Rodent-Bench

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present Rodent-Bench, a novel benchmark designed to evaluate the ability of Multimodal Large Language Models (MLLMs) to annotate rodent behaviour footage. We evaluate state-of-the-art MLLMs, including Gemini-2.5-Pro, Gemini-2.5-Flash and Qwen-VL-Max, using this benchmark and find that none of these models perform strongly enough to be used as an assistant for this task. Our benchmark encompasses diverse datasets spanning multiple behavioral paradigms including social interactions, grooming, scratching, and freezing behaviors, with videos ranging from 10 minutes to 35 minutes in length. We provide two benchmark versions to accommodate varying model capabilities and establish standardized evaluation metrics including second-wise accuracy, macro F1, mean average precision, mutual information, and Matthew's correlation coefficient. While some models show modest performance on certain datasets (notably grooming detection), overall results reveal significant challenges in temporal segmentation, handling extended video sequences, and distinguishing subtle behavioral states. Our analysis identifies key limitations in current MLLMs for scientific video annotation and provides insights for future model development. Rodent-Bench serves as a foundation for tracking progress toward reliable automated behavioral annotation in neuroscience research.

1 Introduction

Behavioral analysis is fundamental to neuroscience and biomedical research, yet manual annotation of animal behavior videos remains a time-consuming bottleneck that limits research scale and reproducibility (Sturm et al., 2020; Mathis & Mathis, 2020). While Multimodal Large Language Models (MLLMs) have shown impressive capabilities in vision-language tasks (Fu et al., 2024; Yin et al., 2024), their application to specialized scientific domains like behavioral analysis remains largely unexplored. MLLMs offer particular promise for scientific annotation tasks as they can potentially handle diverse behavioral paradigms through natural language instructions without requiring specialized model training for each new behavior or experimental setup.

Unlike general computer vision tasks, behavioral analysis requires models to identify subtle, context-dependent actions, maintain temporal coherence across extended sequences, and produce structured outputs aligned with ethological frameworks. Traditional computer vision approaches require training specialized models for each behavioral task, but MLLMs could streamline this process by accepting task descriptions in natural language and adapting to new behaviors without retraining. Existing video understanding benchmarks inadequately address these scientific requirements, creating a significant gap between current MLLM capabilities and practical research applications.

We present **Rodent-Bench-Short** and **Rodent-Bench-Long**, the first comprehensive benchmarks for evaluating MLLMs on rodent behavioral annotation tasks. Our benchmarks encompasse diverse

Submitted to the AI for Science workshop (NeurIPS 2025)

datasets spanning multiple behavioral paradigms and provides standardized evaluation metrics to assess current model capabilities. We evaluate state-of-the-art MLLMs including Gemini-2.5-Pro, Gemini-2.5-Flash, and Qwen-VL-Max, revealing significant performance gaps that limit their utility as research assistants. While some of these models show fair performance on some datasets, our analysis identifies specific challenges in temporal segmentation, long video processing, and handling varied experimental conditions, providing insights for future improvements in scientific applications of multimodal models.

2 Related Work

The emergence of Multimodal Large Language Models (MLLMs) has opened new possibilities for video understanding tasks across diverse domains. Recent comprehensive benchmarks such as Video-MME (Fu et al., 2024) have evaluated state-of-the-art MLLMs including GPT-4 and Gemini on video analysis tasks, revealing significant challenges in temporal reasoning and long-form video understanding. Surveys on video understanding with large language models (Tang et al., 2025; Wang et al., 2024) highlight the emergent capabilities of these systems for multi-granularity reasoning, while identifying key limitations in handling long-form videos and maintaining alignment between visual and textual modalities. Despite these advances, the application of MLLMs to specialized scientific domains remains under-explored, with recent work suggesting significant potential for leveraging these models in natural science research (Yin et al., 2024; Testard et al., 2024).

Traditional animal behavior analysis has undergone significant transformation with the advent of deep learning and computer vision techniques. Deep learning-based behavioral analysis systems have demonstrated the ability to reach human-level accuracy in recognizing specific ethological behaviors (Sturm et al., 2020), with markerless pose estimation tools like DeepLabCut enabling robust tracking of individual body parts in freely moving rodents (Mathis & Mathis, 2020). Specialized tools such as DeepBehavior (Arac et al., 2019), ezTrack (Pennington et al., 2019), MoSeq (Wiltschko et al., 2015), SLEAP Pereira et al. (2022) and real-time behavior recognition systems (de Chaumont et al., 2022) have been developed specifically for automated analysis of animal behavior. However, these systems typically require task-specific training and lack the flexibility and generalization capabilities that modern MLLMs potentially offer. The specific application of MLLMs to behavioral annotation tasks in laboratory settings remains largely unexplored, representing a significant gap that our Rodent-Bench benchmark aims to address.

3 Rodent-Bench

We produced two benchmarks: **Rodent-Bench-Short**, with videos up to 10 minutes long; and **Rodent-Bench-Long**, with videos up to 35 minutes long. We created these two versions because current MLLMs have varying video length limitations—while models like Gemini can process videos up to 1 hour, others like Qwen-VL-Max are restricted to 10 minutes or less. This dual-benchmark approach ensures compatibility across all evaluated models and enables investigation of how video length affects annotation performance.

Along with this we suggest evaluation metrics. The task posed to the MLLM is to annotate each video, determining which of a fixed set of behaviours is occurring and to produce a JSON file segmenting each video into discrete non-overlapping time segments with behaviour labels.

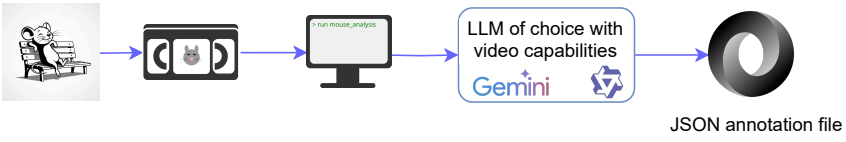


Figure 1: Workflow for annotating rodent videos.

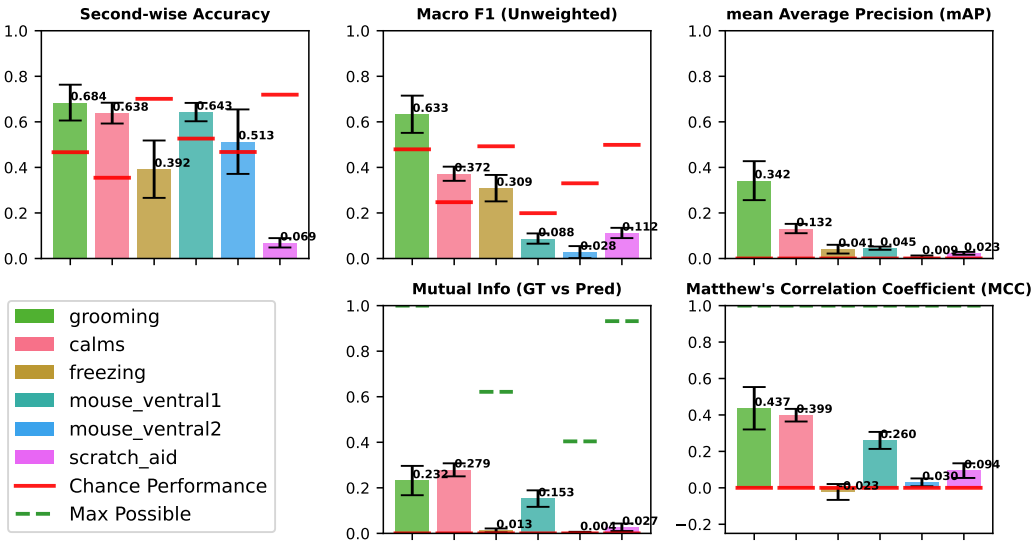


Figure 2: Performance metrics for Gemini-2.5-Pro across all datasets. Each metric shows substantial variation across behavioral paradigms, with the grooming detection dataset achieving the highest performance across most metrics. Social behaviors (CalMS21) show moderate performance, while challenging datasets like freezing and scratch detection exhibit poor performance approaching chance levels. Dashed lines indicate theoretical maximum performance where applicable. Error bars represent $2\times$ standard error across videos within each dataset. The consistently low performance on certain datasets highlights the difficulty of fine-grained temporal behavioral annotation for current MLLMs.

3.1 Data Collection

We collected our data from several openly available datasets, as well as one private dataset which we now make freely available.

Caltech Rodent Social Interactions (CalMS21):

The CalMS21 dataset (Sun et al., 2021) is intended for multi-agent behaviour modelling. It consists of footage of multiple rodents interacting socially, with 6 million un-labelled frames and 1 million labelled frames. The labelled frames consist of both frame level behaviour and pose tracking annotations. For our purposes we use the first 25 labelled videos in the training set.

Rodent Grooming Detection Annotated Dataset:

The rodent grooming dataset (Geuther et al., 2021) was collected in order to train a neural network rodent grooming classifier. It consists of 1,253 video clips with 2,637,363 frames. Each frame is labelled "Grooming" or "Not Grooming". We use the first 25 videos in the training set for our eval.

Mouse-Ventral 1&2: We use the Mouse-Ventral 1&2 subsets of the Deep Ethogram dataset (Bohnslav et al., 2021). These consists of 30 minute videos of a rodents shot from below, 16 for MV2 and 28 for MV1, the videos are annotated with behaviour labels. In the Mouse-Ventral1 subset the rodents are either "grooming", "digging", "scratching", "licking" or "background" (neither scratching nor licking). In the Mouse-Ventral2 subset the rodents are either "scratching", "licking" or "background".

Scratch-AID: The Scratch-AID dataset (Yu et al., 2022) was collected to train a neural network CRNN rodent scratching classification model. The dataset consists of 40 videos of rodents shot from below, the rodents were injected with an itching agent causing them to scratch compulsively. The model trained especially for this task achieved 97.6% recall and 96.9% precision on previously unseen test videos.

Freezing: Our collaborators have given us access to nine videos of rodents displaying a "freezing" behaviour. This is a behavior distinct from resting, and is characterised by the ears being oriented towards the front indicating alert but immobile behavior (Blanchard & Blanchard, 1969). There are three types of videos, and three videos of each type: Low freezing, high freezing and extinction.

Extinction is a behavioral paradigm where the conditioned freezing response is gradually reduced through repeated exposure to the conditioned stimulus without the unconditioned stimulus.

This dataset is particularly important for evaluating MLLMs because freezing behavior presents challenges that traditional pose estimation approaches cannot address. While tools like DeepLabCut excel at tracking body parts and movements, they cannot distinguish between freezing (an active fear response) and other motionless states such as sleeping, resting, or general inactivity. These behaviors are not easily distinguishable when relying solely on pose or movement data. MLLMs, with their ability to integrate visual context, temporal patterns, and behavioral understanding, may offer advantages for this subtle but scientifically important distinction.

Rodent-Bench-Short: Some MLLMs will not accept long video files (30 minutes to an hour), so to evaluate these models we produce a shortened version of the dataset in which any file longer than 10 minutes is shortened to that length. We evaluate all models on both datasets for comparison.

3.2 Metrics

Second-wise accuracy: We treat each second as a binary classification problem: is the behaviour in that second correctly classified or not. We then report the proportion of seconds in which the behaviour was correctly classified.

Macro F1: We calculate the F1 score for each class and average with no weighting.

For each behavior class c:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{1}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{2}$$

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{3}$$

Macro F1 is the unweighted average across all classes:

$$\text{Macro F1} = \frac{1}{|C|} \sum_{c \in C} \text{F1}_c \tag{4}$$

where $|C|$ is the number of behavior classes, TP_c is true positives for class c, FP_c is false positives, and FN_c is false negatives.

