Zero-shot animal behavior classification with vision-language foundation models

Gaspard Dussert[1], Vincent Miele[1, 2], Colin Van Reeth[3], Anne Delestrade[3], Stéphane Dray[1] and Simon Chamaillé-Jammes[*4]

[1]Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive (LBBE), Villeurbanne, France

[2]Laboratoire d'Ecologie Alpine (LECA), Université Grenoble Alpes, CNRS, Université Savoie Mont Blanc, France

[3]Centre de Recherches sur les Ecosystèmes d'Altitude (CREA Mont-Blanc), Observatoire du Mont-Blanc, Chamonix, France

[4]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

July 7, 2024

Abstract

1. Understanding the behavior of animals in their natural habitats is critical to ecology and conservation. Camera traps are a powerful tool to collect such data with minimal disturbance. They however produce very a large quantity of images, which can make human-based annotation cumbersome or even impossible. While automated species identification with artificial intelligence has made impressive progress, automatic classification of animal behaviors in camera trap images remains a developing field.

[*]simon.chamaille@cefe.cnrs.fr

2. Here, we explore the potential of foundation models, specifically Vision Language Models (VLMs), to perform this task without the need to first train a model, which would require some level of human-based annotation. Using an original dataset of alpine fauna with behaviors annotated by participatory science, we investigate the zero-shot capabilities of different kind of recent VLMs to predict behaviors and estimate behavior-specific diel activity patterns in three ungulate species.

3. Our results show that using these methods, it is possible to achieve accuracies over 91% in behavior classification and produce activity patterns that closely align with those derived from participatory science data (overlap indexes between 84% and 90%).

4. These findings demonstrate the potential of foundation models and vision-language models in ecological research. Ecologists are encouraged to adopt these new methods and leverage their full capabilities to facilitate ecological studies.

Keywords : behavior classification, activity pattern, camera trap, machine learning, zero-shot, vision-language model, foundation model, downstream task

1 Introduction

Camera-traps are a powerful tool for studying wildlife in its natural habitats with minimal disturbance. As they are increasingly deployed throughout the world, a growing need for efficient methods to automatically analyze the vast amount of images they generate emerges. This is for instance true for one common goal of camera-trapping, which is to describe the diel activity pattern and time budget of some target species, i.e. the distribution of their level of activity and behaviors throughout the day (J. M. Rowcliffe et al. 2014; Gilbert et al. 2023).

Deriving activity patterns from camera-traps images first requires identifying the species present in the pictures (Whytock et al. 2021). This task can now often be fully automatized, as recent machine learning models achieve very high performance in species

classification (Rigoudy et al. 2023; Willi et al. 2019; Mitterwallner et al. 2023). Then, activity level is generally assumed to be proportional to the number of sightings. One simply counts the number of sightings of the species of interest over time intervals (e.g. hours) and build activity curves (Ridout and Linkie 2009), or make more complex modelling of the distribution of sightings (Gerber et al. 2024). However, one might want to study more informative, behavior-specific patterns, for instance to understand when are animals foraging, resting or moving. In contrast with species identification, classification of animals' behavior in camera-trap images or video is still in its infancy (Norouzzadeh et al. 2018; Liu et al. 2023; Schindler and Steinhage 2021; Schindler, Steinhage, et al. 2024).

In addition to identifying the best models for this task, several key constraints for developing such models exists: firstly, a large amount of training data is needed, but obtaining expert labels of behaviors on large sets of images is very time-consuming and costly. Behaviorally-annotated camera-trap datasets are currently very rare; secondly, a lot of computational resources are required for training. Such resources might not be available to ecologists, and training has a large carbon footprint; thirdly, the model need to be re-trained each time a new species or behavior class is added.

However, research groups in artificial intelligence worldwide are publishing "foundation models", capable of executing zero-shot tasks, i.e. capable of making predictions on tasks for which they have not been explicitly trained (Radford et al. 2021; Kirillov et al. 2023). These pre-trained foundation models can solve most of the above-mentioned problems, since they don't require training data, inference requires much less energy than training and the billions of examples seen during training theoretically allow them to predict and generalize on various species and behavior of interest (Fabian et al. 2023).

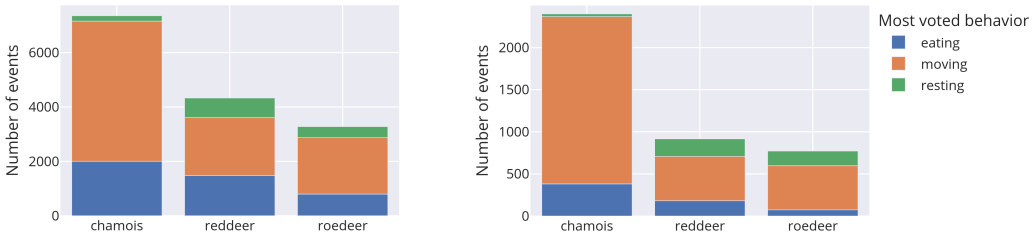
In this article, we investigate whether foundation models are already performant enough to be used in ecology for animals' behavioral classification, from camera-trap images, without any fine-tuning. We use an original dataset of camera trap images of

European fauna with behavior labels provided by participatory sciences. We explore the zero-shot capabilities of six recent Vision Language Models (VLM)(Bordes et al. 2024) to predict the behavior of three different species. As it is known that the formulation of the prompt request has a significant impact on the output of the models, we explore the importance of this effect for the proposed task. In particular, as we hypothesize that the models have not been trained on many images of wild animals active at night and may be biased towards certain behaviors. We therefore also investigate the effect of specifying to the model whether the image was taken during nighttime or daytime. In addition to accuracy measures, we use overlap measures to quantify to what extent diel activity patterns, constructed independently for each behavior and from human annotations, can be well estimated by relying only on models' predictions. Finally, we discuss how our study demonstrates the power of vision-language models that can immediately assist ecologists in many cumbersome tasks.

2 Material and Methods

2.1 Dataset

The CREA Mont-Blanc camera trap dataset is collected by CREA Mont-Blanc (Centre de Recherches sur les Ecosystèmes d'Altitude) in the valley of Chamonix in the French Alps. The dataset is made of 131 537 images collected between 2017 and 2020 during 24 580 events, events being defined as 1-minute periods. Images were labelled on https://www.zooniverse.org/ using participatory science. Participants were asked to identify the species and could also vote for the animal's behavior using five non-exclusive classes: *eating*, *resting*, *moving*, *standing* and *interacting*. Each event had been viewed by at least 30 people at the end of the process. The distribution of species and behaviors is shown in figure Figure 1a. We have focused our work on three species that are actively studied by the CREA Mont-Blanc: chamois (*Rupicapra rupicapra*), red deer (*Cervus*



(a) complete dataset (b) high agreement dataset

Figure 1: Species and behavior distribution of the CREA Mont-Blanc camera trap dataset for the three species of interest. (a) with all events, (b) with events for which the most voted behavior has at least 90% of the votes

elaphus) and roe deer (*Capreolus capreolus*), and focused on 3 behaviors: eating, moving and resting. We excluded standing from our analysis due to inconsistent interpretation by participants and its frequent combination with other behaviors. Additionally, we did not focus on interacting since there were only 71 events with this behavior in the entire dataset.

2.2 Zero-shot Vision-Language Models

Vision-Language Models (VLMs) are multi-modal deep learning models that can process both image and text simultaneously. These models, often trained on very large datasets, demonstrate very good zero-shot performances in downstream tasks. Here, we investigate the zero-shot capabilities of two categories of VLMs: contrastive vision-language models and multimodal Large Language Models (LLMs). A brief overview of these two categories is given below, for a comprehensive review, see Bordes et al. (2024).

Contrastive Vision-Language Models. These models are trained to predict the similarity between a text and an image. Formally, a vision encoder and a text encoder are used to extract embeddings of images and texts, and the similarity of an image-text pair is obtained through the cosine similarity of the normalized embeddings (see Figure 3a). During the training, a contrastive image-text objective is used to bring closer

Figure 2: Sample images from the dataset, of the three species and three behaviors of interest. When an event has multiple image, only the first one is shown. The Zooniverse label and proportion of vote is shown at the top in green, and the prediction of the best zero-shot model (CogVLM) with its confidence score is shown at the bottom in white.

