Evaluating Large Multimodal Language Models for Automated Ethogram Generation from Public Zoo Footage

Sriman Ratnapu
University of California, Santa Cruz

Abstract

Ethograms are key behavioral research tools allowing for structured analysis of observable behaviors over time. However, their construction can be time consuming, subjective, and cannot easily scale to large video datasets. With the emergence of large multimodal language models (LLMs), automatic approaches to components of the ethogram annotation pipeline may now be possible. However, their accuracy compared to human annotators has not been rigorously tested.

In this work, we present a controlled evaluation of LLM-generated ethogram labels by comparing them against human annotations across captive animal species using publicly accessible zoo footage. We adopt a focal-animal sampling protocol, a coarse-grained behavior taxonomy, fixed temporal windowing, and a coverage-driven clip selection strategy designed explicitly to evaluate annotation reliability rather than behavioral prevalence. We further examine the effect of visual temporal context through a frame-sampling ablation, comparing one-frame, three-frame, and five-frame representations of each temporal window. Our results show that LLMs achieve moderate agreement with human annotators for visually salient behaviors, while performance degrades for socially complex and underrepresented categories. Error analysis reveals systematic and interpretable failure modes rather than random noise, suggesting that multimodal LLMs may be useful as assistive tools for ethogram pre-labeling and triage, but not as replacements for expert human annotation.

1. Introduction

Ethograms (systematic time-indexed catalogs of observable animal behavior) are used in the study of animal behavior as well as zoological welfare assessment and captive habitat design. Ethograms allow researchers to quantify behavior patterns, compare them between individuals and environments, and develop objective measures of enrichment, health, and management interventions. Typically, ethograms are constructed by human experts who manually annotate footage with relevant behavior codes from a pre-curated list.

Manual construction of ethograms allows for high-quality results but can be laborious, expensive, and prone to inter-annotator error. For this reason, much long-term zoo monitoring footage as well as user-uploaded public footage are left unlabeled. Animal behavior automatic analysis has been sought after for this reason.

Previous attempts at automatically classifying animal behaviors have utilized computer vision and machine learning approaches under heavy supervision. Methods using species specific models, pose estimation pipelines, or massive labeled datasets filmed in controlled environments have demonstrated high performance on their narrow task but require expensive re-training and careful domain constrained data collection.

Recent multimodal large language models (LLMs), trained on visual inputs paired with explicit structured natural-language instructions, represent one promisingly more-general alternative. Instead of training classifiers for species-specific behavior vocabularies, one could ask LLMs to produce ethogram labels conditioned directly on visual inputs by deferring to its vast visual and semantic priors. However, despite recent enthusiasm around multimodal models, there has been scarce empirical study of whether LLM-produced ethograms qualitatively agree with human annotations in a controlled, reproducible setting. Motivated by these practical considerations experienced during volunteer-based deployment with monitoring of captive animals, we were faced with the challenge of obtaining scalable, quantitative ethograms from video footage. In this work, instead of introducing a novel recognition model, we ask a simpler question: how well can a general-purpose multimodal LLM replicate human ethogram annotations, and in what ways does it fail systematically?

Code and dataset metadata are publicly available in the accompanying repository.

2. Related Work

2.1 Ethogram Construction and Behavioral Sampling

Early ethology work defined observational sampling methods like focal-animal sampling and scan sampling, advocating for principled operationalization of behavior classes and explicit treatment of observer bias and variation (Altmann, 1974). These practices continue to influence behavioral research today, including research taking place in zoos.

For example, common textbook advice still includes using mutually exclusive behavior categories, providing explicit rules for annotators, and calculating inter-annotator statistics to measure agreement. (Martin & Bateson, 2007) In this case specifically, Cohen's κ is often employed as a statistical measure of annotator agreement above chance for nominal labels. (Cohen, 1960)

There have been numerous proposals for assisting with ethogram annotation by offloading some of the effort required to humans. These approaches have traditionally focused on laboratory settings or other highly constrained situations. These systems may accelerate video playback for human annotators or automatically segment videos, but ultimately defer to a human decision. Many of these approaches make strong assumptions about the target species, require stationary cameras, or rely on extensive manual tuning and therefore do not generalize to arbitrary zoo footage.