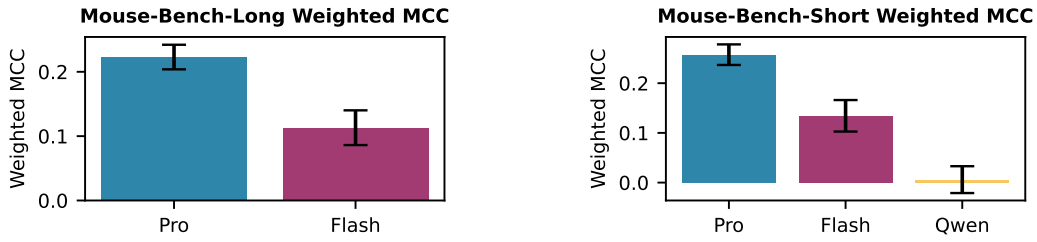
mean Average Precision (mAP): This is calculated by comparing predicted and ground truth behaviour segments across a range of IoU (Intersection over Union) thresholds (from 0.1 to 0.9) (Henderson & Ferrari, 2016). For each threshold, we match predicted segments to ground truth segments of the same behaviour if their IoU exceeds the threshold, counting true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Precision and recall are computed at each threshold, and the average precision is accumulated as the sum of precision values weighted by the change in recall, this is to approximate the area under the precision-recall curve. The final mAP is the total of these values, providing a single metric that summarizes how well the predictions align with the ground truth across different levels of overlap.

Mutual Information: We calculate the mutual information between the ground-truth second-wise labels and the predicted labels.

Matthew's Correlation Coefficient (MCC): This is a correlation coefficient between -1 and 1. It is calculated:

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Dataset label entropy weighted Matthew's Correlation Coefficient: To provide a singular score which takes into account differing datasets "difficulty":



(a) (b)

Figure 3: Weighted Matthew's Correlation Coefficient (MCC) performance across models. (a) Rodent-Bench-Long: Gemini-2.5-Pro achieves the highest performance with lower variance compared to Gemini-2.5-Flash. (b) Rodent-Bench-Short: Similar performance hierarchy with Gemini-2.5-Pro outperforming Flash, while Qwen-VL-Max shows near-chance performance. Error bars represent $2\times$ standard error across datasets. All models show modest performance levels, indicating substantial room for improvement in behavioral annotation tasks.

$$w_i = \frac{(H_i + \epsilon) \cdot T_i}{\sum_j (H_j + \epsilon) \cdot T_j}$$

Where, H_i is the entropy of dataset i, T_i is the duration of dataset i in seconds, $\epsilon = 10^{-8}$ is a small constant to avoid zero weights. This weighting scheme assigns higher weight to longer datasets with more diverse labels, which should be more challenging.

We prioritize mutual information, MCC, and mAP metrics for our primary analysis. Mutual information and MCC provide interpretable baselines, both equalling zero when predictions are statistically independent of ground truth labels, making chance-level performance easily identifiable across all datasets. The mAP metric is valuable for evaluating temporal segmentation quality, as it directly measures how well predicted behavioral segments align with ground truth boundaries across multiple IoU thresholds. In contrast, metrics like second-wise accuracy and macro F1 have dataset-dependent chance baselines that vary with class distributions and choice of random baseline strategy (uniform random vs. frequency-matched random prediction), complicating cross-dataset comparisons and performance interpretation.

4 Experiments

To provide an idea of how models currently perform we evaluate some of the available MLLM's on this benchmark.

4.1 Experimental Setup

Models We evaluate our benchmark on three MLLMs. As of July 2025, a number of MLLMs which claim support for video actually just sample frames from the video at regular intervals and use these (e.g. Qwen-VL-Max). We use Gemini-2.5-Flash, Gemini-2.5-Pro and Qwen-VL-Max. Specifications of these models can be found in Appendix B.

Prompting Strategy Because our dataset is heterogenous we use different prompts for each of the sub-datasets, these appear in appendix G. Despite this we tried to use a similar prompt for each dataset in the interest of fairness. The same prompt is used for all models.

Computational Cost: For the complete benchmark, costs are approximately $7 for Gemini-2.5-Pro, $7 for Qwen-VL-Max, and under $1 for Gemini-2.5-Flash.

4.2 Results

We evaluate three MLLMs on our benchmark. Further figures showing individual dataset performance for each model can be found in appendices D, E and F.

In figure 3a you can see that Gemini Pro outperforms Flash in both absolute performance, as well as variability.

In 3b, evaluating on the shortened dataset, we see a similar comparison between Pro and Flash, while Qwen-VL-Max performs no better than chance.

We notice that both Flash and Qwen models struggle with correct formatting, sometimes the wrong key for certain segments ("end_long_time" instead of "end_time") or in the case of Qwen-VL it will simply stop partway through a segment, rendering the JSON file unreadable (without modifications).

We speculate that these models perform strongest on datasets with shorter videos on average, that have clearly defined labels which depend only on behaviour (i.e do not require the rodent to be in a particular position in the cage when acting for that label to apply), and have behaviours that last at least a few seconds. They perform weakly on videos which have visual filters applied but that require some degree of colour recognition for labelling (i.e "the feeding box is at the back of the cage and is black"), or that are taken from a "non-standard" (i.e not front facing, from above or from below) camera angle, or that have very short or ambiguous behaviours. For instance the freezing dataset features behaviours shorter than a second, and the difference between the rodent "freezing" and simply not moving is quite subtle.

5 Limitations

Our benchmark has several important limitations. First, ground-truth annotations were taken from existing datasets with varying labelling schemes and quality standards, potentially containing inconsistencies that affect evaluation reliability. Second, we lack human annotator baselines to contextualize model performance—while current MLLMs perform poorly, we cannot determine how their accuracy compares to average human annotators on these specific videos.

Third, our evaluation uses zero-shot inference without fine-tuning or extensive prompt optimization. While this ensures fair comparison across models, it may underestimate achievable performance through model adaptation or specialized prompting strategies. Additionally, dataset-specific prompts introduce variability that could advantage certain models.

Finally, our evaluation is limited to three commercially available models with native video processing, and the rapid pace of model development means newer capabilities may alter these findings. Despite these limitations, Rodent-Bench provides a valuable initial assessment of current MLLM capabilities for scientific behavioral annotation tasks.

6 Conclusion

We introduced Rodent-Bench, the first comprehensive benchmark for evaluating Multimodal Large Language Models on scientific behavioral annotation tasks. Our evaluation of state-of-the-art MLLMs—Gemini-2.5-Pro, Gemini-2.5-Flash, and Qwen-VL-Max—reveals that current models perform substantially below the accuracy levels required for practical deployment as research assistants in behavioral neuroscience.

While MLLMs showed modest success on certain datasets (notably grooming detection), performance varied dramatically across behavioral paradigms. Models struggled particularly with subtle temporal distinctions, brief behavioral episodes, and tasks requiring integration of spatial context with behavioral understanding. The freezing behavior dataset exemplified these challenges, where distinguishing between active freezing responses and passive inactivity proved difficult even for advanced multimodal systems.

Our findings highlight several critical areas for improvement. First, enhanced temporal reasoning capabilities are needed to handle the fine-grained segmentation required for behavioral analysis. Second, models must develop better contextual understanding to distinguish between visually similar but behaviorally distinct states. Finally, output formatting consistency remains a practical barrier, with some models frequently producing malformed JSON responses that complicate automated processing.

Despite current limitations, Rodent-Bench establishes a foundation for tracking progress in scientific applications of multimodal AI. The benchmark's diverse behavioral paradigms and standardized

evaluation framework provide a testbed for future model improvements. As MLLMs advance, their potential to democratize behavioral annotation, eliminating the need for specialized model training for each experimental paradigm, remains promising. Rodent-Bench will enable researchers to objectively assess when these models achieve the reliability threshold necessary for practical scientific deployment.

The gap between current capabilities and scientific requirements underscores the need for continued research at the intersection of multimodal AI and domain-specific applications. Our benchmark contributes to this effort by providing concrete evaluation targets and highlighting the unique challenges that scientific video understanding presents to current generation models.

References

Arac, A., Zhao, P., Dobkin, B. H., Carmichael, S. T., and Golshani, P. Deepbehavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data. *Frontiers in Systems Neuroscience*, 13:20, 2019.

Blanchard, R. J. and Blanchard, D. C. Passive and active reactions to fear-eliciting stimuli. *Journal of comparative and physiological psychology*, 68(1p1):129, 1969.

Bohnslav, J. P., Wimalasena, N. K., Clausing, K. J., Dai, Y. Y., Yarmolinsky, D. A., Cruz, T., Kashlan, A. D., Chiappe, M. E., Orefice, L. L., Woolf, C. J., and Harvey, C. D. Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *eLife*, 10:e63377, sep 2021. ISSN 2050-084X. doi: 10.7554/eLife.63377. URL `https://doi.org/10.7554/eLife.63377`.

de Chaumont, F., Coura, R. D. S., Serreau, P., Cressant, A., Chabout, J., Granon, S., and Olivo-Marin, J.-C. Real-time analysis of the behaviour of groups of mice via a depth-sensing camera and machine learning. *bioRxiv*, pp. 2022–02, 2022.

Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Geuther, B. Q., Peer, A., He, H., Sabnis, G., Philip, V. M., and Kumar, V. Action detection using a neural network elucidates the genetics of mouse grooming behavior. *eLife*, 10:e63207, mar 2021. ISSN 2050-084X. doi: 10.7554/eLife.63207. URL `https://doi.org/10.7554/eLife.63207`.

Henderson, P. and Ferrari, V. End-to-end training of object class detectors for mean average precision. In *Asian conference on computer vision*, pp. 198–213. Springer, 2016.

Mathis, M. W. and Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*, 60:1–11, 2020.

Pennington, Z. T., Dong, Z., Feng, Y., Vetere, L. M., Page-Harley, L., Shuman, T., and Cai, D. J. eztrack: An open-source video analysis pipeline for the investigation of animal behavior. *Scientific reports*, 9(1):19979, 2019.

Pereira, T. D., Tabris, N., Matsliah, A., Turner, D. M., Li, J., Ravindranath, S., Papadoyannis, E. S., Normand, E., Deutsch, D. S., Wang, Z. Y., et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022.

Sturm, G., Friede, T., Müller, S., Mathis, A., and Mathis, M. W. Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacology*, 45(11):1942–1952, 2020.

Sun, J. J., Karigo, T., Chakraborty, D., Mohanty, S. P., Wild, B., Sun, Q., Chen, C., Anderson, D. J., Perona, P., Yue, Y., and Kennedy, A. The multi-agent behavior dataset: Mouse dyadic social interactions. *arXiv preprint arXiv:2104.02710*, 2021.

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Liu, P., Feng, M., Zheng, F., Zhang, J., Luo, P., Luo, J., and Xu, C. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. doi: 10.1109/TCSVT.2025.3566695.

Testard, C., Buch, A., Gauthier, J., Marvin, J., Yartsev, M., and Fairhall, A. Data science opportunities of large language models for neuroscience and biomedicine. *Neuron*, 112(4):499–510, 2024.

Wang, J. et al. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*, 2024.