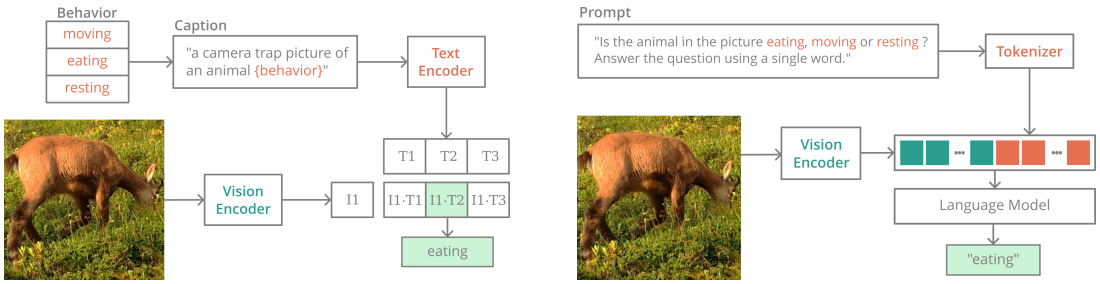
the embeddings of positive image-text pairs and push apart those of negative pairs. In our study, we decided to use CLIP (Radford et al. 2021) for our baseline as the first contrastive vision-language model of large scale, and SigLIP (Zhai et al. 2023) which improved on CLIP with a new contrastive objective. We also include WildCLIP(Gabeff et al. 2023), a fine-tuned version of CLIP trained exclusively on camera trap images of African fauna. WildCLIP uses the large Snapshot Serengeti dataset (Swanson et al. 2015), with image captions built from attributes labeled through participatory science, including animal behavior.

Multimodal Large Language Models. These models process both image and text tokens at the input of a Large-Language Model (LLM). The image tokens are obtained using a vision encoder, typically a pretrained contrastive model such as CLIP. The integration of a language model allows these models to perform more complex reasoning compared to contrastive models. We examine three multimodal LLMs of varying sizes: CogVLM (Wang et al. 2024), which has a very large vision encoder (EVA2-CLIP-E); MobileVLM V2 (Chu et al. 2024), a lightweight VLM which uses CLIP ViT-L/14; and the more recent PaliGemma (Beyer et al. 2024), which utilizes SigLIP-SO.

2.3 Zero-shot inference

For all methods, images are first cropped using MegaDetector v5 (Beery et al. 2019) to ensure that the image encoder of each model extracts information related to the animal and not to the rest of the scene.

Contrastive vision-language models such as CLIP are able to perform zero-shot classification (Radford et al. 2021) using their ability to predict the similarity score between an image and a text. Here, we follow this methodology and match the images with the following captions: "a camera trap picture of an animal {behavior}". This process is illustrated on Figure 3a. We calculate similarity scores for each behavior, normalize them using the softmax function, and select the behavior with the highest score as the



(a) Contrastive VLM (b) Multimodal LLM

Figure 3: Overview of the zero-shot inference process for the two model families we tested. Contrastive vision-language model (a) predict the similarity between the image and the three captions, one for each behavior. The predicted behavior is the one with the highest similarity score (light blue). Multimodal LLMs (b) are language models that can process visual tokens in addition to the prompt, and predict the behavior as the next word in the sequence.

final prediction. We use the normalized similarity score of the chosen behavior as the confidence score.

For the multimodal LLMs, we perform a Visual Question Answering task (Agrawal et al. 2016). As shown on Figure 3b, we prompt the assistant with the following question: "Is the animal in the image eating, moving or resting? Answer the question using a single word.". For each word prediction in the answer, the model considers all possible words in the vocabulary, assigns a score to each one and predicts the one with the highest score. In our case, since the model predicts directly the behavior with the first word, we can use it to retrieve the score of each behavior. Then, following the methodology described by Jiang et al. (2021), the score of the predicted behavior is normalized by the sum of the scores of the candidates words only (the three behaviors words).

We also experiment to give additional context to the models to improve their performances: we inform the model whether the image was taken during daytime or nighttime. For CLIP-based methods, we change the caption to "a camera trap picture of an animal {behavior}, captured during {context}." with *context* being either "nighttime" or "daytime". For LLMs, we add "This image is captured during {context}." before the

question.

For each behavior, the classification performance of the model are evaluated with the accuracy and the calibration of the scores is measured with the ECE (Guo et al. 2017). For these two metrics, we use a subset of the dataset that removes ambiguity in the labels (e.g. sequences with an animal that is seen both eating and moving). This subset is obtained by keeping only the images for which the most voted for behavior has at least 90% of the votes, and will be called the *high agreement* dataset hereafter. To address the imbalance between behaviors (Figure 1b), we report the macro accuracy in addition to the accuracy of each behavior. This metric is calculated by taking the average of the accuracy of the three behaviors, each given equal weight, ensuring that no single behavior disproportionately influences the overall accuracy. We also aggregate individual images predictions to event predictions to improve the accuracy of the models (Dussert et al. 2024). Confidence scores at the sequence level are calculated by averaging the scores of the images in the sequence.

2.4 Behavior-specific activity patterns

Diel activity patterns are computed either on the Zooniverse annotations or on the models predictions using the time at which the camera trap events were triggered. As proposed by Vazquez et al. (2019), times are first converted to radian and standardized using two anchor time (the average sunrise time and the average sunset time) to account for the variation of the length of the day during the year. A circular kernel density (of the von Mises distribution) is fitted to the radian times to obtain the activity pattern. The overlap index Δ is used to compare two activity patterns by computing the integral of the minimum between the two density function (Ridout and Linkie 2009). An index of 1 is obtained for a perfect overlap and 0 for no overlap at all. In practice, it is computed with the estimator $\hat{\Delta}_4$ of Ridout and Linkie (2009). These operations were done with the functions *solartime*, *fitact* and *compareCkern* of the ACTIVITY package in R (M. Rowcliffe

2023), with 200 bootstrap replicates for the estimation of $\hat{\Delta}_4$.

Activity patterns are computed using the complete dataset for our three species and three behaviors of interest. For Zooniverse annotations, we keep only the events for which a behavior has at least 50% of the votes. For models prediction, we used the same threshold of 0.75 on the confidence scores of all the models to reduce false positives. We did not apply any filter to ensure independence between event (e.g. 30 minutes interval between two events) as it has been shown to bias the activity patterns (Peral et al. 2022).

3 Results

Starting with the contrastive vision-language models, the baseline CLIP with ViT-Base architecture performs poorly for this behavior recognition task, with a low accuracy of 71.75% (Table 1). It particularly fails at correctly predicting the eating behavior. The SigLIP model represents a significant improvement over the CLIP model, with a much better accuracy of 91.14%. The accuracy of the eating behavior is however still lower than the one for the other behaviors (Table 1). The specialized WildCLIP model offers very similar performance to SigLIP despite a much smaller number of parameters. For the multimodal LLMs, CogVLM gives the best performance, with an accuracy of 97.45%. PaliGemma which uses SigLIP as vision encoder is only slightly better than SigLIP but performs better than MobileVLM V2 despite having less parameters. Per-species accuracies are shown in Supporting Information Table 1.

Giving additional day-night context improves the accuracy of all the contrastive models (Table 1), but decreases the accuracy of the LLMs. None of the models exhibited a good score calibration, with the best model (WildCLIP) having an ECE of 6.85%. Moreover, adding context slightly deteriorates the calibration of almost all the models.

As expected from the heterogeneity in the models' accuracy, the quality of the

Model	Context (day & night)	Accuracy			Macro accuracy	ECE (%)	#Params
		Eating	Moving	Resting			
Contrastive models							
CLIP (ViT-B/16)	×	31.76	92.39	91.08	71.75	17.15	150M
	✓	48.11	88.67	93.98	76.92	19.61	
SigLIP (ViT-SO400M)	×	84.28	94.20	94.94	91.14	10.78	878M
	✓	94.65	89.20	95.18	93.01	14.15	
WildCLIP (ViT-B/16)	×	89.47	**94.70**	93.49	92.55	**6.85**	150M
	✓	90.09	94.17	95.42	93.23	7.80	
Multimodal LLMs							
CogVLM	×	99.37	93.71	**99.28**	**97.45**	11.93	18B
	✓	**99.53**	87.95	99.04	95.50	11.29	
MobileVLM V2	×	95.75	82.23	**99.28**	92.44	9.55	7B
	✓	96.07	78.89	**99.28**	91.41	10.35	
PaliGemma	×	90.57	91.80	96.87	93.08	7.46	3B
	✓	92.30	91.41	95.42	93.04	9.38	

Table 1: Behavioral classification accuracy and expected calibration error (ECE) of the six studied models on the high agreement dataset.

behavior-specific diel activity patterns estimated from their predictions varies. Figure 4 presents the activity patterns estimated using CLIP, SigLIP and CogVLM models without day/night context. These models were chosen to provide a comprehensive comparison across different classification accuracy levels. For other models, see Supporting Information Figure 1-6 and Table 2. Activity patterns computed with CogVLM predictions are very close to those obtained with participatory science labels, and even small variations in activity along the course of the day are captured. The activity patterns predicted by CLIP are generally unreliable. In particular, the model is biased towards predicting moving at nighttime. To a lesser extent, SigLIP also presents a night bias, which is particularly visible using the red deer data (Figure 4). This bias seems to be considerably reduced by the addition of the day/night context (Supporting Information Table 2).

