Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5894

Advancing Biodiversity Monitoring by Integrating Multimodal AI Models into Camera Trap Workflow

Luiz Alencar [**Federal University of Amazonas** | ***luiz.alencar@icomp.ufam.edu.br***]
Fagner Cunha [**Federal University of Amazonas** | ***fagner.cunha@icomp.ufam.edu.br***]
Eulanda M. dos Santos [**Federal University of Amazonas** | ***emsantos@icomp.ufam.edu.br***]

Institute of Computing. Av. Gen. Rodrigo Octávio, 6200 Setor Norte do Campus Universitário - Coroado, Manaus - AM, 69080-900

Received: 12 April 2025 • **Accepted:** 26 November 2025 • **Published:** 15 April 2026

Abstract Camera trap is an important non-invasive technique for wildlife monitoring. A typical camera-trap workflow involves various relevant tasks, such as filtering empty images, classifying animal species and identifying animal behavior. In this study, we explore the application of large-scale multimodal language models (MLLMs) for processing camera trap images to perform these three tasks. We evaluate the performance of four state-of-the-art models across these tasks, precisely BLIP, CLIP, Gemini, and GPT with zero-shot and few-shot learning methodologies. Our experiments showed several interesting results. First, few-shot learning significantly enhanced model performance in filtering empty images, with BLIP achieving a much higher accuracy (91.0%) compared to only 7.61% of its zero-shot counterpart. In the task of animal species classification, Gemini showed strong baseline performance, reaching 75.89 % of accuracy with zero-shot. In terms of identifying animal behavior, two scenarios were investigated: using single image or sequences of images. The results indicate that sequence-based processing improves behavioral analysis, with BLIP attaining the highest accuracy (75.57 %) in this task. In general, our study emphasizes the limitations of the zero-shot approach in complex tasks while highlights the effective potential of few-shot and sequence-based learning to address challenging problems such as empty images, and species misclassifications. These findings demonstrate the efficacy of advanced MLLMs in automating biodiversity monitoring, offering a scalable and accurate solution for processing large-scale datasets, and advancing conservation science.

Keywords: Multimodal language models, Camera trap data, Animal detection, Species identification, Behavior analysis

1 Introduction

Monitoring biodiversity is critical for understanding and preserving ecosystems, especially in the face of escalating environmental pressures. Camera traps have become indispensable tools for wildlife monitoring, providing rich datasets that offer insights into species distribution, abundance, and behaviors [Yang *et al*., 2021; Tan *et al*., 2022; Guo *et al*., 2024]. These datasets are pivotal for addressing conservation initiatives, such as identifying population declines and monitoring ecological interactions. However, extracting meaningful information from these datasets poses significant challenges, such as dealing with data imbalance [Cunha *et al*., 2023], the prevalence of empty images [Swanson *et al*., 2015b], poor image quality, and errors in species labeling [Iannarilli *et al*., 2021; Alencar *et al*., 2024]. These and other challenges must be addressed in the camera-trap workflow.

This workflow is a structured process used in biodiversity monitoring to collect, process, and analyze images captured by automated cameras in natural environments. It includes several key tasks essential for extracting ecological information. One primary task is filtering empty images [Cunha *et al*., 2021; Alencar *et al*., 2023, 2024], as a significant portion of the collected data (often up to 83% in some datasets, as in Elephant Expedition for example [Willi *et al*., 2019]) consists of non-informative images captured due to false triggers caused by environmental factors such as wind. Another critical task is species classification [Vecvanags *et al*., 2022; Choiński *et al*., 2021; Willi *et al*., 2019; Tabak *et al*., 2019; Beery *et al*., 2018], which involves identifying the species present in an image, such as distinguishing between a jaguar and a puma. Additionally, the workflow includes animal behavior identification [Norouzzadeh *et al*., 2018], where single images or image sequences are analyzed to determine behaviors such as movement, resting, or feeding.

Traditional approaches applied in the camera trap workflow, including supervised learning techniques like convolutional neural networks (CNNs), have achieved substantial progress in some tasks, such as detecting animals and identifying species [Cunha *et al*., 2021; Binta Islam *et al*., 2023; Vecvanags *et al*., 2022]. These methods, however, face limitations. They require extensive labeled training datasets, which are often difficult to obtain for rare species. Moreover, CNN-based models struggle with complex environmental features, such as dense vegetation or low lighting [Alencar *et al*., 2023], which obscure animals in images [Vélez *et al*., 2023]. Consequently, despite their potential, these approaches often exhibit variable performance across different ecological settings and require human verification to address classification errors [Vélez *et al*., 2023].

Zero-shot and few-shot learning approaches offer the opportunity to use MLLMs to reduce dependence on extensive

labeled datasets while integrating visual and textual data for enhanced contextual understanding [Li *et al*., 2022; Radford *et al*., 2021; Liu *et al*., 2024]. It has been shown that these models are able to achieve high performance in some tasks, such as filtering out empty images, with minimal supervision [Alencar *et al*., 2024]. Therefore, it is expected that automating other key tasks with multimodal AI models can improve efficiency, reduce manual effort, and enhance the accuracy of biodiversity assessments in general.