Wiltschko, A. B., Johnson, M. J., Iurilli, G., Peterson, R. E., Katon, J. M., Pashkovski, S. L., Abraira, V. E., Adams, R. P., and Datta, S. R. Mapping sub-second structure in mouse behavior. *Neuron*, 88 (6):1121–1135, 2015.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Yu, H., Xiong, J., Ye, A. Y., Cranfill, S. L., Cannonier, T., Gautam, M., Zhang, M., Bilal, R., Park, J.-E., Xue, Y., Polam, V., Vujovic, Z., Dai, D., Ong, W., Ip, J., Hsieh, A., Mimouni, N., Lozada, A., Sosale, M., Ahn, A., Ma, M., Ding, L., Arsuaga, J., and Luo, W. Scratch-aid, a deep learning-based system for automatic detection of mouse scratching behavior with high accuracy. *eLife*, 11:e84042, dec 2022. ISSN 2050-084X. doi: 10.7554/eLife.84042. URL `https://doi.org/10.7554/eLife.84042`.

A Implementation Details

A.1 Model Access and Configuration

Gemini Models: We access Gemini-2.5-Pro and Gemini-2.5-Flash through Google's Vertex AI GenAI SDK. Videos are processed directly from Google Cloud Storage URIs using the native video input capabilities. The response format is constrained to JSON using structured output schemas specific to each dataset's behavior categories.

Qwen-VL-Max: We access Qwen-VL-Max through Alibaba's DashScope API using the OpenAI-compatible interface.

A.2 Video Processing Pipeline

Each video is processed independently with dataset-specific prompts that include behavior definitions, temporal annotation requirements, and output format specifications.

Structured Output Schemas: We define Pydantic models for each dataset's behavior categories to ensure consistent JSON output formatting. The following is the model for CaLMS21:

```
class RodentBehaviorSegment(BaseModel):
    segment_number: int = Field(..., description="Segment number in order")
    start_time: str = Field(..., description="Start time in MM:SS format")
    end_time: str = Field(..., description="End time in MM:SS format")
    behavior: str = Field(..., description="Behavior label (e.g., attack, investigation, mount, other)")
```

A.3 Batch Processing Implementation

For Gemini models, we implement both individual and batch processing modes. Batch mode generates JSONL files conforming to Gemini's batch API requirements, uploads input files to Google Cloud Storage, and monitors job completion through the batch API. This approach significantly reduces API costs for large-scale evaluations while maintaining identical model configurations.

Error Handling: The system logs all API responses, including malformed outputs, to facilitate debugging. For models producing incomplete JSON (particularly Qwen-VL-Max), we save raw responses to text files for manual inspection. Output validation ensures all required fields are present and temporal segments are non-overlapping.

B Model Specifications

We evaluate three state-of-the-art multimodal large language models with native video understanding capabilities.

B.1 Gemini-2.5-Pro

Key Specifications:

- Maximum video length: 1 hour (without audio), 45 minutes (with audio)
- Context window: 1,048,576 tokens (input), Maximum 65,535 tokens (output)
- Maximum video file size: 2 GB

B.2 Gemini-2.5-Flash

Key Specifications:

- Maximum video length: 1 hour (without audio), 45 minutes (with audio)
- Context window: 1,048,576 tokens (input), Maximum 65,535 tokens (output)
- Maximum video file size: 2 GB

B.3 Qwen-VL-Max

Qwen-VL-Max is Alibaba Cloud's most advanced vision-language model. Unlike the Gemini models which process video natively, Qwen-VL extracts frames from video files for analysis, extracting one frame every 0.5 seconds when using the OpenAI SDK.

Key Specifications:

- Maximum video length: 10 minutes (Qwen2.5-VL series)
- Context window: 129,024 input tokens, 8,192 output tokens
- Maximum video file size: 1 GB (via URL), 10 MB (Base64 encoded)
- Video processing: Frame extraction (no audio support)

B.4 Model Selection Rationale

We selected these models based on three criteria: (1) native video processing capabilities, (2) availability through stable APIs for reproducible evaluation, and (3) demonstrated performance on complex reasoning tasks.

C Datasets

We include screenshots from each dataset, demonstrating each behavior. We additionally include a chart showing the proportion of each behavior in each dataset.

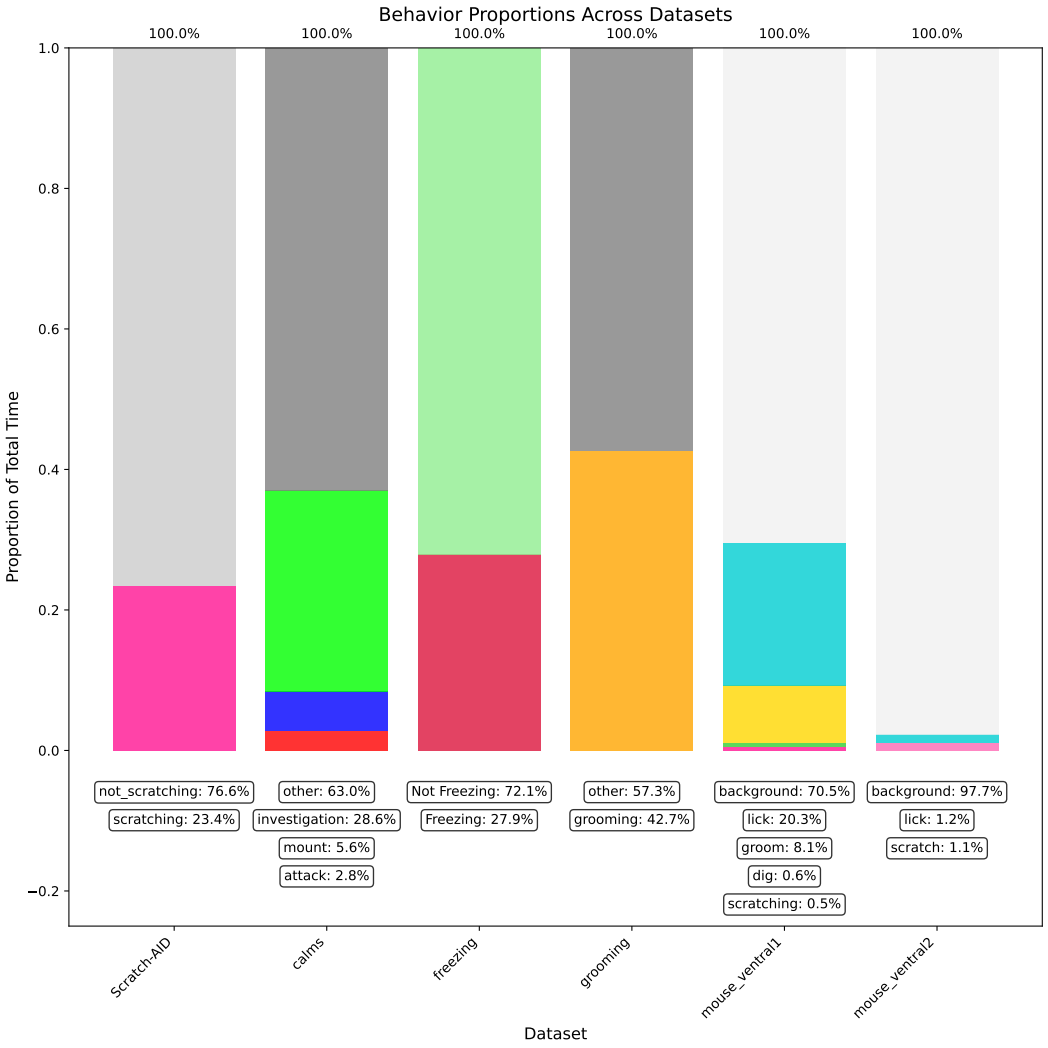


Figure 4: Behavior Proportions for each dataset.

C.1 Rodent-Bench

Table 1: Timing Statistics by Dataset

Dataset	Average Time (Mins)	Minimum Time (Mins)	Maximum Time (Mins)	Total Time (Mins)
CalMS21	4.41	1.02	11.87	110.22
Freezing	14.35	4.01	32.67	129.17
Grooming	1.32	0.34	5.99	32.88
Rodent Ventral 1	8.33	8.28	8.33	233.25
Rodent Ventral 2	29.97	29.97	29.97	479.57
Scratch-AID	20.00	20.00	20.00	300.03

C.2 Rodent-Bench Short

Table 2: Timing Statistics by Dataset

Dataset	Average Time (Mins)	Minimum Time (Mins)	Maximum Time (Mins)	Total Time (Mins)
CalMS21	4.30	1.02	9.98	107.57
Freezing	9.04	4.01	9.98	81.33
Grooming	2.87	0.44	9.98	71.68
Rodent Ventral 1	6.57	0.34	8.33	183.89
Rodent Ventral 2	9.26	8.33	9.98	148.19
Scratch-AID	9.98	9.98	9.99	149.77