Our experiments follow typical ethogram practices with regards to sampling from video. We do not introduce a novel annotation protocol. Given standard sampling of focal-animal behavior, we ask if general-purpose multimodal models can replicate labels produced by human observers.

2.2 Vision-Based Animal Behavior Recognition

Large amounts of literature exists on automated animal behavior recognition through computer vision and machine learning techniques. Classical work consists of handcrafted features and trajectory-based approaches while more modern work implements deep learning pipelines for pose estimation and action recognition. Pose based systems like DeepLabCut allow for frame-by-frame user specified body part localization and have been used to measure fine-grained behaviors in laboratory animals (Mathis et al., 2018).

Past work has also trained supervised models on either single behaviors or temporal sequences of behaviors ("behavior bouts") with temporal convolutional or recurrent models after pose estimation. These models can reach high accuracies in controlled environments but often require extensive labeled training data specific to a species and trained under consistent recording settings. Generalization has proven difficult between species, enclosure types, camera angles, and recording quality.

In contrast to past work, the method we test here does not train a behavior recognition model. Instead we query whether a pretrained multimodal large language model can serve as a zero-shot ethogram annotator if provided a handful of representative frames from a video as well as explicit labeling instructions. For this reason we frame our work as supplementary to supervised vision pipelines instead of a replacement.

2.3 Multimodal Language Models for Visual Annotation

Recent multimodal large language models have achieved promising results when conditioned on visual inputs, performing natural-language directed tasks that output machine-readable information. This capability has led to interest in using LLMs as "digital labelers" to annotate content in an agnostic manner.

However, explicit application of LLMs to behavior data has been largely unexplored. We are aware of no prior work that has directly compared LLM-produced ethograms to human-generated labels within the structure of traditional behavioral sampling methods or agreement metrics. Most publications that mention LLMs within animal contexts cite expected behavior either qualitatively in natural-language summaries or make predictions about welfare in hypothetical or forecasting scenarios, as opposed to directly assessing labeling agreement within behavioral budgets.

For this reason, we do not frame this work as establishing LLMs as a replacement for behavior analysts. Rather, we assess LLMs as a potential annotation tool. We report agreement with human experts, characterize systematic errors, and caveat limitations.

3. Dataset and Sampling Strategy

3.1 Video Sources and Ethics

All video clips used in this study were obtained from publicly accessible zoo videos, primarily from official zoo channels. Each clip is referenced by a stable URL and timestamp to ensure reproducibility. Short excerpts were temporarily downloaded solely for internal annotation and model inference, after which the video files were deleted.

To avoid redistribution of copyrighted material, raw video files are not included with this paper. Metadata sufficient to reconstruct the dataset—including source URLs, timestamps, and clip identifiers—along with analysis code and the LLM prompt used in this study, are available at: https://github.com/srimanratnapu/ethograms_with_llms

3.2 Coverage-Driven Clip Selection

The objective of this study is to evaluate **annotation reliability**, not to estimate behavioral prevalence or time budgets. Accordingly, clip selection follows a **coverage-driven strategy** rather than random sampling.
Specifically:

- Short clips of 60–120 seconds are selected to increase the likelihood of observing multiple behavior types.

- Clips dominated entirely by a single behavior, such as prolonged inactivity, are avoided when possible.

- Selection is performed prior to any model evaluation and does not depend on LLM outputs.

This design intentionally alters the natural frequency distribution of behaviors. As a result, reported behavior frequencies should not be interpreted as population-level estimates. Instead, the dataset is constructed to stress-test annotation agreement across behavior categories.

3.3 Species Scope

Elephants, giraffes, and giant pandas are the three species included in our dataset. All three species are from zoos. We chose these species due to the existence of high-quality public footage. Our dataset is not meant to generalize to all species.