Quantitative measures of the overlap in activity patterns between the one derived from Zooniverse annotations and those derived from the various models are shown in Table 2. In accordance with previous results, CogVLM has the best overall overlap values, with an average overlap index of 90.04%. SigLIP, WildCLIP, PaliGemma and MobileVLM V2 exhibit slightly lower performance compared to CogVLM, while CLIP performs significantly worse. Once again, the additional day-night context helps with

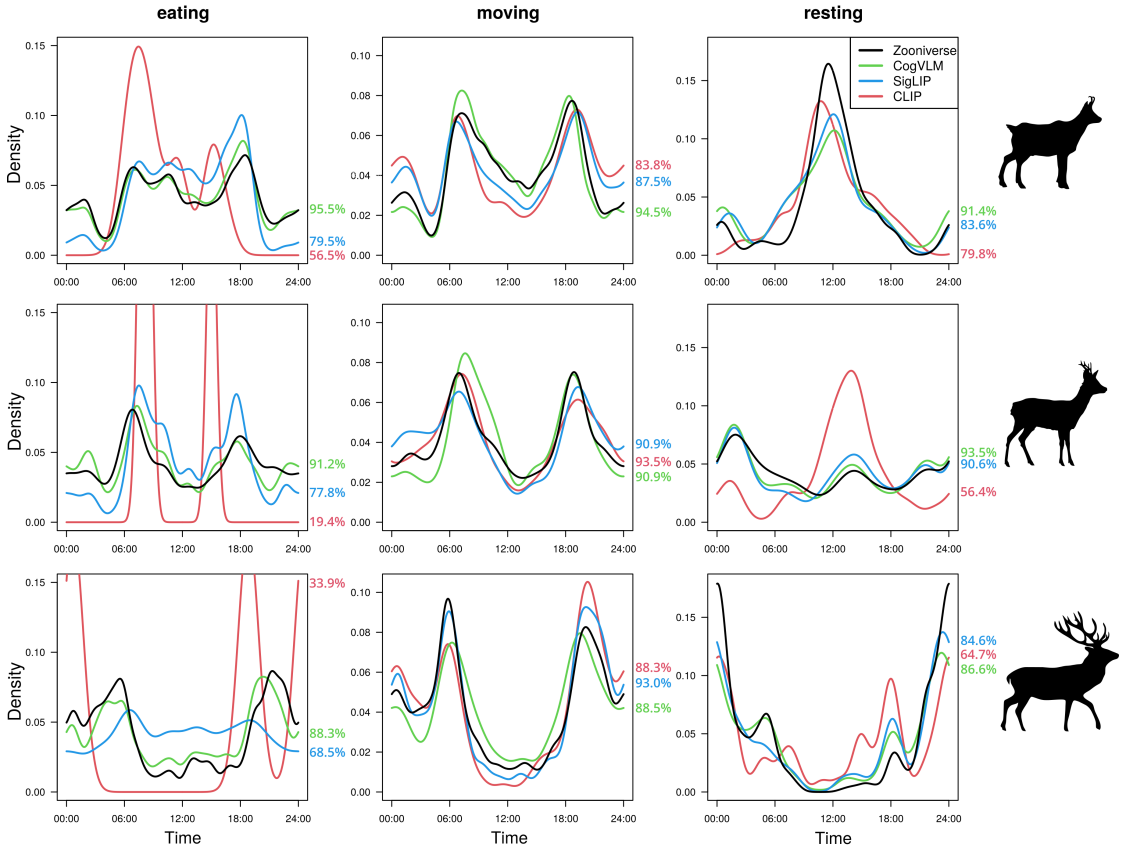


Figure 4: Comparison of the activity patterns using either the participatory science labels from Zooniverse (black) or predictions of three models (colors). Each row corresponds to a different species (top to bottom : chamois, roe deer, red deer). Overlap values are indicated on the right of the curves.

contrastive models, but not with LLMs. In particular, SigLIP with context emerges as the second-best model.

4 Discussion

This study evaluates the ability of zero-shot modern VLMs to predict the behavior of three mountain ungulates in camera trap images. We show that these models can have very good performances, with macro accuracy exceeding 91% for all models except the baseline (CLIP ViT-B). Furthermore, the diel activity patterns generated by these models

Model	Context	Average overlap index			
		Eating	Moving	Resting	All
Contrastive models					
CLIP (ViT-B/16)	×	36.60	66.97	88.53	64.03
	✓	39.73	75.80	91.27	68.93
SigLIP (ViT-SO400M)	×	75.27	86.27	90.47	84.00
	✓	89.97	85.13	**93.73**	89.61
WildCLIP (ViT-B/16)	×	91.03	80.63	92.27	87.98
	✓	91.20	80.23	91.00	87.48
Multimodal LLMs					
CogVLM	×	91.67	87.17	91.30	**90.04**
	✓	93.00	87.37	85.27	88.54
MobileVLM V2	×	**93.43**	83.37	87.67	88.16
	✓	93.10	86.86	83.63	87.86
PaliGemma	×	88.17	**90.47**	82.83	87.15
	✓	89.00	89.13	83.30	87.14

Table 2: Overlap index averaged over the three species of interest on the complete dataset.

closely align with those obtained through participatory science, exhibiting overlap indexes between 84% and 90%, again with the exception of CLIP.

Using an automated and flexible method, such as the one presented in this study, to classify animal behaviors on pictures can save considerable time compared to having images annotated by citizen scientists or experts. For instance, the very good SigLIP model can process tens of thousands of pictures in an hour (see below for further discussion). Therefore, this approach could enable large-scale annotation of camera trap datasets and facilitate or automate the study of behaviors, which could for example be particularly useful for studying the effects of climate change or human activities on the behavior of wild animals. These research topics are increasingly studied using camera-traps (Bison et al. 2024; Blount et al. 2024). Furthermore, even though this study illustrates the effectiveness of foundation vision-language models for a specific task, the flexibility of the method makes it easily adaptable to other ecological tasks. For example, it could be possible to predict human behaviors or landscapes attributes by changing the prompt (or the caption) given to the model.

While some existing datasets such as MammalNet (Chen et al. 2023) and AnimalK-

ingdom (Ng et al. 2022) are available and have more species and behaviors, we decided not to use them as they consist of very high quality video footage. Good model performance on these datasets may not transfer to camera trap datasets given the distribution gap between the two. In addition, we decided not to focus on video processing, as many camera trap studies use single shot or burst mode, requiring the development of methods to obtain the most accurate behavior possible from one or a few images (Rovero and Zimmermann 2016).

A visual analysis of discrepancies between model predictions and participatory science annotations reveals a degree of inherent ambiguity and that both may be correct in certain instances. In fact, some sequences of pictures clearly contains images in which the behavior of the animal differ (e.g. Figure 2, bottom right), with eating and moving being the most frequent combination. While not common, these ambiguities may explain the slight deviation in the diel activity curves obtained from CogVLM, despite its near-perfect accuracy on the high agreement dataset. This suggests that exploring a zero-shot multilabel classification approach might be useful. Furthermore, the models occasionally make errors attributable to a poor scene understanding: an animal walking in the foreground might be misclassified as "eating" if its head aligns with background vegetation, or misclassified as "resting" if its legs are hidden by deep snow or rocks. These errors could potentially be eliminated using VLMs capable of simultaneously processing multiple images from the same sequence, rather than independently.

Given that several of the models achieve very good performance, it is interesting to study the trade-off between model size and performance. When considering multimodal LLMs, CogVLM appears to be the best model if inference time is not a concern while MobileVLM V2 is clearly not interesting, as its performance is similar to the one of WildCLIP, which is 50 times smaller. The difference in performance is highly likely due to the differing sizes of the vision encoders between CogVLM (10B parameters) and MobileVLM (428M parameters). For contrastive models, WildCLIP seems to be

the most efficient model. It performs considerably better than CLIP ViT-Base, despite having the same number of parameters. WildCLIP's performance is impressive given that it has been fine-tuned on African data only, which demonstrates the effectiveness of the VR-LwF "Learning without Forgetting" loss that was used. However, the model is not performing better than SigLIP despite having already been exposed to camera trap images and similar captions depicting the same behaviors. This may be attributed to WildCLIP having much fewer parameters and the distribution shift between the training data and our dataset. Nonetheless, the release of our data will facilitate the development of specialized models more robust to the changes in species, habitats and behaviors. Finally, SigLIP seems to be the most efficient non-specialized foundation model and has a reasonable inference time (1 hour for 40 000 images on one NVIDIA V100 GPU), which could also make it useful for analyses conducted on standard desktop CPU-only computers.