The study presented in this paper aims to explore the potential of four state-of-the-art MLLMs to improve three important tasks involving camera trap images: filtering empty images, classifying animal species and identifying animal behavior. The models investigated are CLIP, BLIP, GPT (version: GPT4-o), and Gemini (version: Gemini 2.0 flash-exp). Our work makes a significant contribution to the field of wildlife monitoring by integrating advanced multimodal AI methods with the camera trap workflow. This approach not only addresses long-standing challenges in data processing but also facilitates large-scale biodiversity assessments with high efficiency and accuracy [Wu *et al*., 2023; Fu *et al*., 2024]. By reducing manual labor and enhancing model performance, we may provide a robust framework for supporting conservation initiatives and advancing ecological research [Iannarilli *et al*., 2021; Vélez *et al*., 2023]. Our work seeks to answer the following questions:

- RQ1: How effectively can large-scale MLLMs improve the accuracy of three key camera trap workflow tasks: filtering empty images; classifying animal species; and identifying animal behavior?
- RQ2: What are the comparative advantages and limitations of zero-shot and few-shot learning approaches in enhancing the performance of MLLMs when handling complex visual tasks such the three tasks investigated in this paper?

This paper extends our previous work [Alencar *et al*., 2024] with the following new contributions: 1) we broaden the camera-trap workflow from a single task (empty-image filtering) to a multi-task setting by including species classification and behavior identification; 2) we evaluate four state-of-the-art MLLMs under both zero-shot and few-shot paradigms, while in [Alencar *et al*., 2024] only three models were investigated; and 3) we expand the dataset from seasons 1–4 to seasons 1–6 of the Snapshot Serengeti project, providing a larger and more diverse benchmark. In addition, the experimental evaluation has been largely revised and extended, including a detailed class-wise analysis of recall, precision, and F1 scores, and the analysis of the impact of sequence-based inputs for behavior recognition. These contributions go beyond the scope of our previous study and establish a more comprehensive and reproducible framework for applying MLLMs to biodiversity monitoring.

The remainder of this paper is organized as follows. Section 2 reviews related work, highlighting advancements and limitations in AI-driven biodiversity monitoring. Section 3 details our methodology, describing the multimodal models used, datasets, preprocessing steps, and the experimental setup for the zero-shot and few-shot learning approaches. Section 4 presents the results, analyzing model performance across the tasks investigated. Finally, Section 5 concludes the study, summarizing key insights and outlining potential improvements for integrating MLLMs into conservation workflows.

2 Related Work

Camera traps are invaluable tools for monitoring wildlife, enabling the collection of extensive ecological data. Recent advancements in AI, particularly MLLMs [Wang *et al*., 2024; Muhtar *et al*., 2024; Ma *et al*., 2024; Driess *et al*., 2023; Koh *et al*., 2024], have been fundamental in dealing with the challenges involved in the camera trap workflow. This section reviews key aspects from recent literature, highlighting gaps that motivated this research.

Most studies on camera trap data focus on a single task, addressing only one specific challenge of the camera trap workflow. For instance, Cunha *et al*. [2021] investigated empty image filtering to reduce the proportion of images without animals in large datasets. Similarly, Binta Islam *et al*. [2023] concentrated on species classification. In contrast, Norouzzadeh *et al*. [2018] conducted a comprehensive study that tackled three key tasks: empty image filtering; species classification; and behavior identification. They applied deep learning models to detect empty images, to identify 48 possible species, as well as to classify six animal behaviors in camera trap images. Their results showed that the investigated models performed well on the three tasks, indicating that AI models can save manual labor significantly, advancing wildlife behavior analysis.

It is important to note that deep learning methods were employed in all the aforementioned works. In particular, CNNs have been widely adopted to accurately perform these tasks. More recently, however, there is an increasing interest in MLLMs to take advantage of the capacity of these models to deal with multimodal data. Most of the approaches integrate visual and textual data, enhancing the automation and scalability of biodiversity monitoring workflows. In the next section, we describe recent works focused on using large language models (LLMs) or MLLMs in different tasks of the camera trap workflow.

2.1 Multimodal models in the camera trap workflow

The most common task investigated is animal species classification, as done in [Fabian *et al*., 2023]. The authors introduce a zero-shot animal species classification framework employing multimodal foundation LMMs. Their approach uses instruction tuning on vision-language models to generate descriptive textual representations of animals in camera trap images. These descriptions are then matched with a knowledge base of species information, enabling classification without labeled training data. The study evaluates LLaVA-7B models on the curated Magdalena Camera Traps dataset. Results demonstrate that instruction tuning significantly improves classification accuracy, with the best model achieving 70.12% micro accuracy, surpassing naive CLIP-based baselines by over 25%. The results highlight the potential of LMMs for

species identification in wildlife monitoring, while also identify key limitations, including high computational costs and the dependency on comprehensive, high-quality knowledge bases.

In [Vyskočil and Picek, 2024], the problem of automatic species categorization is tackled focusing on reducing overfitting to location-specific backgrounds and improving zero-shot classification. They evaluated state-of-the-art CNN and Transformer-based classifiers (such as BEiTV2 and EfficientViT) on three datasets: WCT (New Zealand), CCT20 (USA), and CEF (Europe). Their best-performing approach combined MegaDetector [Fennell *et al*., 2022] for object detection with two independent classifiers, reducing the error by 42% (CCT20), 48% (CEF), and 75% (WCT) compared to a single classifier. They also tested Segment Anything Model (SAM) for background removal, but the results indicated that it slightly reduced classification accuracy. For zero-shot classification, they explored DINOv2, BioCLIP, BLIP, and ChatGPT, with DINOv2G achieving near-supervised accuracy (Top1: 83.2% CCT20, 87.5% CEF) in an image retrieval setting using similarity search. Their results reinforce the potential of the zero-shot approach in reducing the need for dataset-specific retraining, while maintaining high classification performance.