4. Annotation Protocol

4.1 Focal-Animal Sampling

To reduce ambiguity caused by multiple animals in one scene, our dataset follows the focal-animal sampling paradigm. For each video clip, one focal animal is predetermined for annotation. Ideally, the predetermined focal animal should be visible throughout the entire clip. If multiple animals are present, the largest animal or the animal with the largest visible portion at the beginning of the clip is typically chosen to be the focal animal. All labels are associated with this single focal animal throughout the duration of the clip.
If the focal animal is not visible for most of a timestamp, we label that timestamp as **OutOfFrame**.

4.2 Temporal Windowing

Each clip is divided into non-overlapping windows of 10 seconds each. For each window, we sample one behavior that best represents the behavior the animal performs for most of the window duration. If no behavior is performed for more than half of the window, we label that window as **Uncertain**.
We chose to use single labels to facilitate agreement study and to remove ambiguity caused by using multiple labels.

4.3 Behavior Taxonomy

We employ a coarse-grained, mutually exclusive behavior taxonomy:

- **Resting**: Lying or sitting with no locomotion
- **Locomotion**: Walking, climbing, or swimming
- **Feeding**: Eating or manipulating food
- **Social**: Directed interaction with another animal
- **ObjectInteraction**: Manipulating toys or enclosure objects
- **OutOfFrame**: Focal animal not visible
- **Uncertain**: Behavior cannot be confidently determined

The taxonomy prioritizes observability and inter-annotator consistency over fine-grained behavioral distinctions.

5. LLM-Based Annotation Method

5.1 Visual Representation and Frame Sampling

Multimodal LLM APIs do not uniformly support streaming videos as input. Therefore, we represent each 10-second interval with few uniformly sampled frames extracted using ffmpeg.
Unless otherwise mentioned, we use three frames (start, middle, end) of the ten-second interval. We perform a frame-sampling ablation to test the impact of visual temporal context using::

- one frame (middle only),
- three frames (start, middle, end),
- five frames (uniformly sampled).

5.2 Prompt Design

The model is instructed to assign exactly one behavior label per interval using only visible evidence and without inferring internal states. The prompt explicitly defines each behavior category and requires structured JSON output including a confidence score.
The prompt is frozen across all experiments to ensure reproducibility.

6. Evaluation Metrics

We compute:

- **Human–human agreement** using Cohen's κ (Cohen, 1960),
- **Human–LLM agreement** using per-class precision, recall, and F1,
- **Confusion matrices** to identify systematic errors.

Intervals labeled **OutOfFrame** or **Uncertain** by human consensus are excluded from the primary evaluation subset, as these labels often reflect visibility constraints rather than interpretable behavior.

7. Results

7.1 Human–Human Agreement

Human annotators exhibit very high agreement (Cohen's $\kappa \approx 0.95$), indicating that the focal-animal protocol and coarse-grained taxonomy support reliable annotation across diverse scenes.

7.2 Human–LLM Agreement

On the filtered evaluation set, the LLM achieves moderate agreement with human consensus. Performance varies by behavior category:

- Visually salient behaviors such as resting and feeding achieve the highest agreement.
- Locomotion and object interaction show moderate performance.
- Social behavior exhibits substantially lower recall.

Class-wise performance and systematic confusions are quantified in Table 2 and visualized in Figures 1 and 2.

7.3 Effect of Frame Sampling Density

Sampling density of frames used as input to the LLM does affect performance. Single frame input has the lowest accuracy and macro-F1 suggesting the temporal context is lacking. Gains in accuracy can be seen going from one frame to three frames but macro-F1 does not increase and can decrease due to shifts in class-wise errors when there is class imbalance. Five frames had the highest accuracy and macro F1 but showed diminishing returns compared to three frames.
Based on these results it can be inferred that not much temporal context is needed to provide coarse labels for ethograms and that there are limited benefits past including only a few frames.