Providing the context of whether the image was taken during the day or at night only seems to improve the contrastive models. This can be explained by the fact that giving a more precise caption reduces the distribution gap with the captions in the training dataset (Radford et al. 2021). Current multimodal LLMs, on the other hand, are trained with rigid conversation templates and can be very sensitive to the structure of the prompt. Most of the datasets they were trained on do not include context before the question. Hence, in contrast to the contrastive models, giving this additional context has increased the distribution gap and may have made confused the models. Moreover, current LLMs are not invariant to permutations in the order of choices presented in the prompt (Zong et al. 2023). In other words, the sequence in which options are listed can impact the prediction of the model. For example, given choices A, B, and C, the model may output different predictions if the options are presented as A-B-C or C-A-B. This is for instance the case in our study (see Supporting Information Table 3), although it does not alter our general conclusions. LLMs may also exhibit positional bias, being more likely to predict

one choice based on its position rather than its content (Zheng et al. 2024). Therefore, while multimodal LLMs demonstrate impressive performance, it is crucial to acknowledge and account for these limitations in their application and interpretation of results.

Overall, our study raises questions regarding the necessity of training new models for ecological applications, especially considering the rapidly accelerating rate of publication of foundation models. We can expect significant improvements and capabilities in the near future, such as the ability to analyze multiple consecutive images, as the commercial GPT4-Vision model already does (OpenAI 2023). Ecologists should embrace these new methods, evaluate their performance, and learn how to exploit them to their full potential.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD010113729 made by GENCI.

Conflict of interest

None of the authors has a conflict of interest.

Author contributions

G.D., S.C.J., S.D. and V.M. conceived the ideas and designed the methodology. A.D. and C.V.R gathered, annotated and processed the CREA-Mont Blanc dataset. G.D. coded and performed the analysis. G.D. wrote the first version of the manuscript, S.C.J., S.D., A.D and C.V.R and V.M. contributed critically to the drafts and gave final approval for publication.

Data availability statements

The code and dataset will be made available on Zenodo:
https://doi.org/10.5281/zenodo.10925926

References

Agrawal, Aishwarya, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh (Oct. 2016). *VQA: Visual Question Answering*. arXiv:1505.00468 [cs]. URL: http://arxiv.org/abs/1505.00468.

Beery, Sara, Dan Morris, and Siyu Yang (July 2019). *Efficient Pipeline for Camera Trap Image Review*. arXiv:1907.06772 [cs]. DOI: 10.48550/arXiv.1907.06772. URL: http://arxiv.org/abs/1907.06772.

Beyer, Lucas, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Xiaohua Zhai, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, and et al. (2024). "PaliGemma". In: DOI: 10.34740/KAGGLE/M/23393. URL: https://www.kaggle.com/m/23393.

Bison, Marjorie, Nigel G. Yoccoz, Bradley Z. Carlson, Arthur Bayle, and Anne Delestrade (2024). "Camera traps reveal seasonal variation in activity and occupancy of the Alpine mountain hare Lepus timidus varronis". en. In: *Wildlife Biology* 2024.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wlb3.01186, e01186. ISSN: 1903-220X. DOI: 10.1002/wlb3.01186. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wlb3.01186.

Blount, J. David, Austin M. Green, Mark Chynoweth, Kyle D. Kittelberger, Dário Hipólito, Katarzyna Bojarska, Emrah Çoban, Josip Kusak, and Çağan H. Şekercioğlu (2024). "Seasonal activity patterns and home range sizes of wolves in the human-dominated landscape of northeast Türkiye". en. In: *Wildlife Biology* n/a.n/a. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wlb3.01257, e01257. ISSN:

1903-220X. DOI: 10.1002/wlb3.01257. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wlb3.01257.

Bordes, Florian, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra (2024). *An Introduction to Vision-Language Modeling*. arXiv: 2405.17247.

Chen, Jun, Ming Hu, Darren J. Coker, Michael L. Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny (June 2023). *MammalNet: A Large-scale Video Benchmark for Mammal Recognition and Behavior Understanding*. arXiv:2306.00576 [cs]. DOI: 10.48550/arXiv.2306.00576. URL: http://arxiv.org/abs/2306.00576.

Chu, Xiangxiang, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen (Feb. 2024). *MobileVLM V2: Faster and Stronger Baseline for Vision Language Model*. arXiv:2402.03766 [cs]. URL: http://arxiv.org/abs/2402.03766.

Dussert, Gaspard, Simon Chamaillé-Jammes, Stéphane Dray, and Vincent Miele (Feb. 2024). *Being confident in confidence scores: calibration in deep learning models for camera trap image sequences*. en. Pages: 2023.11.10.566512 Section: New Results. DOI: 10.1101/2023.11.10.566512. URL: https://www.biorxiv.org/content/10.1101/2023.11.10.566512v2.

Fabian, Zalan, Zhongqi Miao, Chunyuan Li, Yuanhan Zhang, Ziwei Liu, Andrés Hernández, Andrés Montes-Rojas, Rafael Escucha, Laura Siabatto, Andrés Link, Pablo Arbeláez, Rahul Dodhia, and Juan Lavista Ferres (Nov. 2023). *Multimodal Foun-*

dation Models for Zero-shot Animal Species Recognition in Camera Trap Images. arXiv:2311.01064 [cs]. DOI: 10.48550/arXiv.2311.01064. URL: http://arxiv.org/abs/2311.01064.

Gabeff, Valentin, Marc Rußwurm, Devis Tuia, and Alexander Mathis (Dec. 2023). *Wild-CLIP: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models*. en. Pages: 2023.12.22.572990 Section: New Results. DOI: 10.1101/2023.12.22.572990. URL: https://www.biorxiv.org/content/10.1101/2023.12.22.572990v1.

Gerber, Brian D., Kadambari Devarajan, Zach J. Farris, and Mason Fidino (2024). "A model-based hypothesis framework to define and estimate the diel niche via the 'Diel.Niche' R package". In: *Journal of Animal Ecology* 93.2, pp. 132–146. DOI: https://doi.org/10.1111/1365-2656.14035. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.14035. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2656.14035.

Gilbert, Neil A., Kate A. McGinn, Laura A. Nunes, Amy A. Shipley, Jacy Bernath-Plaisted, John D. J. Clare, Penelope W. Murphy, Spencer R. Keyser, Kimberly L. Thompson, Scott B. Maresh Nelson, Jeremy M. Cohen, Ivy V. Widick, Savannah L. Bartel, John L. Orrock, and Benjamin Zuckerberg (Apr. 2023). "Daily Activity Timing in the Anthropocene". In: *Trends in Ecology & Evolution* 38.4, pp. 324–336. ISSN: 0169-5347. DOI: 10.1016/j.tree.2022.10.008.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger (Aug. 2017). *On Calibration of Modern Neural Networks*. arXiv:1706.04599 [cs]. URL: http://arxiv.org/abs/1706.04599.

Jiang, Zhengbao, Jun Araki, Haibo Ding, and Graham Neubig (May 2021). *How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering*. arXiv:2012.00955 [cs]. URL: http://arxiv.org/abs/2012.00955.

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick (Apr. 2023). *Segment Anything*. arXiv:2304.02643 [cs]. URL: http://arxiv.org/abs/2304.02643.

Liu, Dan, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jingdong Zhang (2023). "LoTE-Animal: A Long Time-span Dataset for Endangered Animal Behavior Understanding". In.

Mitterwallner, Veronika, Anne Peters, Hendrik Edelhoff, Gregor Mathes, Hien Nguyen, Wibke Peters, Marco Heurich, and Manuel J. Steinbauer (2023). "Automated visitor and wildlife monitoring with camera traps and machine learning". In: *Remote Sensing in Ecology and Conservation* n/a.n/a. DOI: https://doi.org/10.1002/rse2.367. eprint: https://zslpublications.onlinelibrary.wiley.com/doi/pdf/10.1002/rse2.367. URL: https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1002/rse2.367.

Ng, Xun Long, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu (June 2022). "Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding". en. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, pp. 19001–19012. ISBN: 978-1-66546-946-3. DOI: 10.1109/CVPR52688.2022.01844. URL: https://ieeexplore.ieee.org/document/9879959/.

Norouzzadeh, Mohammad Sadegh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune (June 2018). "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning". en. In: *Proceedings of the National Academy of Sciences* 115.25. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1719367115. URL: https://pnas.org/doi/full/10.1073/pnas.1719367115.