In their turn, Gabeff *et al*. [2024] addressed the problem of building image captions for animal species identification using a vision-language model. They developed WildCLIP, an adaptation of CLIP fine-tuned on wildlife camera-trap data, incorporating an adapter module to enhance flexibility and enable few-shot learning of new attributes. Their approach facilitates open-vocabulary image retrieval, overcoming limitations of traditional classification models that rely on fixed categories and large labeled datasets. WildCLIP was benchmarked on the Snapshot Serengeti dataset, demonstrating improved retrieval of novel attributes compared to standard CLIP. The model's performance was quantified using mean average precision (mAP), achieving 64% for base vocabulary attributes and 40% for novel vocabulary attributes. Adaptation with few-shot learning improved novel vocabulary retrieval to 38%, while also mitigated catastrophic forgetting through a vocabulary replay loss. Their results highlight WildCLIP's potential in aiding automated wildlife annotation by enabling more detailed and flexible image queries.

Santamaria *et al*. [2024] developed CATALOG (Camera Trap Language-guided Contrastive Learning Model) to tackle the domain shift problem in species classification in camera-trap images, where variations in lighting, camouflage, and occlusions make recognition challenging. Their methodology integrates multiple Foundation Models (FMs), including CLIP, LLaVA, BERT, and an LLM, to extract and align visual and textual features using a contrastive learning framework. The model was trained on the Snapshot Serengeti dataset (340,972 images) and tested on the Caltech dataset (45,912 images), evaluating its performance in both in-domain and out-of-domain settings. In a zero-shot evaluation, CATALOG achieved 48.59% accuracy on Cis-Test [1] and 41.92% on Trans-Test [2], significantly outperforming previous models such as WildCLIP (40.38% Cis-Test, 38.90% Trans-Test) and CLIP ViT-B/16 (39.14% Cis-Test, 34.67% Trans-Test). In an in-domain evaluation, CATALOG reached 90.63% accuracy on Snapshot Serengeti, surpassing WildCLIP (61.78%) and WildCLIP-LwF (64.39%). These results highlight CATALOG's superior ability to generalize across domains and enhance species classification in challenging real-world camera-trap scenarios.

The problem investigated in our previous work [Alencar *et al*., 2024] is the application of MLLMs for filtering empty images. The study evaluates CLIP, BLIP, and Gemini under zero-shot and few-shot learning paradigms, comparing their performance against a ResNet50-Siamese model tailored for this task. Experiments were conducted on three datasets: Snapshot Serengeti, Caltech, and WCS. Results indicate that Gemini (zero-shot) achieved the highest accuracy (86.73%), while BLIP (few-shot) performed competitively, in some cases surpassing ResNet50. However, zero-shot models exhibited a notable weakness in misclassifying non-empty images, underscoring the importance of few-shot adaptation. The study highlights the capacity of MLLMs in automated image filtering in ecological research, while also reinforces challenges, such as high computational cost and sensitivity to environmental conditions.

Dussert *et al*. [2024] explored zero-shot animal behavior classification using image-text foundation models. The study evaluates contrastive learning models (CLIP, SigLIP, WildCLIP) alongside MLLMs (CogVLM, MobileVLM V2) for predicting behaviors in camera trap images without fine-tuning. The models were tested on a dataset of European fauna, specifically targeting three behaviors (eating, moving and resting) of three species: chamois, red deer, and roe deer. The results indicated that CogVLM achieved the highest accuracy (97.45%), followed by SigLIP (91.14%) and WildCLIP (92.55%), while CLIP underperformed (71.75%). It is important to mention that, incorporating day-night context improved accuracy for contrastive models but reduced the performances of MLLMs. The study highlights the potential of foundation models for automated behavioral analysis in wildlife studies, offering a scalable alternative to manual annotation while maintaining high predictive accuracy.

The scenario observed considering works devoted to apply LLMs or MLLMs in tasks of the camera trap workflow is the same observed in previous works that use more traditional deep learning methods: only one task is dealt with. Unlike these previous works, Dorm *et al*. [2025] evaluate the ecological knowledge and reasoning abilities of GPT-4o and Gemini 1.5 Pro across five tasks: species presence prediction, range mapping, listing endangered species, threat assessment, and trait estimation. Using expert-derived datasets like IUCN Red List and AVONET bird traits, the models were benchmarked against established ecological data. GPT-4o outperformed Gemini in species presence prediction (78.0% vs. 70.8%), but both struggled with range maps (F1 scores: 24.7% Gemini, 20.9% GPT-4o) and threat classification, performing only slightly above random guessing. For listing endangered species, GPT-4o achieved higher recall (20.1%) than Gemini (18.0%) but still missed many species. While the

[1] Measures in-domain performance—how well the model performs when environmental conditions are consistent with the training data.

[2] Measures out-of-domain generalization—how well the model adapts to new environments (i.e., domain shift).

MLLMs showed potential, the study highlights their spatial inaccuracies, reasoning limitations, and need for domain-specific fine-tuning to enhance ecological applications.

In this paper, we also address different tasks of the camera trap workflow. Here, however, except for species classification, the tasks investigated are different from those studied in [Dorm *et al*., 2025]. Three tasks are dealt with in this work: filtering empty images; classifying animal species; and identifying animal behavior. The first two are the most basic tasks of the camera trap workflow [Leorna and Brinkman, 2022]. The third can also be added to this group, since the results of these three tasks are often used to determine species diversity, distribution, abundance, behavior, etc. [Leorna and Brinkman, 2022]. The methodology employed in this work is detailed in the next section.