C.3 CaLMS21

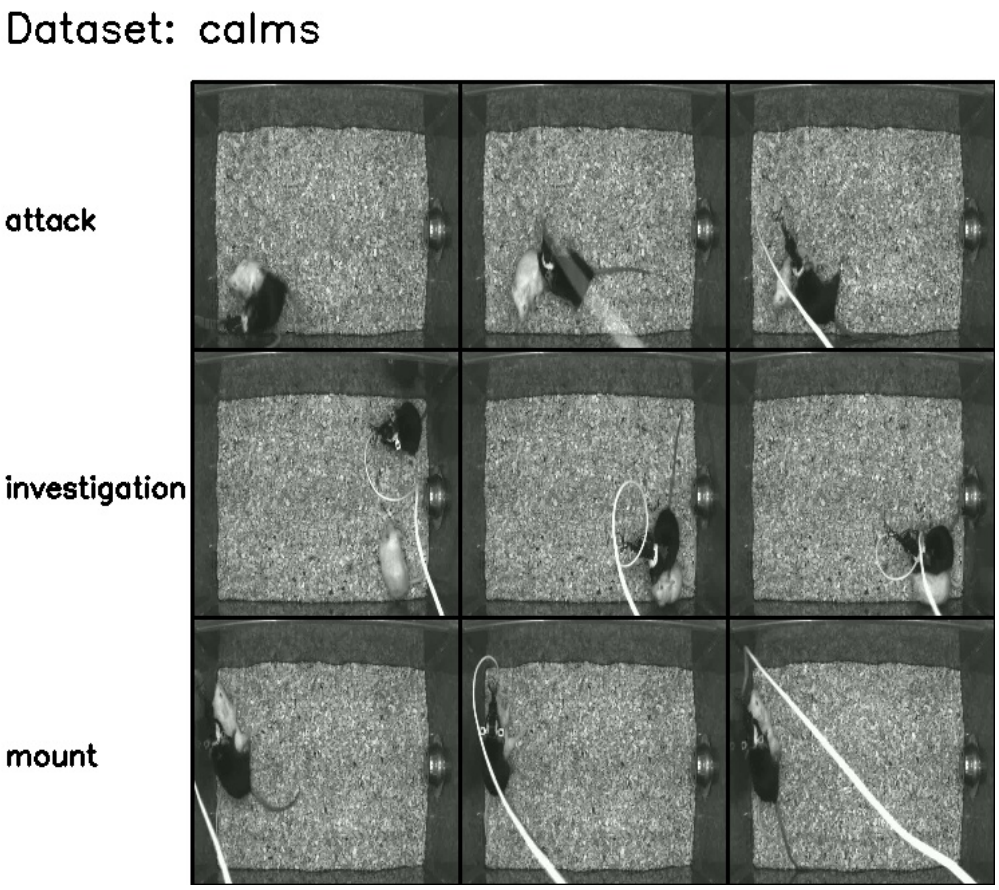


Figure 5: CaLMS21 Behaviors

C.4 Rodent Grooming

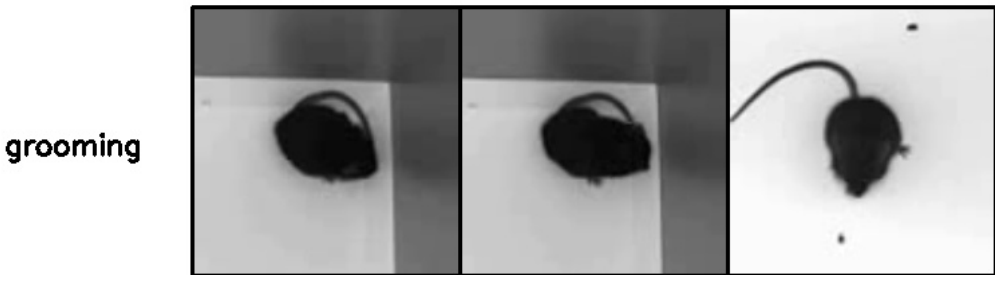


Figure 6: Rodent Grooming Behaviors

C.5 Mouse-Ventral 1&2

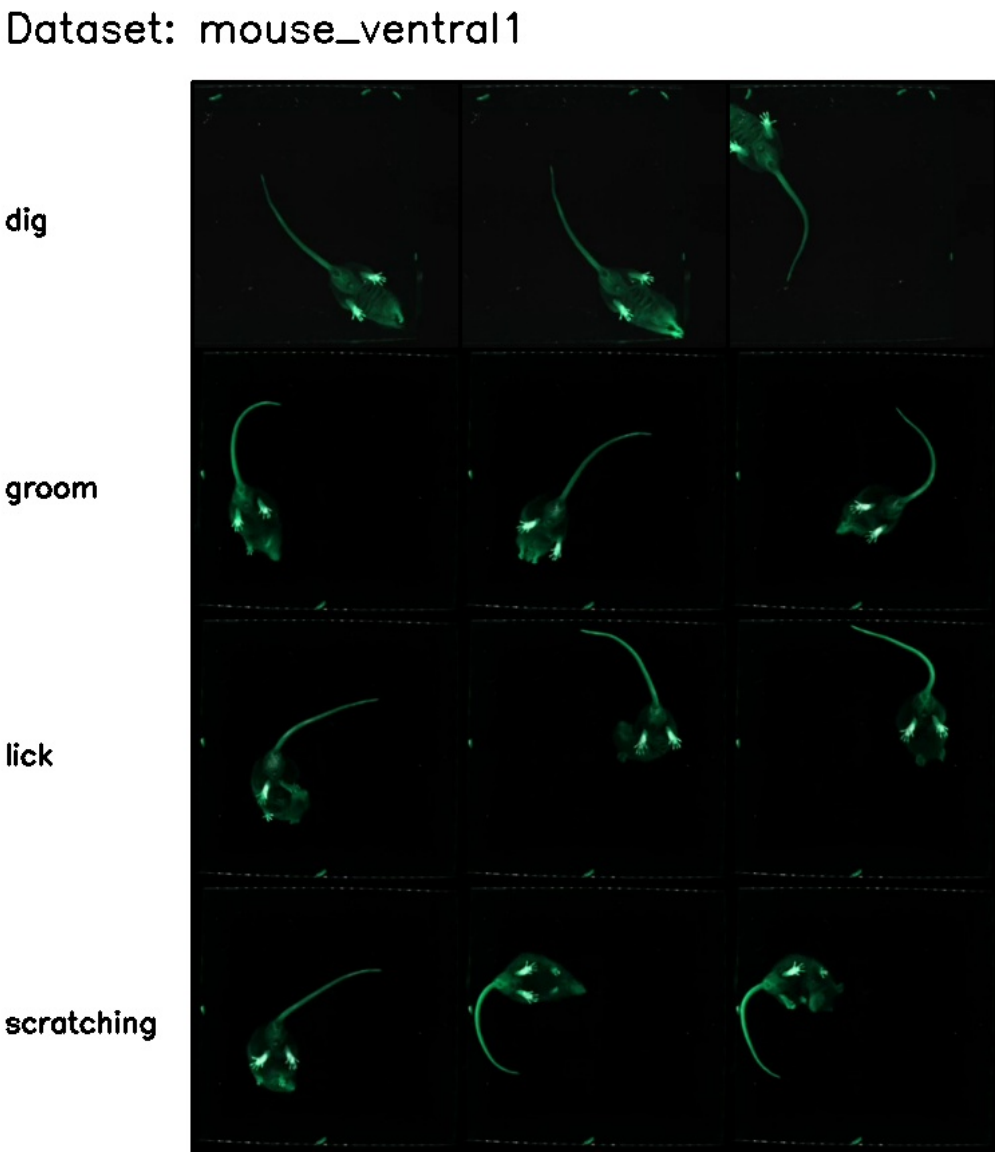


Figure 7: Mouse-Ventral 1 Behaviors

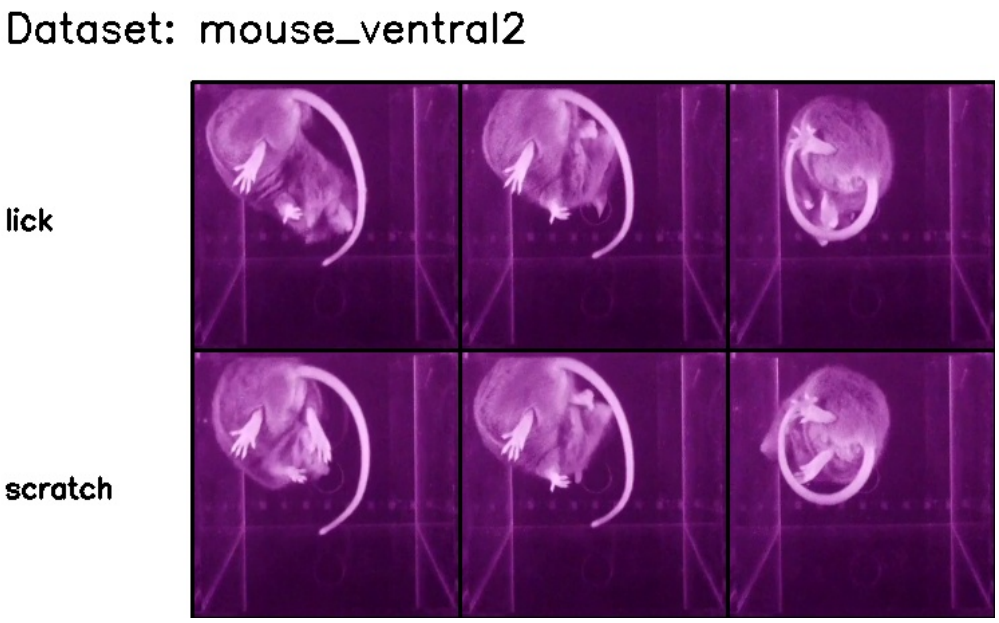


Figure 8: Mouse-Ventral 2 Behaviors

C.6 Scratch-AID

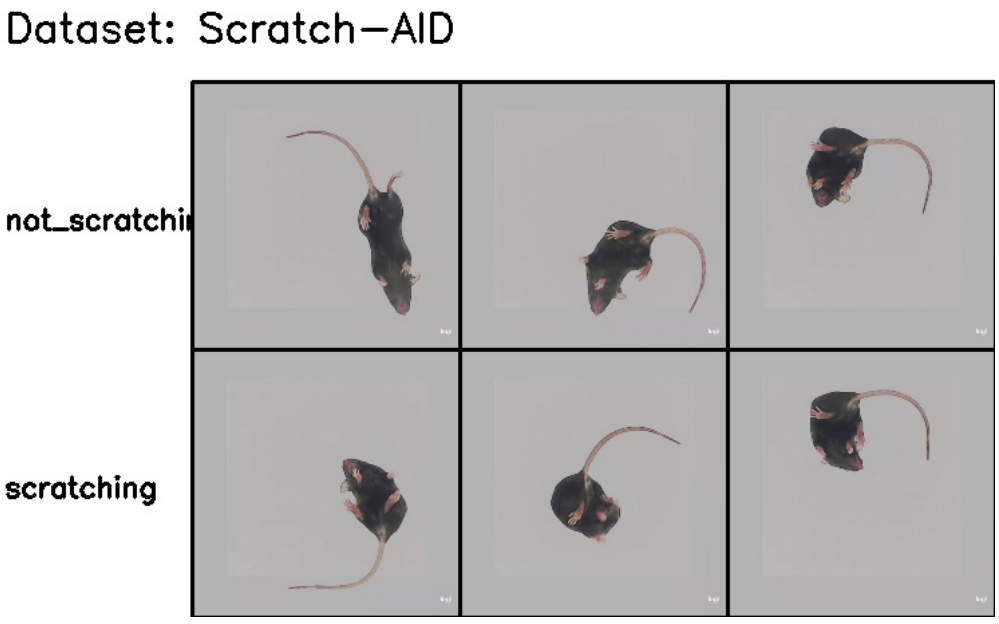


Figure 9: Scratch-AID Behaviors

C.7 Freezing

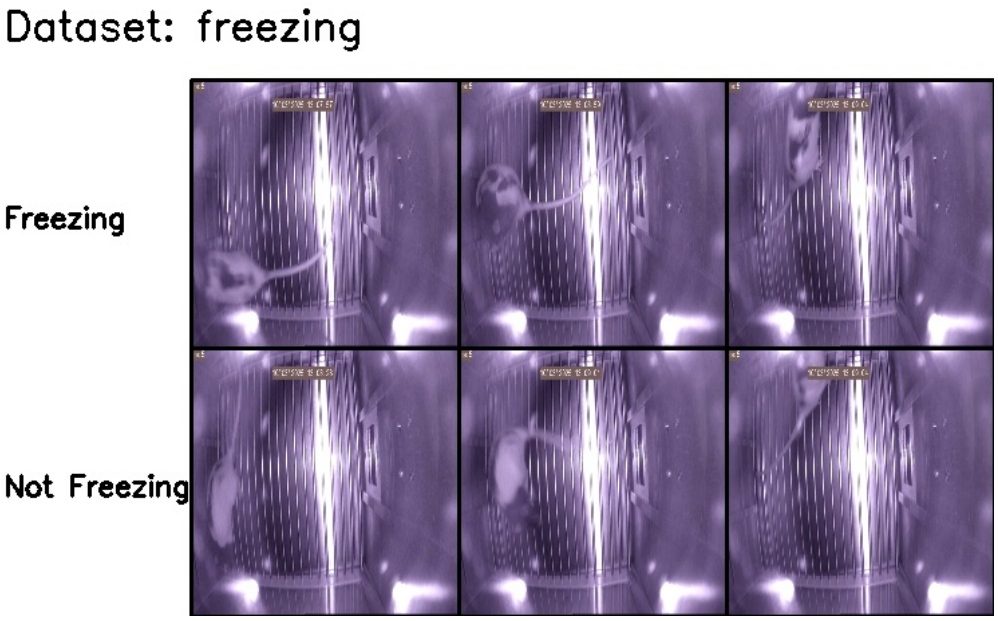


Figure 10: Freezing Behaviors

D Gemini-Pro Results

D.1 Rodent-Bench