OpenAI (2023). “GPT-4 Vision”. In: URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Peral, Christopher, Marietjie Landman, and Graham I. H. Kerley (2022). “The inappropriate use of time-to-independence biases estimates of activity patterns of free-ranging mammals derived from camera traps”. en. In: *Ecology and Evolution* 12.10. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.9408, e9408. ISSN: 2045-7758. DOI: 10.1002/ece3.9408. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.9408.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (Feb. 2021). *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:2103.00020 [cs]. DOI: 10.48550/arXiv.2103.00020. URL: http://arxiv.org/abs/2103.00020.

Ridout, M. S. and M. Linkie (Sept. 2009). “Estimating overlap of daily activity patterns from camera trap data”. en. In: *Journal of Agricultural, Biological, and Environmental Statistics* 14.3, pp. 322–337. ISSN: 1537-2693. DOI: 10.1198/jabes.2009.08038. URL: https://doi.org/10.1198/jabes.2009.08038.

Rigoudy, Noa, Gaspard Dussert, Abdelbaki Benyoub, Aurélien Besnard, Carole Birck, Jérome Boyer, Yoann Bollet, Yoann Bunz, Gérard Caussimont, Elias Chetouane, Jules Chiffard Carriburu, Pierre Cornette, Anne Delestrade, Nina De Backer, Lucie Dispan, Maden Le Barh, Jeanne Duhayer, Jean-François Elder, Jean-Baptiste Fanjul, Jocelyn Fonderflick, Nicolas Froustey, Mathieu Garel, William Gaudry, Agathe Gérard, Olivier Gimenez, Arzhela Hemery, Audrey Hemon, Jean-Michel Jullien, Daniel Knitter, Isabelle Malafosse, Mircea Marginean, Louise Ménard, Alice Ouvrier, Gwennaelle Pariset, Vincent Prunet, Julien Rabault, Malory Randon, Yann Raulet, Antoine Régnier, Romain Ribière, Jean-Claude Ricci, Sandrine Ruette, Yann Schneylin, Jérôme Sentilles, Nathalie Siefert, Bethany Smith, Guillaume Terpereau, Pierrick Touchet,

Wilfried Thuiller, Antonio Uzal, Valentin Vautrain, Ruppert Vimal, Julian Weber, Bruno Spataro, Vincent Miele, and Simon Chamaillé-Jammes (Oct. 2023). "The DeepFaune initiative: a collaborative effort towards the automatic identification of European fauna in camera trap images". en. In: *European Journal of Wildlife Research* 69.6, p. 113. ISSN: 1439-0574. DOI: 10.1007/s10344-023-01742-7. URL: https://doi.org/10.1007/s10344-023-01742-7.

Rovero, Francesco and Fridolin Zimmermann (June 2016). *Camera Trapping for Wildlife Research*. Pelagic Publishing. ISBN: 978-1-78427-048-3.

Rowcliffe, J. Marcus, Roland Kays, Bart Kranstauber, Chris Carbone, and Patrick A. Jansen (2014). "Quantifying levels of animal activity using camera trap data". en. In: *Methods in Ecology and Evolution* 5.11. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12278, pp. 1170–1179. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12278. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12278.

Rowcliffe, Marcus (2023). *activity: Animal Activity Statistics*. R package version 1.3.4. URL: https://CRAN.R-project.org/package=activity.

Schindler, Frank and Volker Steinhage (Mar. 2021). "Identification of animals and recognition of their actions in wildlife videos using deep learning techniques". en. In: *Ecological Informatics* 61, p. 101215. ISSN: 15749541. DOI: 10.1016/j.ecoinf.2021.101215. URL: https://linkinghub.elsevier.com/retrieve/pii/S1574954121000066.

Schindler, Frank, Volker Steinhage, Suzanne T. S. van Beeck Calkoen, and Marco Heurich (Jan. 2024). "Action Detection for Wildlife Monitoring with Camera Traps Based on Segmentation with Filtering of Tracklets (SWIFT) and Mask-Guided Action Recognition (MAROON)". In: *Applied Sciences* 14.2, p. 514. ISSN: 2076-3417. DOI: 10.3390/app14020514.

Swanson, Alexandra, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer (June 2015). "Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna". en. In: *Scientific Data*

2.1. Publisher: Nature Publishing Group, p. 150026. ISSN: 2052-4463. DOI: 10.1038/sdata.2015.26. URL: https://www.nature.com/articles/sdata201526.

Vazquez, Carmen, J. Marcus Rowcliffe, Kamiel Spoelstra, and Patrick A. Jansen (2019). "Comparing Diel Activity Patterns of Wildlife across Latitudes and Seasons: Time Transformations Using Day Length". In: *Methods in Ecology and Evolution* 10.12, pp. 2057–2066. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13290.

Wang, Weihan, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang (Feb. 2024). *CogVLM: Visual Expert for Pretrained Language Models*. arXiv:2311.03079 [cs]. URL: http://arxiv.org/abs/2311.03079.

Whytock, Robin C., Jędrzej Świeżewski, Joeri A. Zwerts, Tadeusz Bara-Słupski, Aurélie Flore Koumba Pambo, Marek Rogala, Laila Bahaa-el-din, Kelly Boekee, Stephanie Brittain, Anabelle W. Cardoso, Philipp Henschel, David Lehmann, Brice Momboua, Cisquet Kiebou Opepa, Christopher Orbell, Ross T. Pitman, Hugh S. Robinson, and Katharine A. Abernethy (2021). "Robust ecological analysis of camera trap data labelled by a machine learning model". en. In: *Methods in Ecology and Evolution* 12.6. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13576, pp. 1080–1092. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13576. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13576.

Willi, Marco, Ross T. Pitman, Anabelle W. Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, and Lucy Fortson (2019). "Identifying animal species in camera trap images using deep learning and citizen science". en. In: *Methods in Ecology and Evolution* 10.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13099, pp. 80–91. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13099. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13099.

Zhai, Xiaohua, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer (Sept. 2023). *Sigmoid Loss for Language Image Pre-Training*. arXiv:2303.15343 [cs]. URL: http://arxiv.org/abs/2303.15343.

Zheng, Chujie, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang (Feb. 2024). *Large Language Models Are Not Robust Multiple Choice Selectors*. arXiv:2309.03882 [cs]. DOI: 10.48550/arXiv.2309.03882. URL: http://arxiv.org/abs/2309.03882.

Zong, Yongshuo, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales (Oct. 2023). *Fool Your (Vision and) Language Model With Embarrassingly Simple Permutations*. en. arXiv:2310.01651 [cs]. URL: http://arxiv.org/abs/2310.01651.

Supporting information

Model	Context	Chamois			Roe deer			Red deer		
		Eating	Moving	Resting	Eating	Moving	Resting	Eating	Moving	Resting
Contrastive models										
CLIP	×	91.65	36.48	93.75	93.51	26.03	89.14	94.11	24.18	92.31
	✓	87.72	54.59	87.50	90.65	43.84	93.71	90.30	36.26	95.19
SigLIP	×	95.22	85.04	96.88	91.60	89.04	94.86	92.97	80.77	94.71
	✓	92.65	94.49	93.75	80.53	97.26	95.43	84.79	93.96	95.19
WildCLIP	×	96.48	88.19	78.12	90.27	87.67	93.71	92.40	92.86	95.67
	✓	95.87	89.24	81.25	90.84	87.67	96.00	91.06	92.86	97.12
Multimodal LLMs										
CogVLM	×	95.87	99.21	100.0	91.03	100.0	99.43	88.21	99.45	99.04
	✓	92.35	99.21	100.0	80.34	100.0	98.86	78.90	100.0	99.04
MobileVLM V2	×	85.41	94.75	100.0	77.48	100.0	99.43	75.29	96.15	99.04
	✓	82.64	94.75	100.0	72.71	100.0	99.43	70.91	97.25	99.04
PaliGemma	×	92.50	90.29	96.88	90.84	90.41	94.86	90.11	91.21	98.56
	✓	92.60	91.86	96.88	88.93	93.15	93.14	89.35	92.86	97.12

Table 3: Behavior classification accuracy of the six models, for each species and behavior, using the high agreement dataset

Model	Context	Chamois			Roe deer			Red deer		
		Eating	Moving	Resting	Eating	Moving	Resting	Eating	Moving	Resting
Contrastive models										
CLIP	×	56.5	83.8	79.8	19.4	93.5	56.4	33.9	88.3	64.7
	✓	65.8	90.4	77.2	18.2	92.4	72.7	35.2	91.0	77.5
SigLIP	×	79.5	87.5	83.6	77.8	90.9	90.6	68.5	93.0	84.6
	✓	92.0	96.8	83.4	92.2	92.6	88.1	85.7	91.8	83.9
WildCLIP	×	93.0	97.0	71.8	88.9	90.7	85.6	91.2	89.1	84.5
	✓	93.3	97.3	68.7	89.9	88.3	86.6	90.4	87.4	85.4
Multimodal LLMs										
CogVLM	×	95.5	94.5	81.4	91.2	90.9	93.5	88.3	88.5	86.6
	✓	96.5	89.3	81.6	92.9	83.2	93.9	89.6	83.3	86.6
MobileVLM V2	×	95.9	94.4	73.8	91.6	84.6	92.4	92.8	84.0	83.9
	✓	95.9	92.7	74.6	91.0	82.3	92.8	92.4	85.5	83.5
PaliGemma	×	89.9	94.1	71.8	90.5	89.4	88.9	84.1	87.9	87.8
	✓	90.4	93.7	75.1	91.6	87.0	87.8	85.0	86.7	87.0

Table 4: Overlap index of the six models, for each species and behavior, using the complete dataset.