3 Methodology

This study evaluates the performance of four multimodal models—Gemini, GPT, BLIP, and CLIP—on three tasks of the camera trap workflow using the few-shot and zero-shot learning paradigms. The objective is to determine how effectively these pre-trained models generalize to the three tasks. Notably, Gemini and GPT were only evaluated using zero-shot learning due to their closed-loop architecture. All investigated models are described below.

3.1 Models

Gemini is a multimodal AI model designed to process and generate both text and image-based outputs. Its architecture integrates visual and linguistic information, allowing it to interpret images and respond to prompts effectively [Team *et al*., 2023]. While Gemini is capable of learning via prompt engineering, the methodology discussed in this article focuses on fine-tuning through model weight adjustments. In this context, since Gemini is a closed-source model, fine-tuning can be performed through specific services (e.g., Vertex AI [3]), but local fine-tuning on external datasets is not feasible, limiting its application in this paper to zero-shot scenarios. As a result, Gemini relies entirely on its pre-trained knowledge when performing the tasks investigated in our work.

GPT, primarily recognized for its advanced text generation abilities, has recently evolved to include multimodal functionalities, enabling it to interpret and generate responses based on visual inputs [Islam and Moushi, 2024]. Its strength lies in processing complex textual inputs and generating detailed, contextually aware responses, making it highly effective for tasks that demand nuanced language comprehension alongside basic visual analysis. Similarly to Gemini, GPT operates as a closed model, meaning local fine-tuning on external datasets is not feasible, although service-based fine-tuning options exist.

GPT also supports in-context learning through prompt conditioning with textual and visual examples using the same API interface employed for zero-shot inference. However, incorporating in-context examples substantially increases the number of input tokens, leading to higher inference costs and reduced scalability across large evaluation datasets. For these reasons, and to ensure a cost-efficient, scalable, and methodologically consistent evaluation setting, GPT is investigated exclusively under the zero-shot learning scenario in this work.

BLIP (Bootstrapped Language-Image Pretraining) is a multimodal model specifically designed for tasks that combine visual and textual information, such as image captioning, visual question answering, and image-text retrieval [Li *et al*., 2022]. BLIP supports in-context learning, allowing it to adapt to few-shot learning with sample prompts and responses without the use of paid APIs. Its flexibility in handling both zero-shot and few-shot learning makes it particularly effective in tasks like species classification and behavior identification, where limited labeled data is available.

CLIP (Contrastive Language–Image Pretraining) is a robust model developed to link images and text through contrastive learning [Radford *et al*., 2021]. It is expected to perform very well on zero-shot image classification by matching visual inputs with textual descriptions, such as “a photo of a zebra” or “an animal eating”. CLIP also supports few-shot learning through light fine-tuning with a small number of labeled examples. Its ability to generalize across diverse image-text pairs may help CLIP to be highly effective in camera-trap tasks, from detecting animal presence to identifying specific species.

Regarding the training process of CLIP and BLIP in our experiments, both models underwent light fine-tuning in a local execution environment. For CLIP, we optimized the contrastive loss on a small subset of labeled samples per class, following its original design for aligning image and text embeddings. BLIP was also fine-tuned using the same labeled samples used with CLIP, updating its weights to adapt the visual-language alignment to the target tasks. In both cases, the fine-tuning was deliberately restricted to a small number of iterations and samples (summarized in Table 1), consistent with the few-shot learning paradigm, in order to highlight the models’ ability to adapt with minimal supervision rather than large-scale retraining.

3.2 Data Collection and Preprocessing

The dataset used in our study is based on the Snapshot Serengeti project [Swanson *et al*., 2015a], which is one of the most comprehensive and publicly available camera trap datasets. Data collection has occurred over multiple years, with motion-activated cameras deployed continuously in the Serengeti National Park, Tanzania, since 2010. These camera traps operate throughout the day and night, capturing images whenever animal movement triggers the sensors. As a result, the dataset encompasses activity across different seasons and environmental conditions. The Snapshot Serengeti dataset has played a crucial role in ecological research, providing insights into species distribution, behavior, and predator-prey interactions. The key characteristics of the dataset include:

- Scale and Data Volume: The dataset contains approximately 2.65 million camera trap image sequences, totaling 7.1 million images, covering seasons one through eleven of the Snapshot Serengeti project. These images

[3] Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-prepare

provide an extensive temporal record of wildlife activity, making the dataset one of the most comprehensive long-term studies of its kind.
- Class Imbalance (empty vs. images with animals): A significant challenge with Snapshot Serengeti is the imbalance between empty images and images containing animals. Approximately 76% of the images are empty, as camera traps often get triggered by movement from non-animal sources. This makes animal detection particularly challenging, demanding robust filtering techniques to reduce false positives.

3.3 Few-shot and Zero-shot Learning Approaches

- Zero-shot Learning: In this setup, models were tested without prior exposure to the dataset, relying entirely on their pre-trained knowledge. CLIP matched images to predefined textual descriptions (e.g., "a photo of a zebra"). Gemini, GPT, and BLIP were prompted with specific instructions tailored to each task. However, as previously explained, due to the nature of Gemini and GPT as closed models, they were exclusively evaluated in the zero-shot context without additional fine-tuning.
- Few-shot Learning: In this approach, models were provided with a small subset of labeled images prior to evaluation. Our few-shot learning approach consisted of providing representative examples per class within the prompt, enabling the model to generalize to unseen samples. CLIP underwent light fine-tuning using a few examples per class, while BLIP leveraged in-context learning with sample prompts and corresponding labels. The number of images per class used to fine-tune both models is summarized in Table 1.