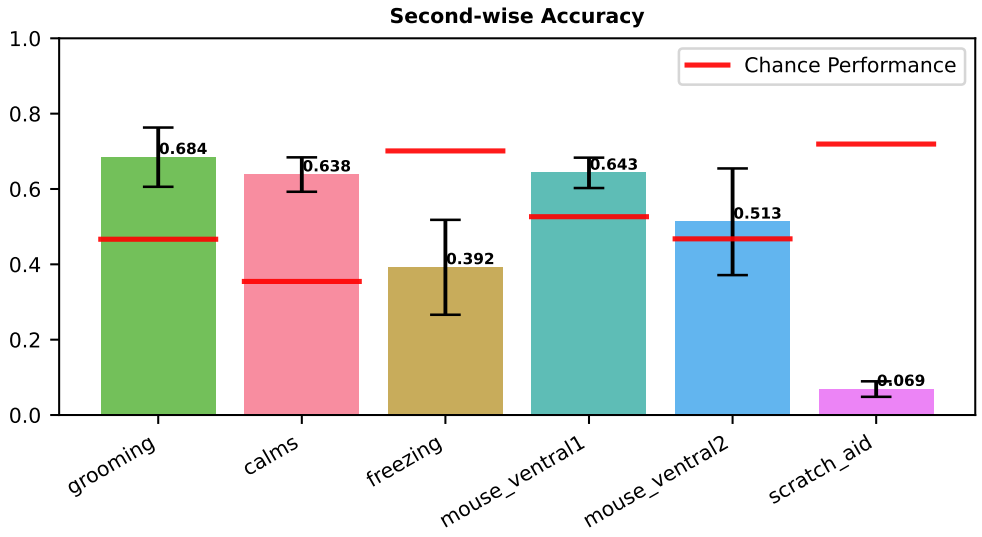


Figure 11: Per second accuracy

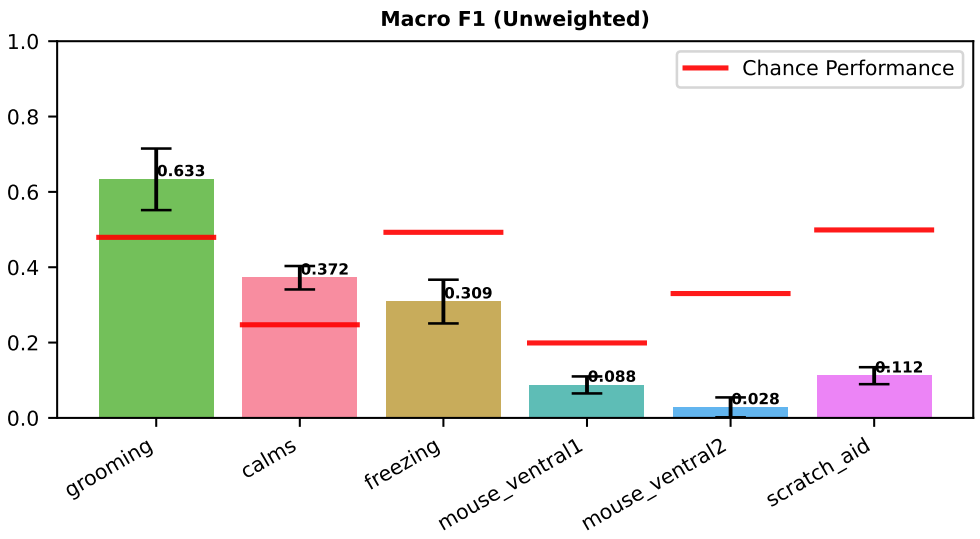


Figure 12: Macro F1 score

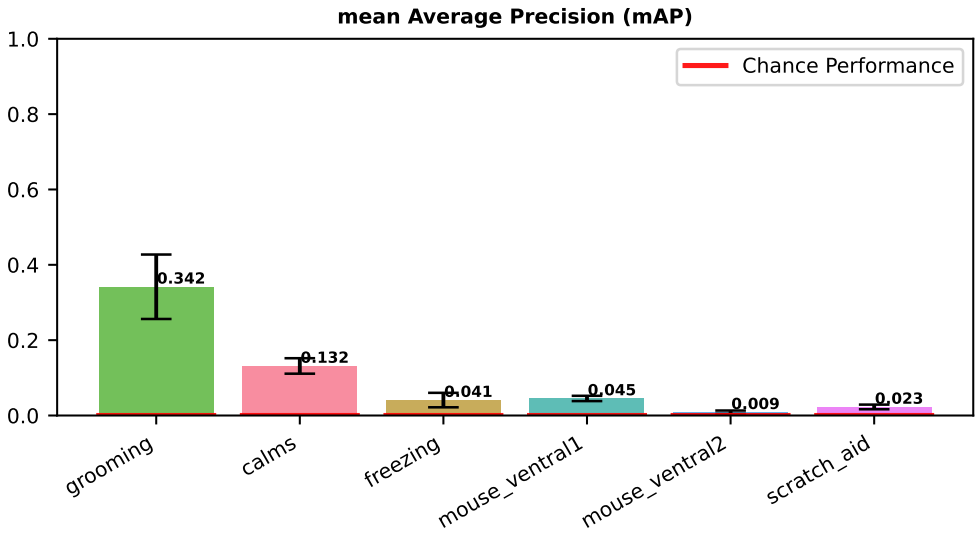


Figure 13: Per second accuracy

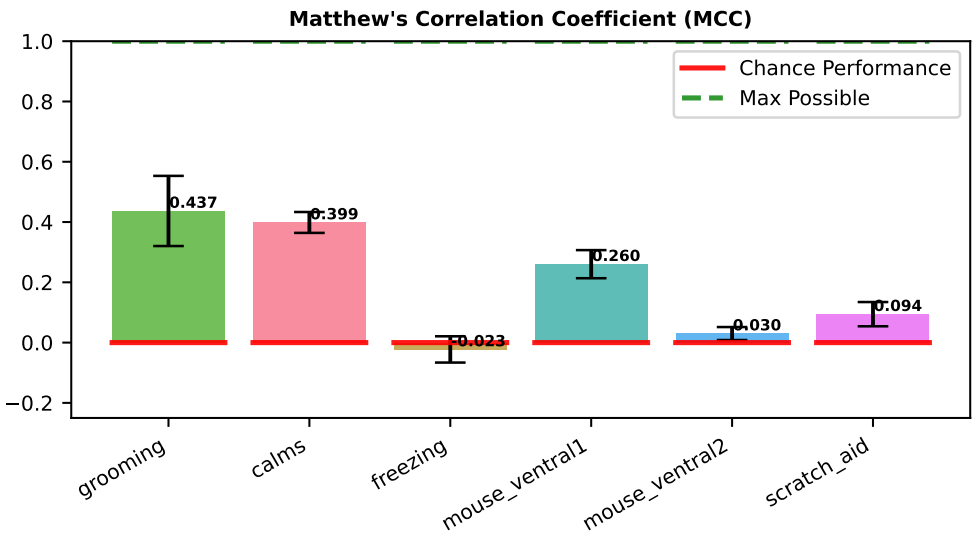


Figure 14: Matthew's Correlation Coefficient

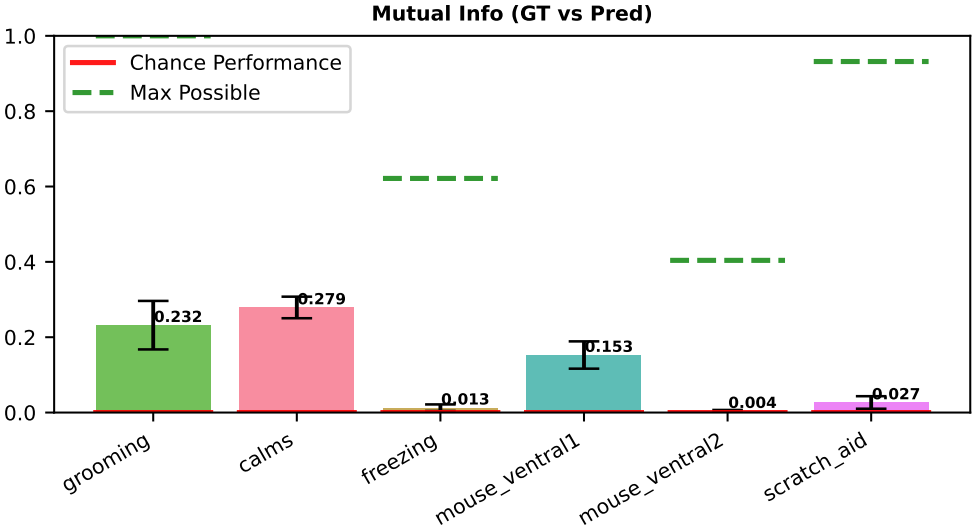


Figure 15: Mutual information between ground truth and predictions

D.2 Rodent-Bench-Short

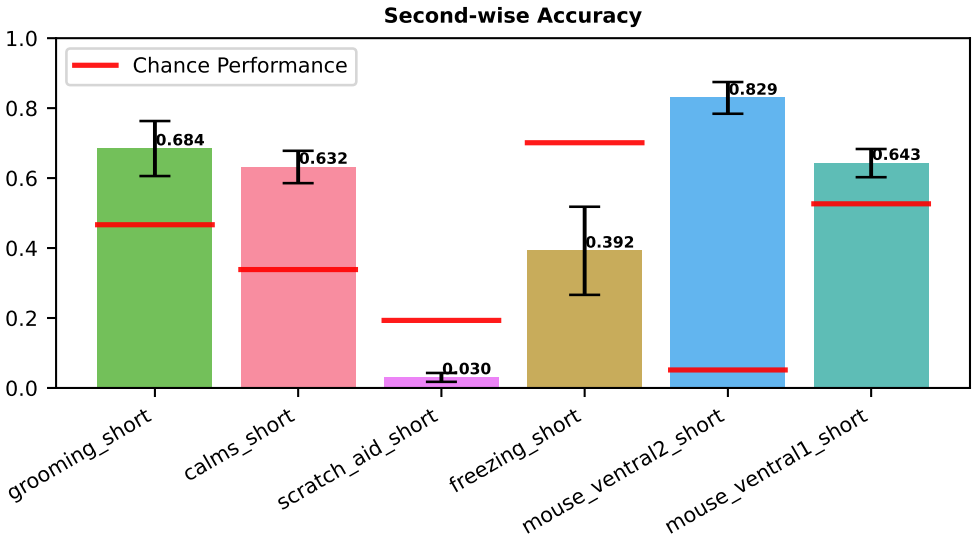


Figure 16: Per second accuracy

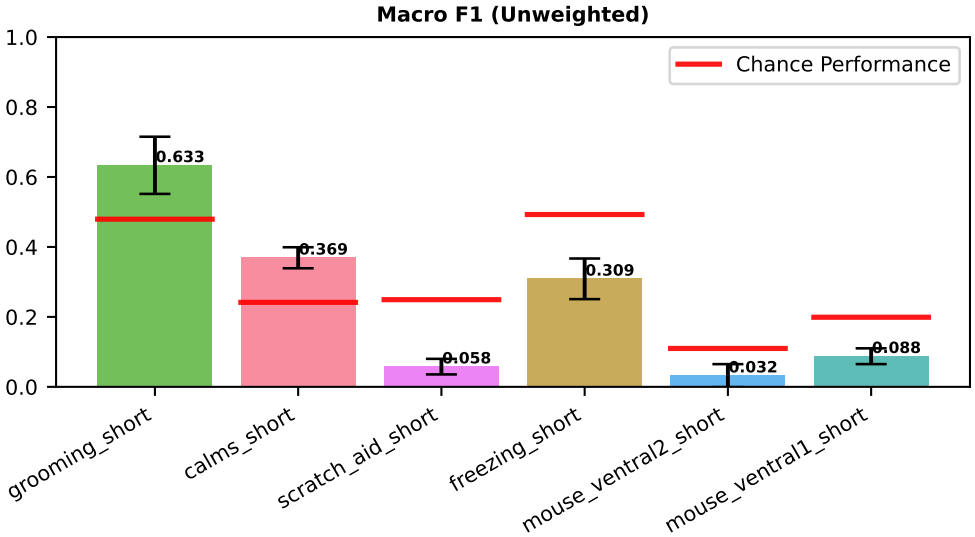


Figure 17: Macro F1 score