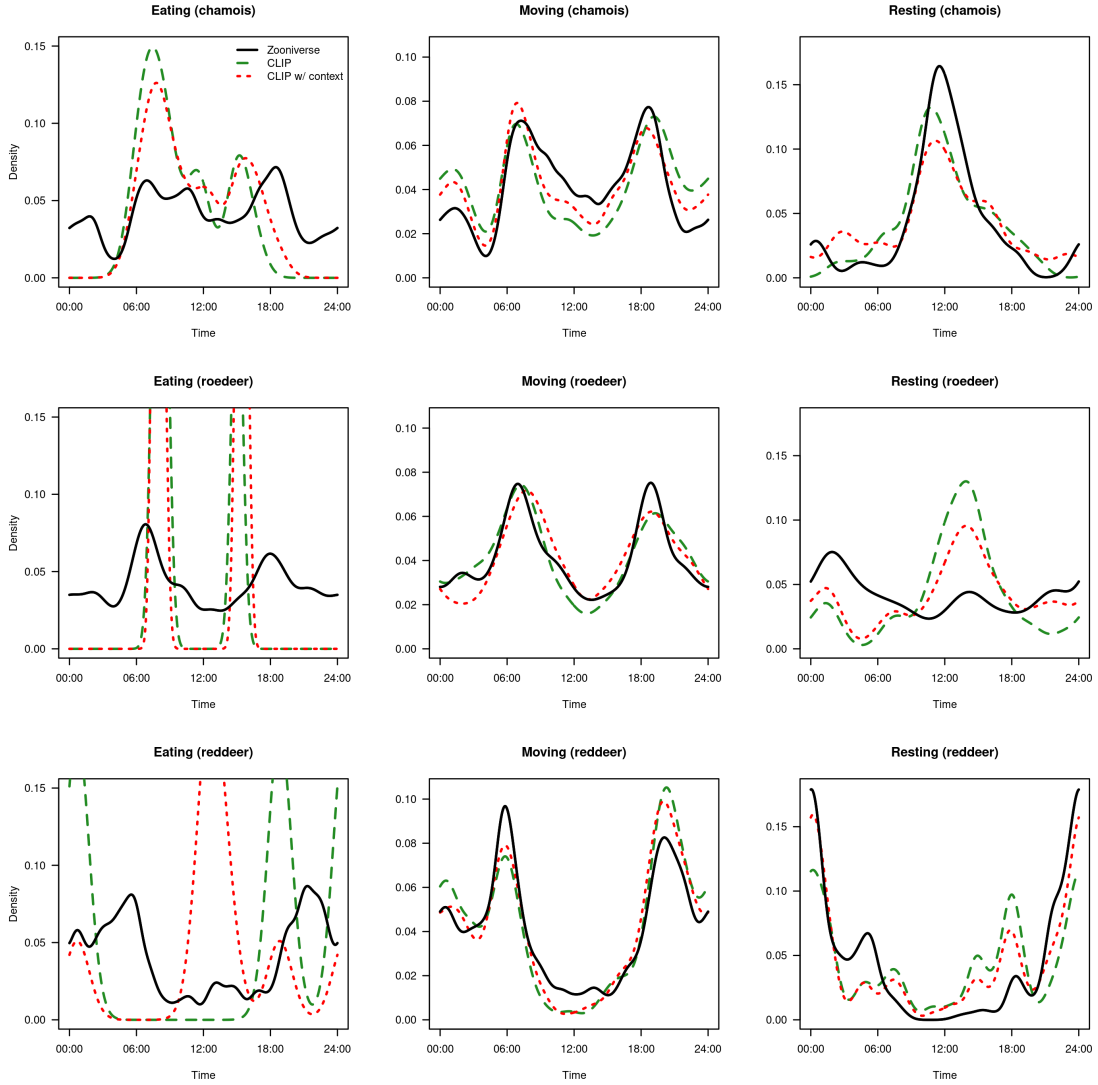


Figure 5: Activity patterns of CLIP, with and without day-night context, using the complete dataset.

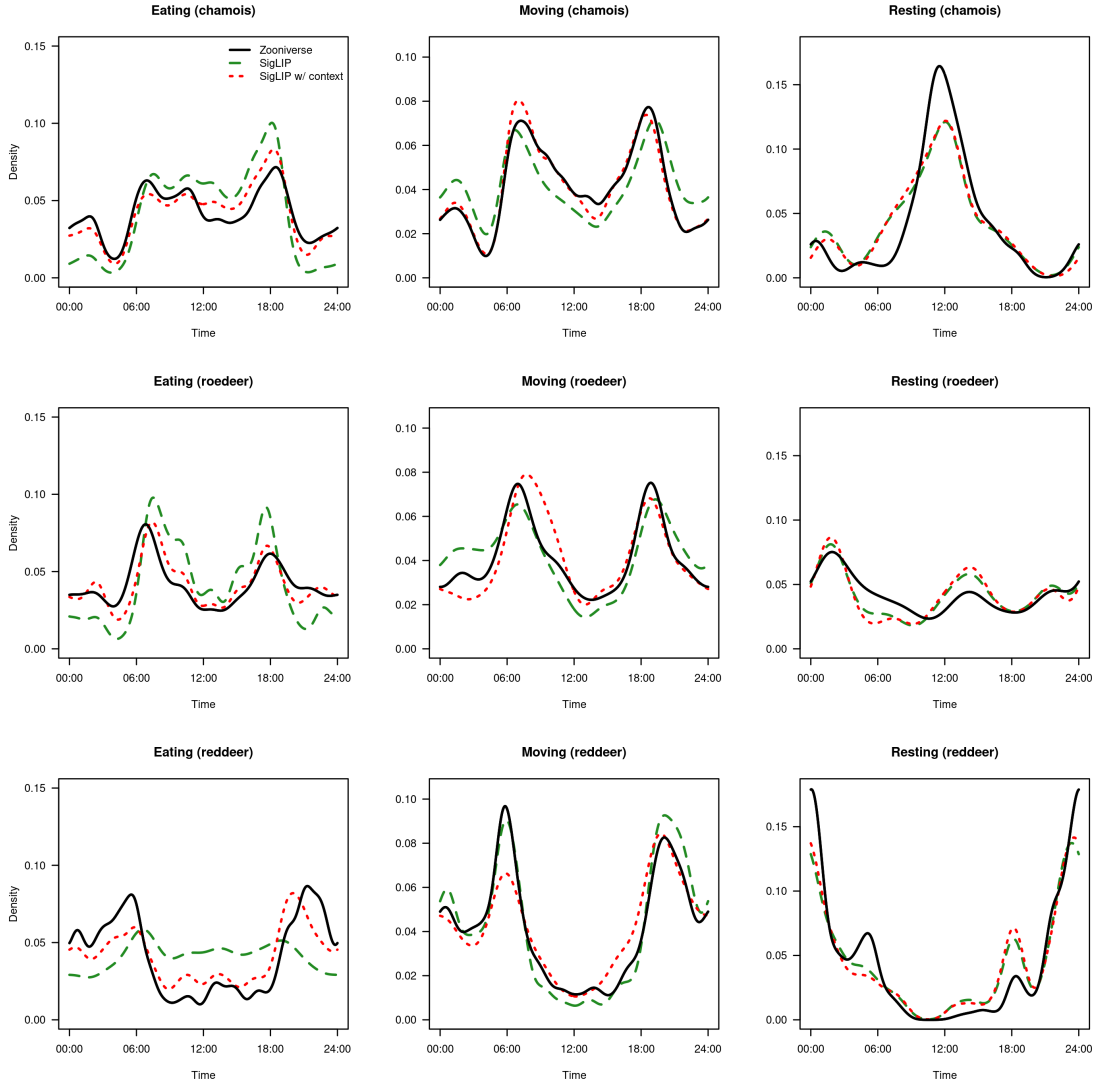


Figure 6: Activity patterns of SigLIP, with and without day-night context, using the complete dataset.