It is important to reinforce that the comparison of zero-shot and few-shot learning conducted in this paper is not entirely symmetrical across all models. Gemini and GPT are closed-source models with local fine-tuning unfeasible. In contrast, CLIP and BLIP allow light fine-tuning or local in-context learning. Therefore, while all models were compared on the same tasks, we explicitly distinguish between open models (CLIP, BLIP), investigated in both zero-shot and few-shot scenarios, and closed models (GPT, Gemini), evaluated only using zero-shot with optimized prompts.

3.4 Experimental Setup

Three tasks of the camera trap workflow were designed to evaluate the performance of four MLLMs. The methodology followed in this work ensures a structured and consistent evaluation of model performance across the tasks, focused on effectively highlighting each model's strengths and limitations in both zero-shot and few-shot learning scenarios. To maximize the number of models included in our experiments while accounting for hardware constraints, we opted for smaller subsets of data. The data spans seasons 1 through 6 of Serengeti, ensuring a diverse range of instances. The dataset was carefully divided into training, validation, and testing subsets to guarantee a fair evaluation and to avoid data leakage across tasks. To minimize bias, we stratified the subsets by location, meaning that images from the same camera location were not divided into different subsets, which is a standard partition strategy in the literature [Beery *et al*., 2018; Schneider *et al*., 2020]. This ensures that the models were evaluated on entirely new locations not seen during training, providing a stronger assessment of generalization.

Table 1 presents the exact distribution of images per class for each of the three tasks investigated: empty image filtering, species classification, and behavior identification. For the filtering task, near-balanced subsets were obtained (e.g., 10,114 animal vs. 9,886 empty in training; and 4,981 vs. 5,019 in testing). For behavior classification, the dataset was divided across the three classes (moving, resting, eating), with approximately 6,586–6,732 instances per class for training and around 3,300 per class for testing. In terms of species classification, we considered nine of the most frequent species in the dataset, with number of training instances per class ranging from 2,161 bird images to 2,257 gazelle images, and near-balanced classes for testing (e.g., 1,125 lion images and 1,136 buffalo images).

The prompts employed were tailored to the specific operational mechanisms of each model. For Gemini, GPT, and BLIP, we employed direct classification prompts restricted to the predefined categories of each task, ensuring consistency across models. CLIP, in contrast, relies on image–text similarity rather than open-ended generation. Therefore, we used predefined textual descriptions for each class and matched them with the given image. These differences reflect the distinct ways models process visual information.

Several prompt variations were evaluated during the validation of this work, such as changes in detail (e.g., "Identify the animal species" vs. "Identify the animal species and return only one option from the following list") and alternative phrasings (e.g., "Is there an animal in the picture?" vs. "Does this image show a background without animals or one with an animal?"). We selected the final prompts, detailed in the next subsections, that consistently produced the most interpretable results. This decision was based exclusively on training and validation performance to prevent tailoring the prompts to the test set and ensure an unbiased evaluation. Prompt engineering was particularly crucial for closed models (GPT and Gemini), where it acted as the only form of adaptation.

3.4.1 Empty Image Filtering

This task focused on indicating whether an animal was present in the image, or whether it was considered an empty image, and was conducted using the following prompts:

- Gemini, GPT, and BLIP: "A photo of an animal: 0) no and 1) yes."
- CLIP: Matched images with phrases like "a photo of a background" (for empty images) or "a photo of an animal" (for images containing animals).

3.4.2 Species Classification

In this task, models were required to identify the species of the animal from a predefined list of nine classes: hyena, zebra, giraffe, buffalo, gazelle, wildebeest, elephant, lion and

Table 1. Number of images in each dataset used for each task.

Subset	Category	Filtering Classifier		Behaviour Classifier			Species Classifier								
		Animal	Empty	Moving	Resting	Eating	Gazelle	Hyena	Elephant	Lion	Bird	Buffalo	Zebra	Wildebeest	Giraffe
Val	Filtering Classifier	1538	1462	-	-	-	-	-	-	-	-	-	-	-	-
	Behaviour Classifier	-	-	1033	988	979	-	-	-	-	-	-	-	-	-
	Species Classifier	-	-	-	-	-	350	347	341	334	333	329	323	322	321
Test	Filtering Classifier	4981	5019	-	-	-	-	-	-	-	-	-	-	-	-
	Behaviour Classifier	-	-	3328	3309	3363	-	-	-	-	-	-	-	-	-
	Species Classifier	-	-	-	-	-	1136	1130	1127	1125	1127	1136	1111	1122	1103
Train	Filtering Classifier	10114	9886	-	-	-	-	-	-	-	-	-	-	-	-
	Behaviour Classifier	-	-	6586	6732	6682	-	-	-	-	-	-	-	-	-
	Species Classifier	-	-	-	-	-	2257	2248	2245	2220	2161	2248	2245	2209	2203

bird. These species were selected based on their high occurrence in the training dataset, ensuring sufficient representation for model training and evaluation. The chosen species also exhibit diverse visual characteristics, contributing to a balanced classification challenge while reflecting the most frequently observed animals in the studied environment. This design separates the empty-image filtering task from species classification to leverage specialized architectures for each. Preliminary tests showed improved accuracy compared to a joint multi-class approach.