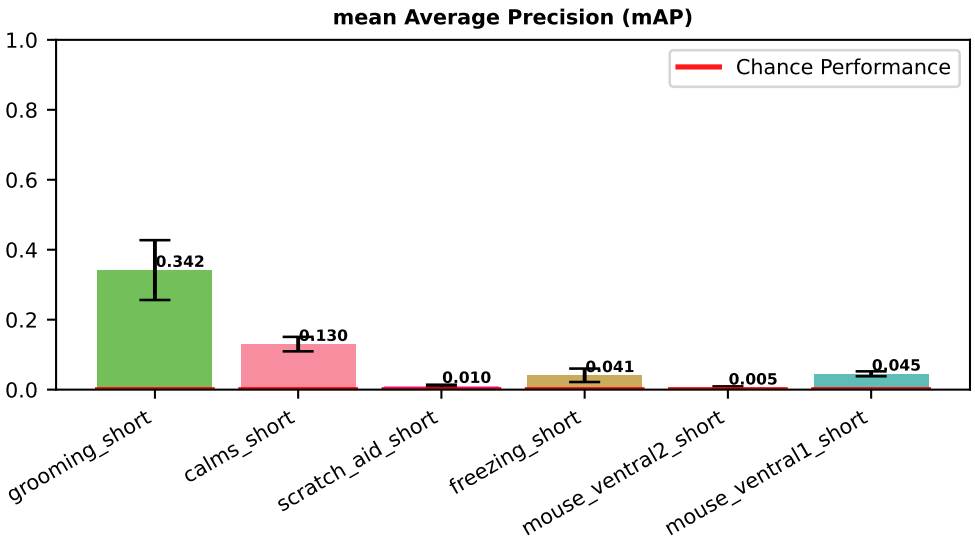


Figure 18: Per second accuracy

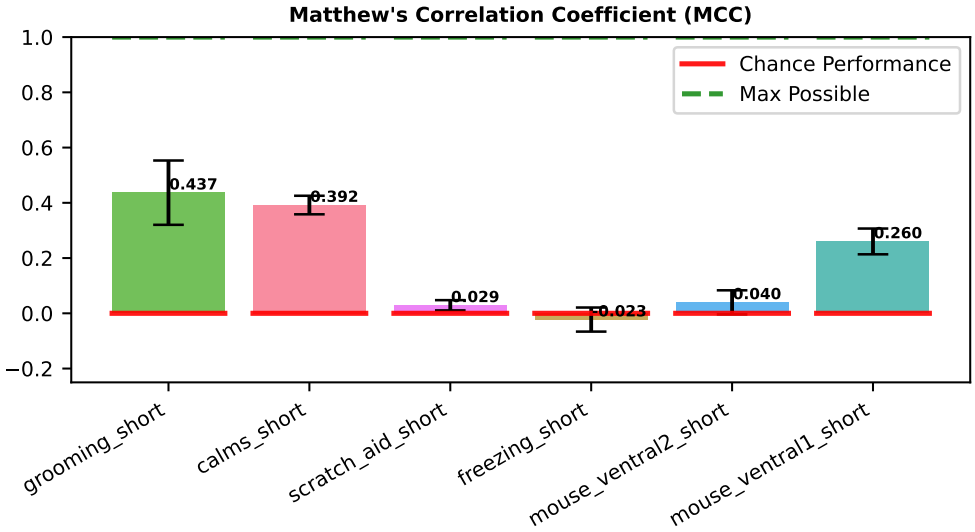


Figure 19: Matthew's Correlation Coefficient

E Gemini-Flash Results

E.1 Rodent-Bench

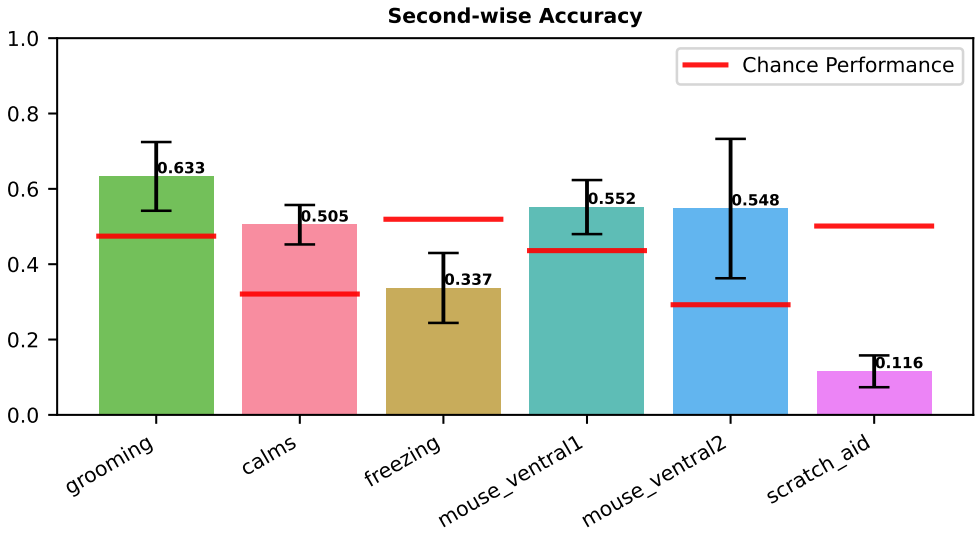


Figure 20: Per second accuracy

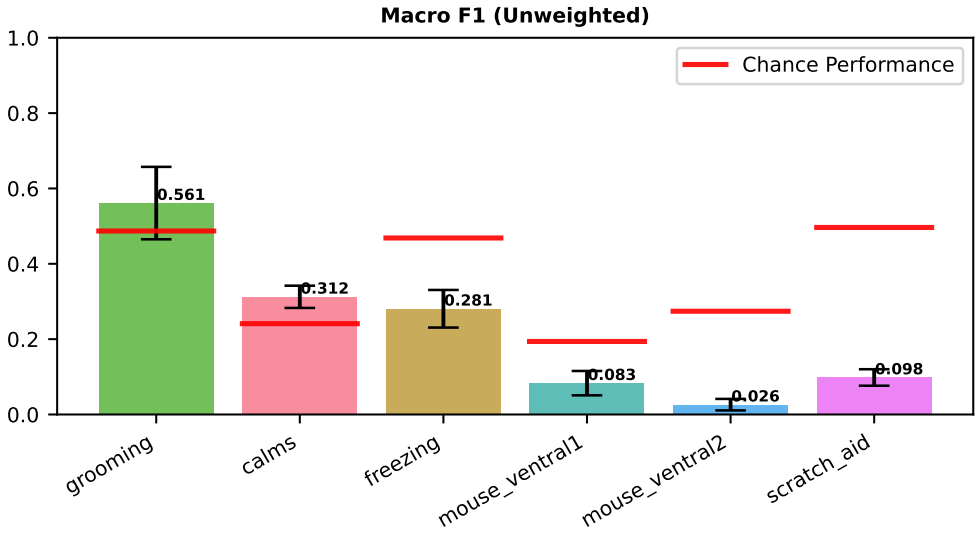


Figure 21: Macro F1 score

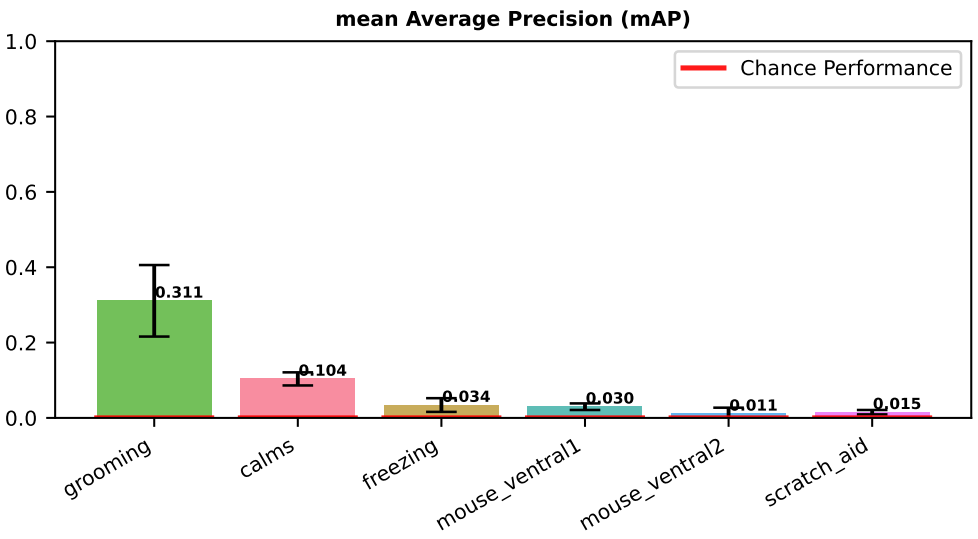


Figure 22: Per second accuracy

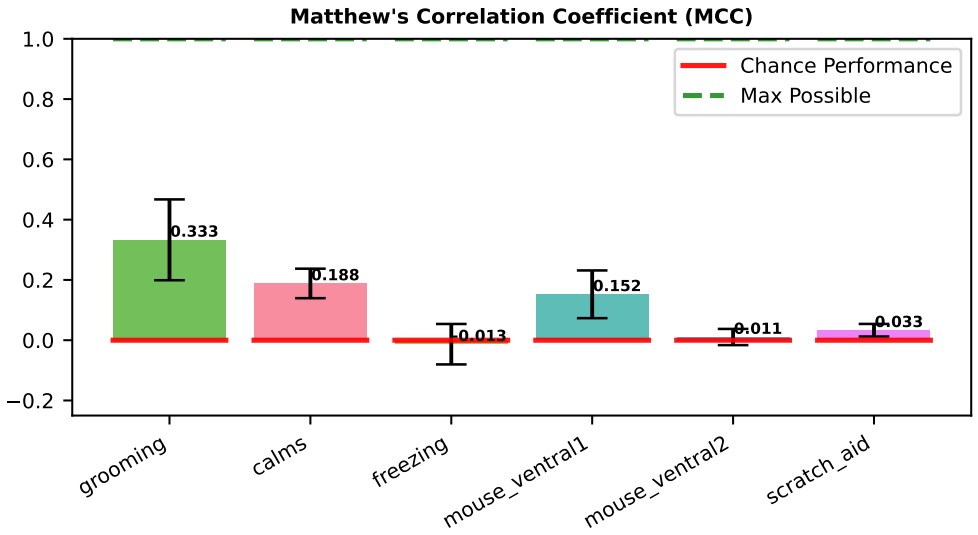


Figure 23: Matthew's Correlation Coefficient

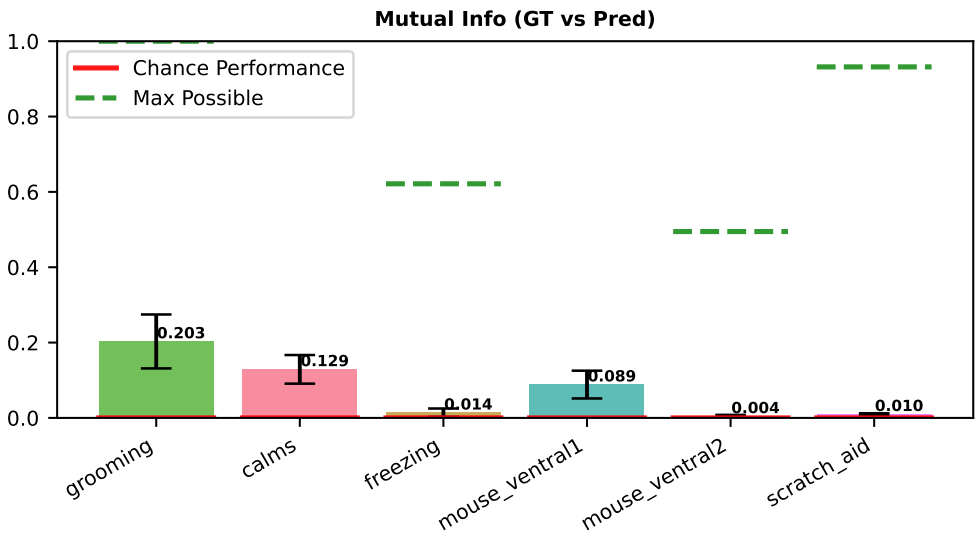


Figure 24: Mutual information between ground truth and predictions