Table 1. Effect of Frame Sampling Density on Performance (Filtered Set)

Frames	N	Accuracy	Macro-F1
1 frame	311	0.370	0.317
3 frames	311	0.424	0.306
5 frames	311	0.463	0.348

7.4 Error Analysis and Underrepresented Behaviors

Misclassifications are systematic, and not necessarily random. Social behaviors are often mistaken as either locomotion or resting especially when there are multiple animals in frame. Behavioral transitions that occur over a short amount of time within the ten second window get omitted and classified as the predominant behavior within that window. Social has limited examples because it occurs less frequently in public zoo datasets and is less likely to be sampled due to our coverage driven sampling method.

Table 2. Per-Behavior Precision, Recall, and F1 (Filtered Set)

Behavior	Support (n)	Precision	Recall	F1
Resting	146	0.932	0.473	0.627
Feeding	70	0.611	0.471	0.532
Locomotion	36	0.410	0.444	0.427
ObjectInteraction	47	0.389	0.447	0.416
Social	12	0.455	0.417	0.435

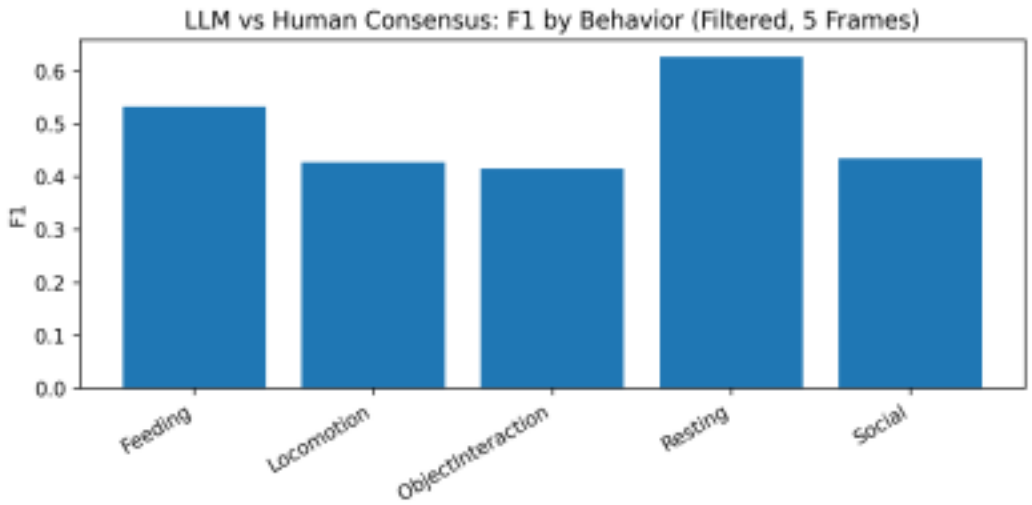

Figure 1. F1 by Behavior (Filtered Evaluation Set, 5 Frames)