- Prompts for Gemini, GPT, and BLIP: "Identify the species of the animal. Respond only with one of the following options: 0) hyena, 1) zebra, 2) giraffe, 3) buffalo, 4) gazelle, 5) wildebeest, 6) elephant, 7) lion, 8) bird."
- CLIP: Used image-text matching with species-specific descriptions, such as "a photo of a giraffe", "a photo of a lion", or "a photo of a zebra."

3.4.3 Behavior Identification

This task involved classifying animal behavior into one of three categories: moving, eating, or resting. The selection of these three behavior classes was based on their frequent occurrence in camera trap datasets, annotation feasibility, and practical application in wildlife monitoring. These categories represent fundamental and easily distinguishable behaviors that appear most frequently in labeled datasets, ensuring sufficient training data for machine learning models. More specific behaviors are either rarely captured in camera trap images or are difficult to label consistently, leading to class imbalance and annotation challenges. Additionally, a simplified classification framework enhances the robustness and generalization of machine learning models, making them more effective for large-scale ecological studies. By focusing on these core behaviors, the classification supports scalable species monitoring and facilitates the analysis of broad activity patterns relevant to conservation efforts.

In this task, models were evaluated using both single and sequences of images captured during the same camera-trap event. On average, the number of images per sequence ranged from 3 to 5, depending on camera trigger conditions. Our hypothesis is that, while single images provide isolated snapshots of animal activity, image sequences offer temporal context, allowing the models to better infer behaviors such as moving, eating, or resting. Therefore, this dual approach tests the models' ability to process not only static visual information but also recognize patterns across consecutive frames, which is critical for accurately interpreting animal behavior in dynamic, real-world environments.

- Prompts for Gemini, GPT, and BLIP: "Identify the behavior the animals are performing. Respond only with one of the following options: 0) moving, 1) eating, 2) resting."
- CLIP: Used image-text matching with descriptions such as "a photo of an animal moving, "a photo of an animal eating," or "a photo of an animal resting."

All inferences with GPT and Gemini were executed via their official APIs. GPT-4o was accessed through the OpenAI API, and Gemini 2.0 flash-exp was accessed through the Google AI Studio API. Images and structured textual prompts were submitted by coded instructions, and the models' outputs were parsed automatically to ensure reproducibility and avoid human intervention during evaluation.

4 Experiments and Results

This section presents the performances reached by the models investigated. The results are shown using the following evaluation metrics: accuracy, precision, recall, and F1-score. Table 2 summarizes all the results, highlighting in bold the best values attained per task. It is important to mention that to better compare the estimates of accuracy, we also report in this table 95% binomial confidence intervals (CI95) for each model and task. A binomial CI treats every test prediction as an independent Bernoulli trial (success = correct prediction, failure = error) and quantifies the uncertainty of the observed accuracy. Using the number of test samples n and the measured accuracy p, the 95% CI is computed from the exact Clopper–Pearson method (or an equivalent normal approximation) to give a lower and upper bound $[p_{\text{low}}, p_{\text{high}}]$ such that, under repeated sampling, the true accuracy would fall inside this range 95% of the time. Non-overlapping intervals suggest significant performance differences between models.

The results of the statistical test are further clarified by the following examples. In the image filtering task, for instance, BLIP-FewShot achieved an accuracy of 0.91 with CI95 = [0.904, 0.915], clearly separated from CLIP's 0.7987 [0.791, 0.806]. Similar non-overlapping intervals appear across most tasks: in species classification, BLIP-FewShot 0.7524 [0.744, 0.761] vs. CLIP 0.4558 [0.446, 0.466]; in behavior classification, BLIP-FewShot-Seq 0.7557 [0.747, 0.764] vs. CLIP-FewShot 0.4817 [0.472, 0.491]. These confidence bounds demonstrate that the observed performance gaps are not attributable to random variation in the test set. Zero-shot models

Table 2. Performance attained by different models on the three tasks investigated.

Task	Model	Precision	Recall	F1 Score	Accuracy	CI95
animal	**BLIP-FewShot**	**0.9100**	**0.9100**	**0.9100**	**0.9100**	**[0.904, 0.915]**
	GPT	0.8693	0.8562	0.8552	0.8566	[0.850, 0.863]
	GEMINI	0.5695	0.5551	0.5536	0.8322	[0.825, 0.839]
	CLIP	0.8150	0.7983	0.7959	0.7987	[0.791, 0.806]
	CLIP-FewShot	0.7594	0.7589	0.7587	0.7588	[0.750, 0.767]
	BLIP	0.0137	0.0035	0.0055	0.0761	[0.071, 0.081]
behavior	**BLIP-FewShot-Seq**	**0.7640**	**0.7555**	**0.7567**	**0.7557**	**[0.747, 0.764]**
	BLIP-FewShot	0.7041	0.7011	0.7011	0.7014	[0.692, 0.710]
	GEMINI-Seq	0.5002	0.4616	0.4336	0.6171	[0.608, 0.627]
	GEMINI	0.4533	0.4312	0.3940	0.5766	[0.567, 0.586]
	CLIP-FewShot-Seq	0.5222	0.5007	0.5016	0.5006	[0.491, 0.510]
	GPT	0.6339	0.4813	0.4195	0.4819	[0.472, 0.492]
	CLIP-FewShot	0.4875	0.4824	0.4766	0.4817	[0.472, 0.491]
	GPT-Seq	0.4886	0.4053	0.3351	0.4058	[0.396, 0.415]
	CLIP-Seq	0.3431	0.3511	0.3397	0.3506	[0.341, 0.360]
	CLIP	0.3348	0.3430	0.3321	0.3425	[0.333, 0.352]
	BLIP-Seq	0.0477	0.0290	0.0361	0.0879	[0.083, 0.094]
	BLIP	0.0312	0.0002	0.0004	0.0011	[0.001, 0.002]
species	**BLIP-FewShot**	**0.7684**	**0.7532**	**0.7556**	0.7524	[0.744, 0.761]
	GEMINI	0.6912	0.6837	0.6859	**0.7589**	**[0.754, 0.764]**
	GPT	0.7036	0.6190	0.6296	0.6868	[0.678, 0.696]
	CLIP	0.5352	0.4580	0.4456	0.4558	[0.446, 0.466]
	CLIP-FewShot	0.3359	0.3201	0.3174	0.3202	[0.311, 0.329]
	BLIP	0.1861	0.0005	0.0011	0.0006	[0.000, 0.001]