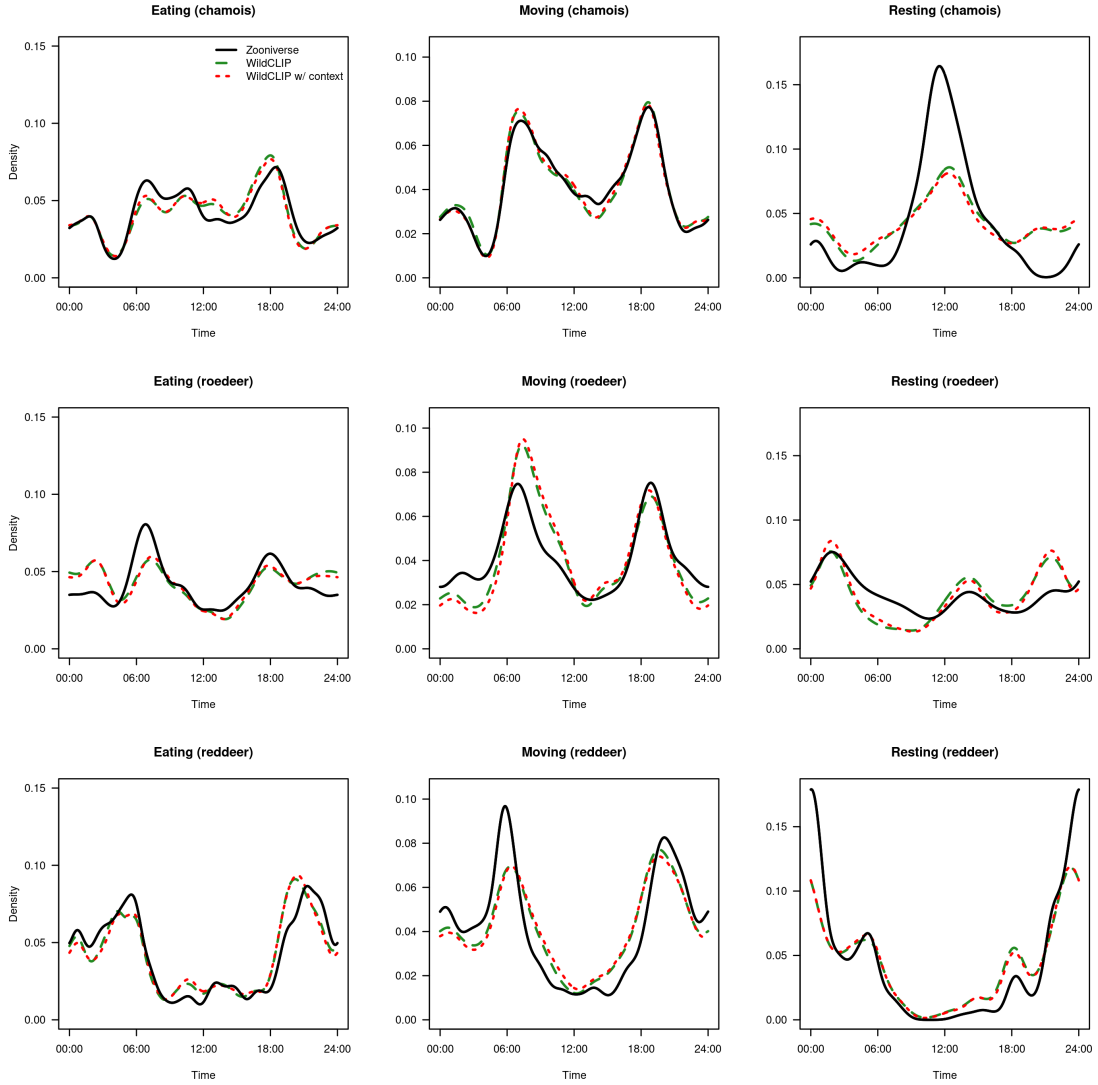


Figure 7: Activity patterns of WildCLIP, with and without day-night context, using the complete dataset.

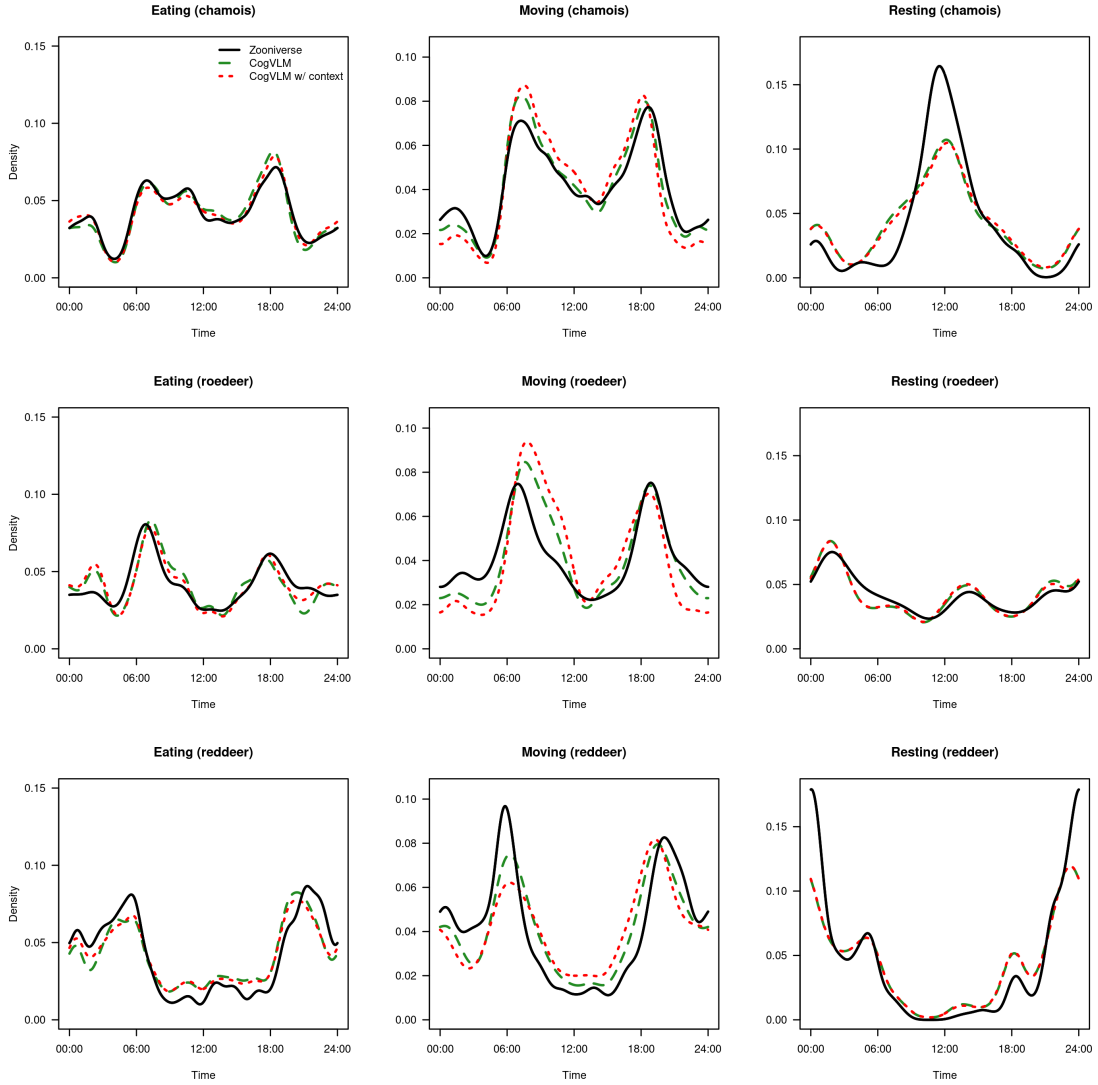


Figure 8: Activity patterns of CogVLM, with and without day-night context, using the complete dataset.