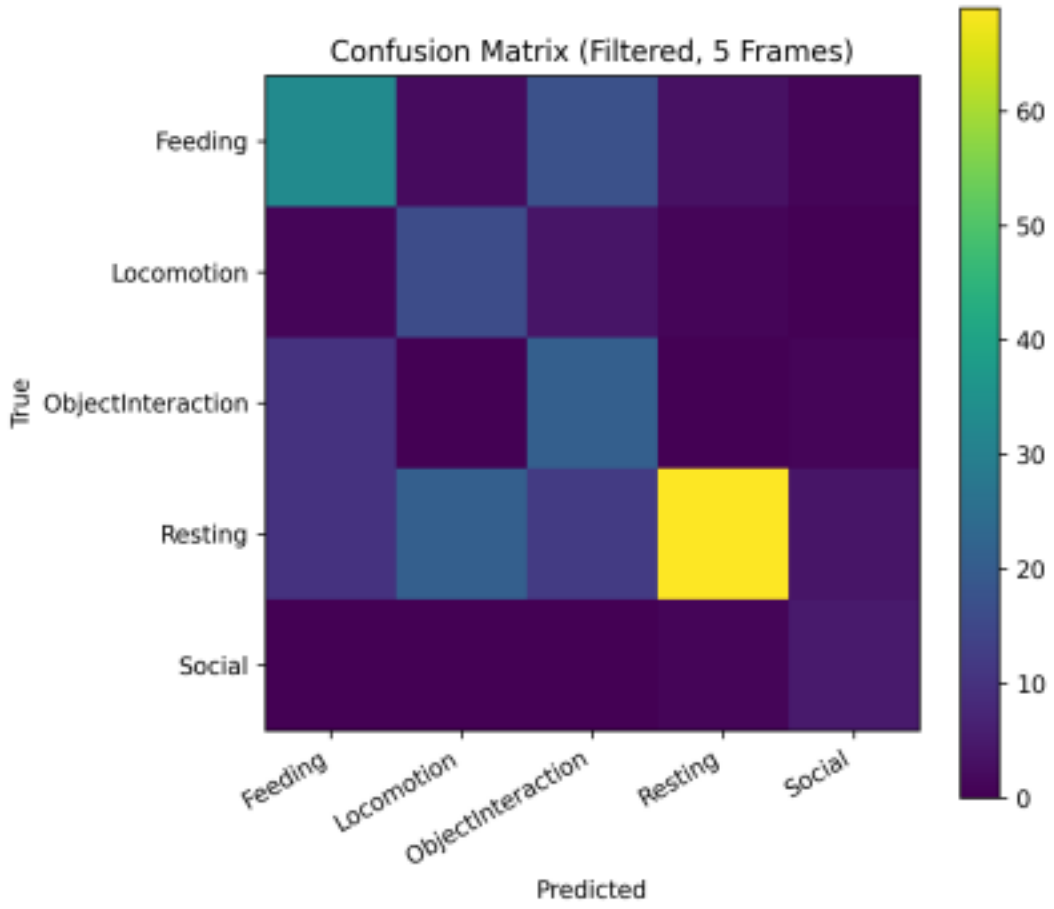

Figure 2. Confusion Matrix (Filtered Evaluation Set, 5 Frames)

8. Discussion

Taken together, these results suggest that multimodal LLMs may be dependable when tasked with detecting coarse, visually salient behavior, but less so for socially mediated or low-representation categories. Crucially, disagreement was interpretable and reliable rather than random.

The evaluation design—short clips, coverage-driven selection, and a coarse-grained taxonomy—strengthens interpretability of agreement metrics while deliberately avoiding claims about behavioral prevalence or welfare outcomes. Within this scope, LLMs appear well-suited as assistive tools for pre-labeling and prioritizing human review rather than as autonomous ethogram annotators.

9. Limitations

This study has several limitations:

- Clip selection is coverage-driven and does not support time-budget analysis.
- Some behavior categories, particularly social interaction, are underrepresented.
- Frame-based representations may miss fine temporal dynamics.
- Results may not generalize across species, camera viewpoints, or enclosure contexts.

10. Conclusion

We present a controlled and reproducible evaluation of LLM-generated ethograms relative to human annotations using publicly accessible zoo footage. While multimodal LLMs show promise for assistive annotation, they exhibit clear and systematic limitations. These findings support hybrid human–LLM workflows for scalable ethogram analysis, complementing rather than replacing human expertise.

Acknowledgements

The author thanks Shyam Agarwal (Carnegie Mellon University) for helpful discussions and feedback on the experimental design and interpretation of results.

References

Altmann, J. (1974). Observational study of behavior: sampling methods. *Behaviour*, 49(3–4), 227–267.
Martin, P., & Bateson, P. (2007). *Measuring Behaviour: An Introductory Guide*. Cambridge University Press.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
Mathis, A., et al. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*.
Anderson, D. J., et al. (2020). Machine learning for animal behavior analysis: a review. *Nature Methods*.