10.1007/s11357-013-9589-9.

[16] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision–language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[17] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision–language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[18] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[19] Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. Can multimodal LLMs do visual temporal understanding and reasoning? the answer is no!, 2025. URL `https://arxiv.org/abs/2501.10674`.

[20] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL `https://arxiv.org/abs/2509.04664`.

[21] Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Improving LLM video understanding with 16 frames per second. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=3H7qAT9Qow`.

[22] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?, 2022. URL `https://arxiv.org/abs/2202.12837`.

[23] Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matt Botvinick, Jane Wang, and Eric Schulz. Meta-in-context learning in large language models. *Advances in Neural Information Processing Systems*, 36: 65189–65201, 2023.

[24] Dyke Ferber, Georg Wölflein, Isabella C. Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar S. M. El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, and Jakob Nikolas Kather. In-context learning enables multimodal large language models to classify cancer pathology images, 2024. URL https://arxiv.org/abs/2403.07407.

[25] Kangsan Kim, Geon Park, Youngwan Lee, Woongyeong Yeo, and Sung Ju Hwang. Videoicl: Confidence-based iterative in-context learning for out-of-distribution video understanding, 2024. URL https://arxiv.org/abs/2412.02186.

[26] Wentao Zhang, Junliang Guo, Tianyu He, Li Zhao, Linli Xu, and Jiang Bian. Video in-context learning: Autoregressive transformers are zero-shot video imitators, 2025. URL https://arxiv.org/abs/2407.07356.