E.2 Rodent-Bench-Short

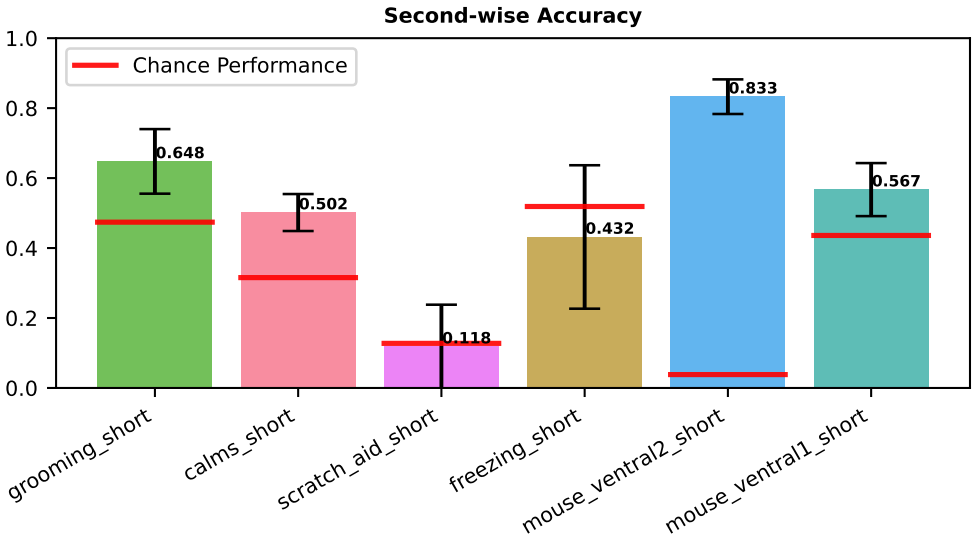


Figure 25: Per second accuracy

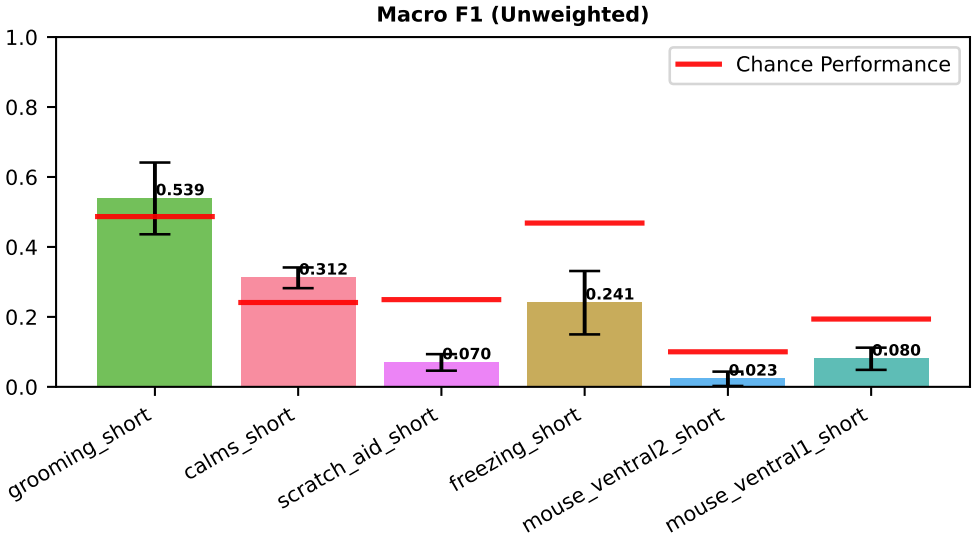


Figure 26: Macro F1 score

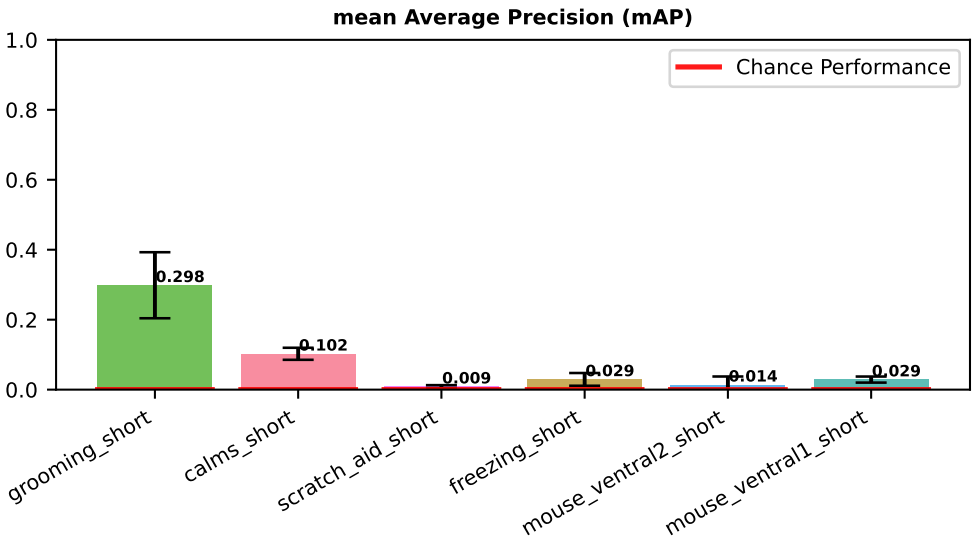


Figure 27: Per second accuracy

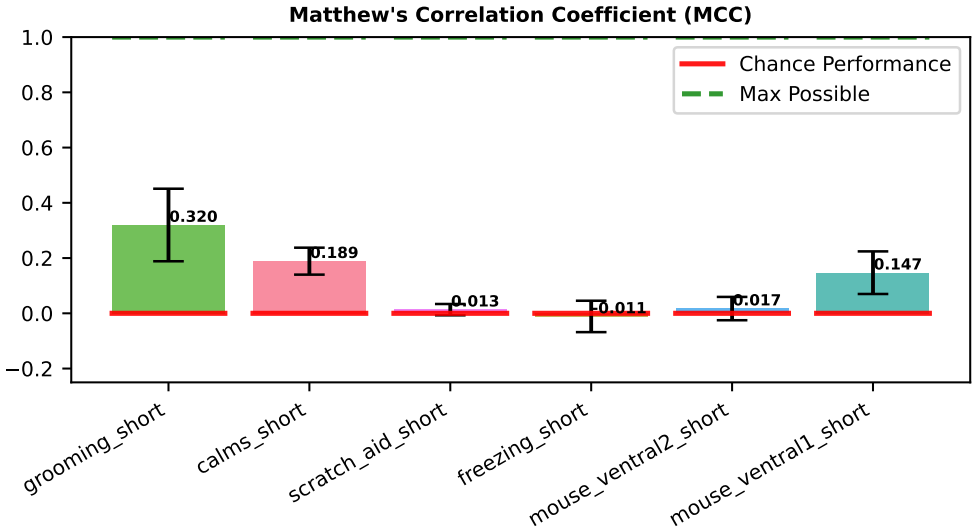


Figure 28: Matthew's Correlation Coefficient

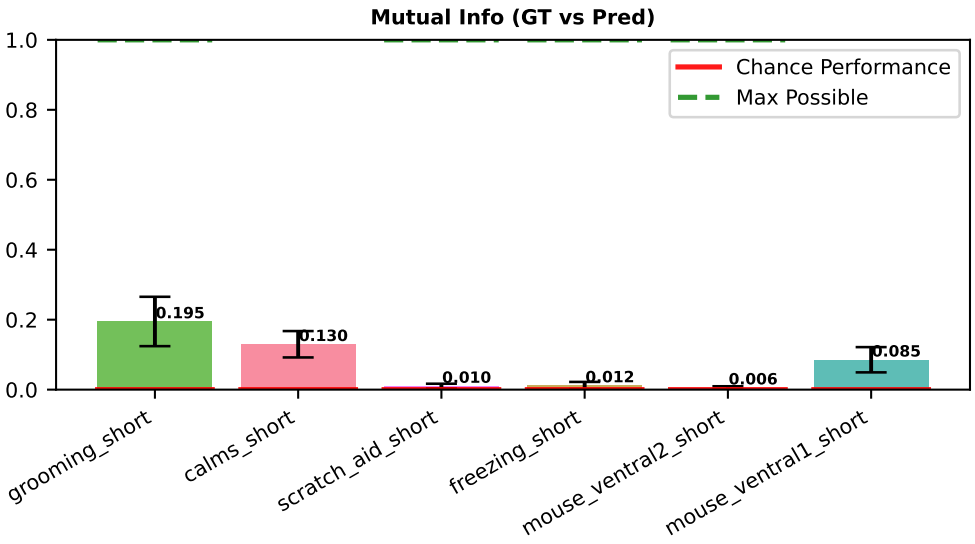


Figure 29: Mutual information between ground truth and predictions

F Qwen-VL-Max Results

Because the Qwen-VL-Max can't ingest videos longer than 10 minutes we only have results for Rodent-Bench-Short.