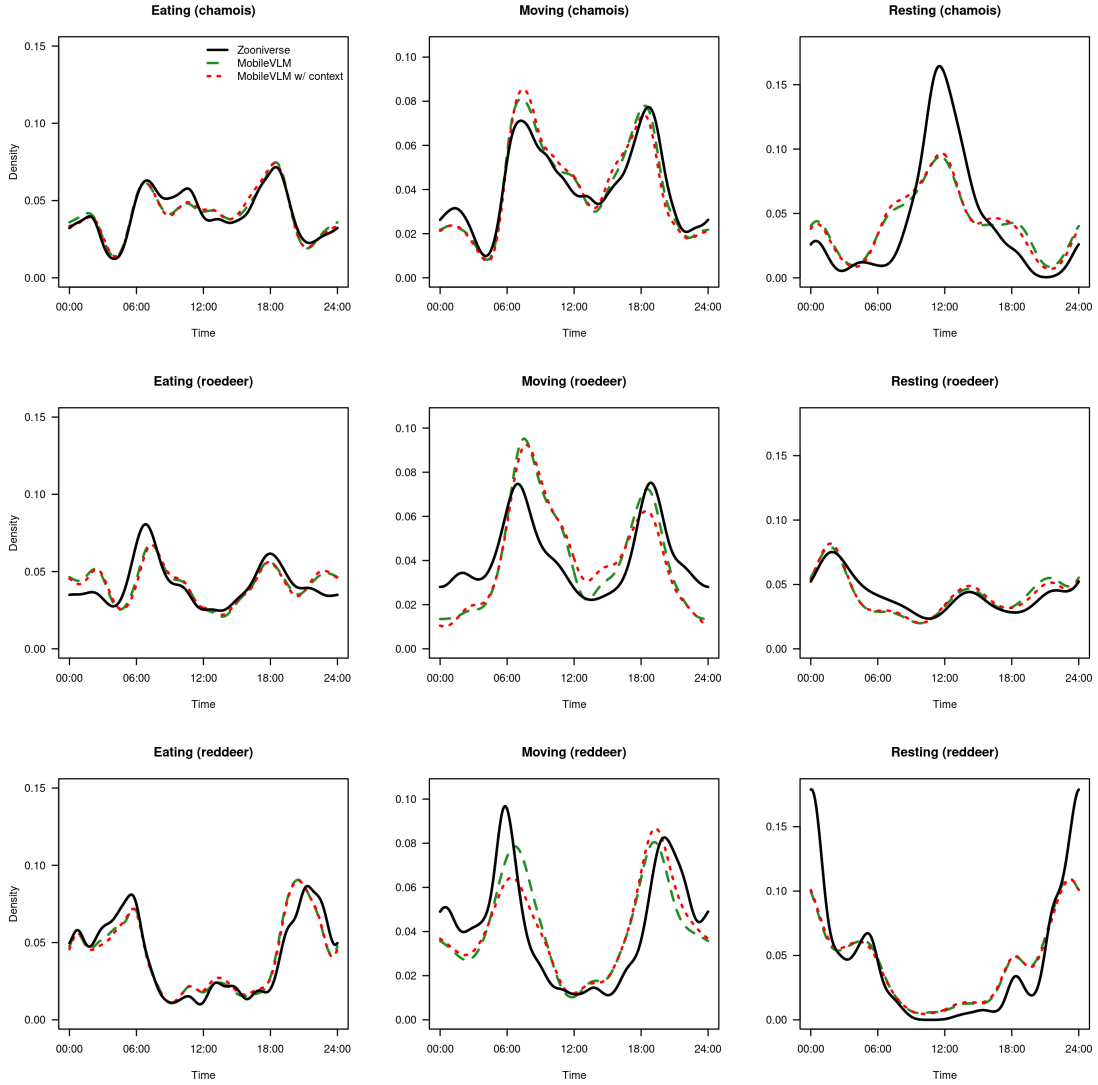


Figure 9: Activity patterns of MobileVLM V2 with and without day-night context, using the complete dataset.

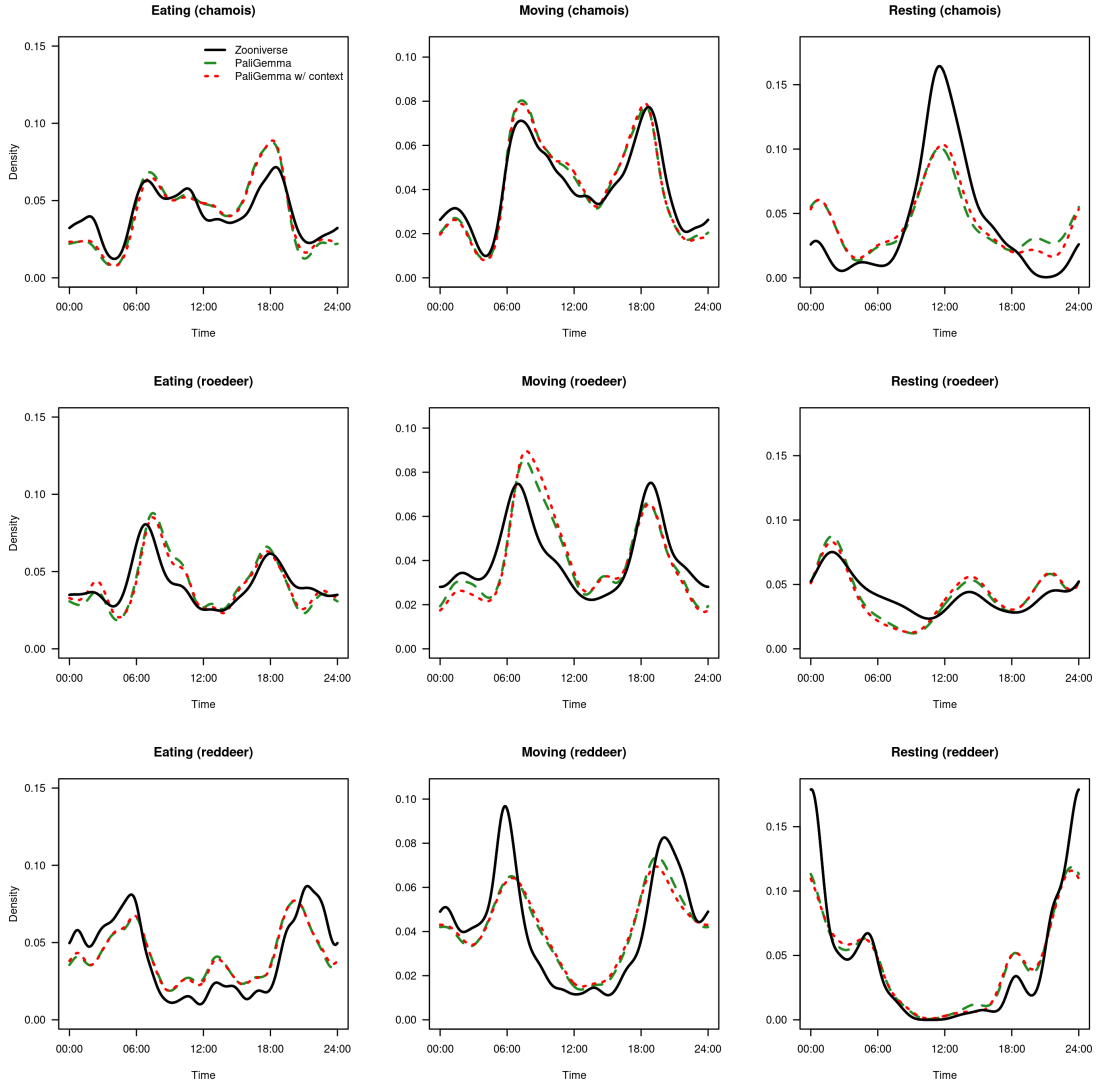


Figure 10: Activity patterns of PaliGemma, with and without day-night context, using the complete dataset.

Prompt order	CogVLM			PaliGemma		
	Eating	Moving	Resting	Eating	Moving	Resting
eating-moving-resting	99.37	93.71	99.28	90.57	91.80	96.87
eating-resting-moving	95.44	98.16	99.28	86.01	96.61	94.46
moving-resting-eating	99.53	89.07	99.28	92.61	91.14	97.35
moving-eating-resting	99.69	84.72	99.28	85.38	93.12	97.83
resting-moving-eating	99.21	94.30	98.80	91.51	85.02	98.80
resting-eating-moving	96.07	98.19	98.80	81.29	95.52	95.90
average	98.22	93.02	99.12	87.89	92.20	96.87

Table 5: Behavior classification accuracy of the CogVLM and PaliGemma for different prompt permutations: in each prompt, the behaviors are given in a different order. For example, in the first row the prompt would be "Is the animal in the image eating, moving or resting?" and in the second row the prompt would be "Is the animal in the image eating, resting or moving?". To avoid complicating the presentation, the results in the study are given for the first permutation, as this is the one we used before studying the permutations.