were evaluated on the same held-out locations, so the intervals quantify their generalization to unseen camera sites. By reporting CI95 values for all models and tasks, the analysis provides a sound basis for concluding that the few-shot fine-tuned models consistently outperform their zero-shot counterparts.

In the following subsections, we discuss the results for each task individually.

4.1 Filtering Classifier

In this task, BLIP on the zero-shot approach exhibited the lowest performance: accuracy of 0.0761, precision of 0.0137, recall of 0.0035, and F1-score of 0.0055. These results highlight BLIP's significant limitations in detecting animal presence without prior exposure to the dataset, making it unreliable for real-world applications in its zero-shot form. However, after fine-tuning BLIP-FewShot with a small number of labeled examples (according to the number of images per class in Table 1), its performance improved significantly, reaching 0.910 across all metrics, which were the highest rates attained in the filtering empty images task. This suggests that BLIP benefits significantly from a small amount of domain-specific data, making it a viable option in few-shot scenarios.

In contrast, CLIP performed considerably better than BLIP in the zero-shot setting, attaining accuracy of 0.7987 and F1-score of 0.7959. However, an unexpected result was observed when CLIP was evaluated on the few-shot scenario: its performance slightly decreased, e.g. accuracy dropped to 0.7588. This suggests that CLIP's initial zero-shot learning capabilities were relatively well-optimized for this task, and fine-tuning it with few instances may have introduced biases or overfitting.

Finally, among the models tested, GPT achieved the highest overall performance in the zero-shot setting, with accuracy of 0.8566, precision of 0.8693, recall of 0.8562, and F1-score of 0.8552. This result indicates that GPT was the most reliable model in distinguishing non-empty from empty frames using zero-shot. Gemini also produced valid results, but with comparatively lower performance (e.g. F1-Score 0.5536) in this binary classification task.

4.1.1 Discussion

For real-world deployment, these results indicate that certain models, particularly BLIP in its zero-shot form, are unsuitable due to their poor accuracy and recall. The BLIP-FewShot and GPT models, on the other hand, achieved high performance, making them strong candidates for practical applications. However, real-world scenarios often require not only high accuracy but also robust performance across different conditions (e.g., varying lighting, occlusions, and animal positions). While accuracy rates around 85-91% (as seen with GPT and BLIP-FewShot) are promising, further testing on unseen datasets would be needed to confirm their reliability in operational settings. It is worth noting that the relatively high performance of GPT may be partly explained by prior exposure to the dataset, since the dataset was released in 2015 and no GPT snapshot predating is available for testing.

In addition, it is important to evaluate whether the models provide similar performance among the two classes involved in the task of filtering empty images. As it is shown in Figure 1, there are images whose correct class was assigned by only one of the investigated methods. For instance, it is possi-

Table 3. Recall results attained by different models in filtering empty images.

Model	Animal	Empty
BLIP	0.1528	0.0000
BLIP-FewShot	0.9119	0.9081
CLIP	0.6840	0.9125
CLIP-FewShot	0.7828	0.7350
Gemini	**0.9538**	0.7115
GPT	0.7635	**0.9490**

ble to observe in this figure that CLIP-FewShot was the only model able to identify that there is an animal centered in the middle of the image in Figure 1(a). In Figure 1(b), BLIP was the only model that correctly assigned the empty class to the image. The same happens with Gemini in Figure 1(c), which was the only model that classified the image as non-empty, since it is difficult to detect the animal in the lower left corner of the image due to the daylight.

In order to evaluate how the models dealt with each of the two classes, Table 3 shows the class-wise recall values. We can see in this table that models perform differently when classifying empty frames versus non-empty ones. It is noteworthy to mention that the ideal performance is a trade-off between both classes. In one hand, retrieving non-empty instances is important, even at the cost of degrading the empty class prediction. On the other, the error rate for the empty class cannot be too high, as this will result in the accumulation of a large amount of empty images.

CLIP, in its zero-shot setting, was not successful in attaining this trade-off: it achieved high recall for the empty class (0.9125) but a much lower rate for the non-empty (0.6840). This indicates that the model is more prone to classify instances as empty. After the few-shot fine-tuning, CLIP improved recall in detecting animals (0.7828) but dropped it for the empty class (0.7350). However, this shift indicates a better trade-off introduced during fine-tuning the model. A similar behavior is observed with GPT. It reached an intermediate recall for the non-empty class (0.7635), but showed the highest recall for empty images (0.9490). An opposite behavior is presented by Gemini, which excelled at recognizing non-empty images (recall = 0.9538), though its performance was significantly lower for the empty category (0.7115).