F.1 Rodent-Bench-Short

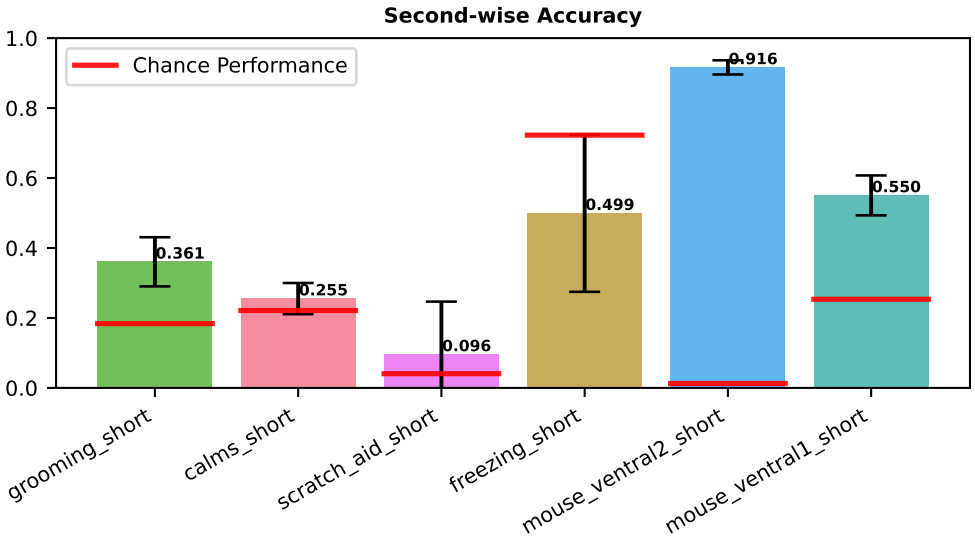


Figure 30: Per second accuracy

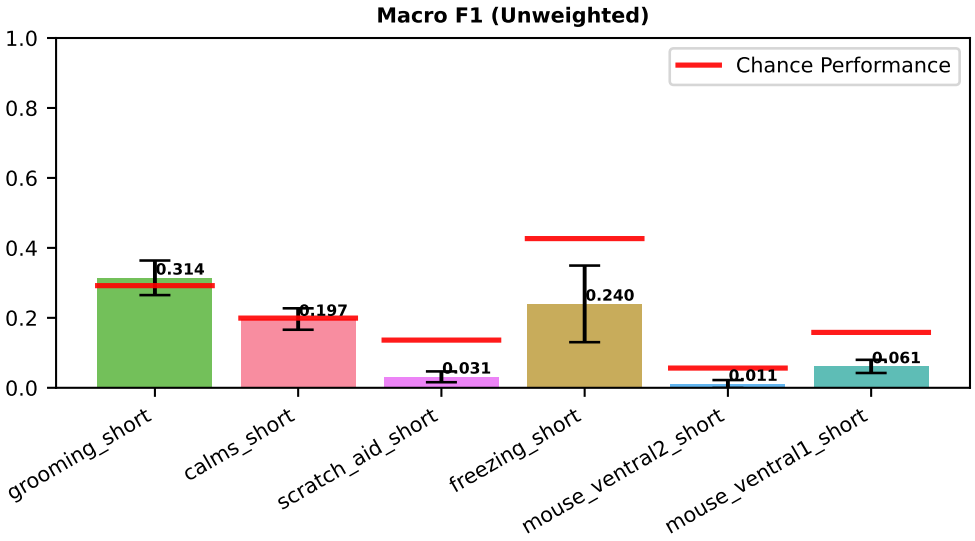


Figure 31: Macro F1 score

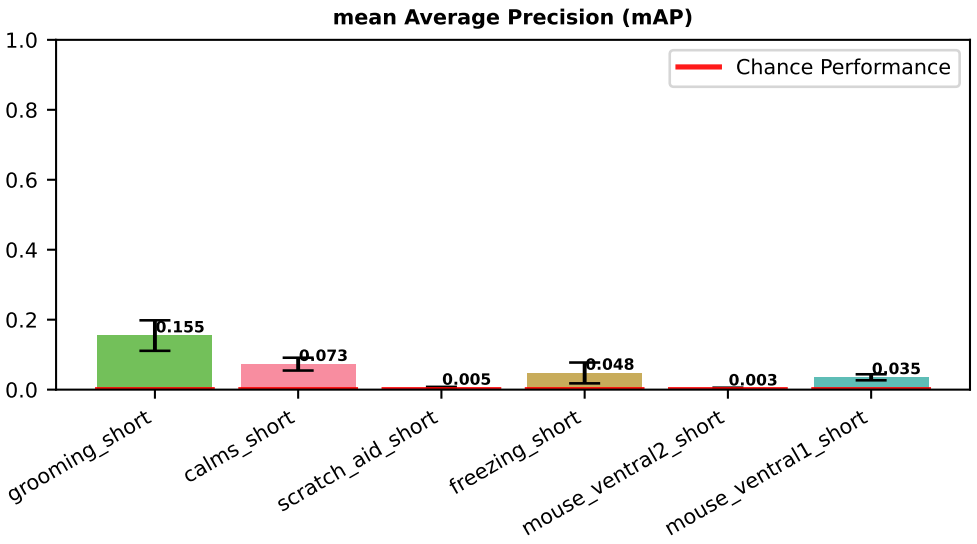


Figure 32: Per second accuracy

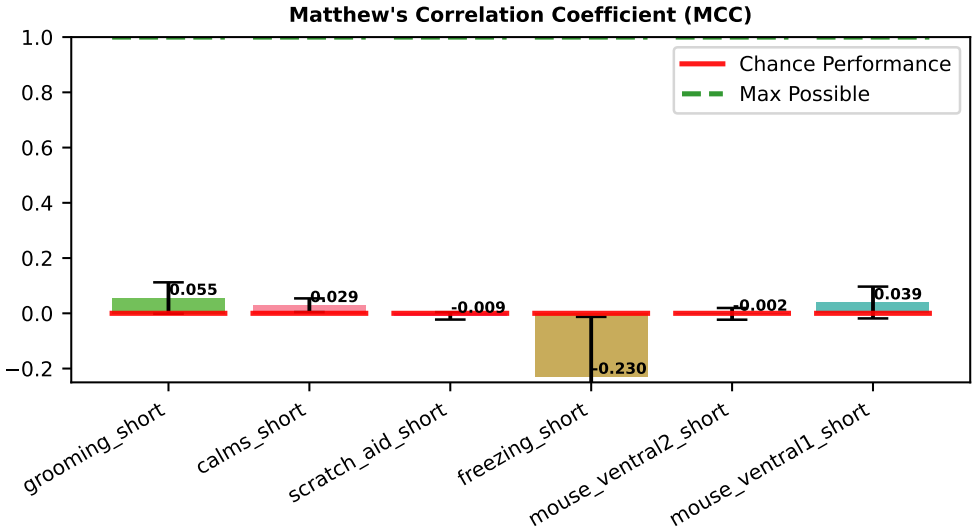


Figure 33: Matthew's Correlation Coefficient

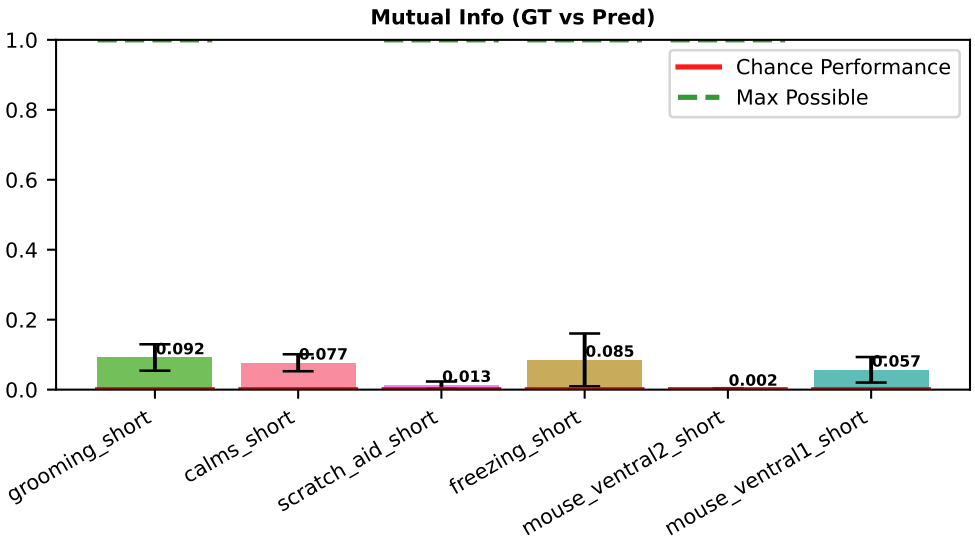


Figure 34: Mutual information between ground truth and predictions

G Prompt Templates

We provide the complete prompt templates used for each dataset. All prompts follow a consistent structure: role definition, task description, available behavior labels, formatting requirements, and JSON output schema.

G.1 CalMS21 Social Behaviors

```
You are a Rodent Behavior Labeler specializing in rodent social behavior.
Your task is to analyze a video of rodents and segment it into periods of distinct behaviors.

Available behavior labels:
- attack - when the black rodent is attacking another rodent
- investigation - when the black rodent is investigating another rodent
- mount - when the black rodent is mounting another rodent
- other - when the black rodent is doing something else

Important:
You must use ONLY the labels listed above. Do not create new labels or modify existing ones.

Start your analysis from the start of the video and continue until the end of the video.

For each segment, provide:
- segment number (in order)
- start and end time in MM:SS format
- behavior label (must be one of the above labels)

Your response must be in JSON format with the following structure:
{
    "segments": [
        {
            "start_time": MM:SS,
            "end_time": MM:SS,
            "behavior": "behavior_label",
            "segment_number": INTEGER,
        },
        ...
    ]
}
```

G.2 Scratch-AID

```
You are a Rodent Behavior Labeler specializing in telling when a rodent is scratching.
Your task is to analyze a video of rodents and segment it into periods of distinct behaviors.

Available behavior labels:
- scratching - when the rodent is scratching, usually with the hind legs
- not scratching - when the rodent is not scratching.

Important:
You must use ONLY the labels listed above. Do not create new labels or modify existing ones.

Start your analysis from the start of the video and continue until the end of the video.
```

```
The video is of a rodent and is taken from below.

For each segment, provide:
- start and end time in MM:SS format
- segment number (in order)

Your response must be in JSON format with the following structure:
{
    "segments": [
        {
            "segment_number": INTEGER,
            "start_time": MM:SS,
            "end_time": MM:SS,
            "behavior": "behavior_label",
        },
        ...
    ]
}
```

G.3 Rodent Grooming Detection

```
You are a Rodent Behavior Labeler specializing in identifying grooming behaviors in rodents.
Your task is to analyze a video of rodents and segment it into periods of distinct behaviors.
The video shows a rodent from above.

Available behavior labels:
- grooming - when the rodent is actively grooming itself (e.g., licking, scratching, cleaning fur)
- other - when the rodent is not grooming (e.g., walking, exploring, resting)

Important:
You must use ONLY the labels listed above. Do not create new labels or modify existing ones.
Grooming behaviors are characterized by:
- Repetitive movements of paws over the face or body
- Licking of fur or paws
- Scratching with hind legs
- Cleaning of specific body parts

Start your analysis from the start of the video and continue until the end of the video.

For each segment, provide:
- segment number (in order)
- start and end time in MM:SS format
- behavior label (must be one of the above labels)

Your response must be in JSON format with the following structure:
{
    "segments": [
        {
            "segment_number": INTEGER,
            "start_time": MM:SS,
            "end_time": MM:SS,
            "behavior": "behavior_label",
        },
        ...
    ]
}
```

G.4 Freezing Behavior

```
You are a Rodent Behavior Labeler specializing in identifying freezing behaviors in rodents.
Your task is to analyze a video of rodents and segment it into periods of distinct behaviors.
The video shows a rodent from above.

Available behavior labels:
- Freezing - when the rodent is Freezing, i.e characterized by the complete cessation of movement, except for respiratory-related movements so no head twitching for instance.
- Not Freezing - when the rodent is not Freezing

Important:
You must use ONLY the labels listed above. Do not create new labels or modify existing ones.

Start your analysis from the start of the video and continue until the end of the video.

For each segment, provide:
- segment number (in order)
- start and end time in MM:SS format
- behavior label (must be one of the above labels)

Your response must be in JSON format with the following structure:
{
    "segments": [
        {
            "segment_number": INTEGER,
            "start_time": MM:SS,
            "end_time": MM:SS,
            "behavior": "behavior_label",
        },
        ...
    ]
}
```