Finally, even though BLIP in zero-shot struggled significantly with the task in general: it reached recall = 0.00 for empty images for instance, BLIP-FewShot demonstrated the best balanced recall for both classes (0.9119 and 0.9081), indicating that fine-tuning enhanced the model's ability to correctly retrieve instances from both categories without bias toward one. These results highlight that identifying the presence of animals in images is usually more difficulty than recognizing empty scenes. Models may misclassify non-empty instances due to the wide variety of appearances, behaviors, and potential occlusions. On the other hand, empty frames tend to be more visually consistent, making them somewhat easier to detect—provided the model is well-trained for the task. In summary, models that can sustain high and balanced recall across both classes—such as GPT and BLIP-FewShot—are ideal for ensuring no critical detections are missed in wildlife monitoring or surveillance systems.

4.2 Species Classification

In this task, again BLIP in zero-shot learning struggled significantly, reaching accuracy of just 0.0006, along with precision of 0.1861, recall of 0.0005, and F1-score of 0.0011. As observed for the previous task, the few-shot approach improved substantially its performance, with accuracy rising to 0.7524 and F1-score to 0.7556. This highlights the model's heavy dependence on task-specific fine-tuning for effective classification. Unlike BLIP, the performance of CLIP in zero-shot was slightly better, achieving accuracy of 0.4558 and F1-score of 0.4456. However, again, the same surprising behavior of CLIP fine-tuned with few-shot learning (CLIP-FewShot) was noticed in this task: the performance declined, with accuracy dropping to 0.3202 and F1-score to 0.3174. This suggests that CLIP's capabilities fail to benefit from limited training samples, possibly due to overfitting or ineffective adaptation.

Among the evaluated models, Gemini demonstrated the best zero-shot performance: accuracy of 0.7589, precision of 0.6912, recall of 0.6837, and F1-score of 0.6859. GPT performed also well. It reached higher precision (0.7036), but the remaining metrics were lower: recall of 0.6190, F1-score of 0.6296 and accuracy of 0.6868. These results suggest that both Gemini and GPT leverage strong pretraining to generalize effectively to new species without additional supervision.

4.2.1 Error distribution across classes

In this task is also crucial to analyze how each model individually handles challenging cases. Figure 2 illustrates examples in which only one model provided the correct species label, while the remaining failed. In Figure 2(a), CLIP-FewShot was the only method able to correctly identify a gazelle. Figure 2(b) presents a similar situation, but in this case, GPT was the model that recognized the gazelle, possibly leveraging textual context in the visual scene. Lastly, Figure 2(c) shows one example where only BLIP-FewShot successfully identified the species as a lion, demonstrating its strength in combining vision and language representations to disambiguate complex visual features.

In addition, a detailed summarization of recall values per species is shown in Table 4). From this table, some observations can be made:

- BLIP (zero-shot) exhibited very limited performance in the species identification task, as reflected in its recall values across all classes. Specifically, BLIP achieved non-zero recall for only three species: Lion (0.0027), Giraffe (0.0018), and Zebra (0.0009).
- BLIP-FewShot showed a significant boost in recall, reaching above 0.57 in all classes. Notably, high recall rates were reached for giraffe (0.8812), hyena (0.8575), and elephant (0.8058). Even traditionally challenging classes like bird (0.6690) and buffalo (0.5669) were handled reasonably well, reinforcing that few-shot learning helped to improve the model's recognition capability of all classes.
- CLIP (zero-shot) achieved moderate recall across most species. It performed best for wildebeest (0.7273), gazelle (0.7051), and hyena (0.6018). However, it

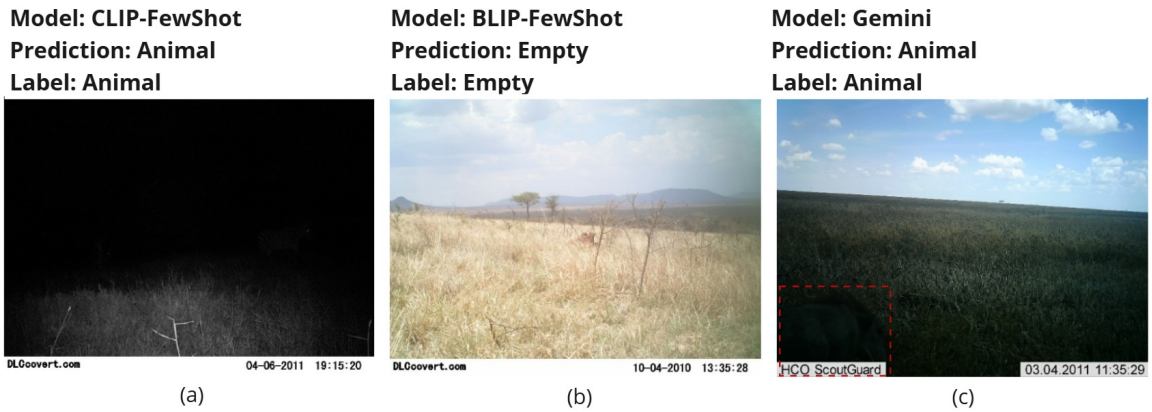

Figure 1. Examples of predictions for the filtering empty images task. In these cases, only one model produced the correct prediction, while the others misclassified the image: (a) Only CLIP-FewShot correctly identified the non-empty class; (b) Only BLIP-FewShot correctly identified the image as empty; (c) Only Gemini correctly assigned the non-empty class (the bounding box in the image is just to help see where the animal is, due to the conditions of some images).

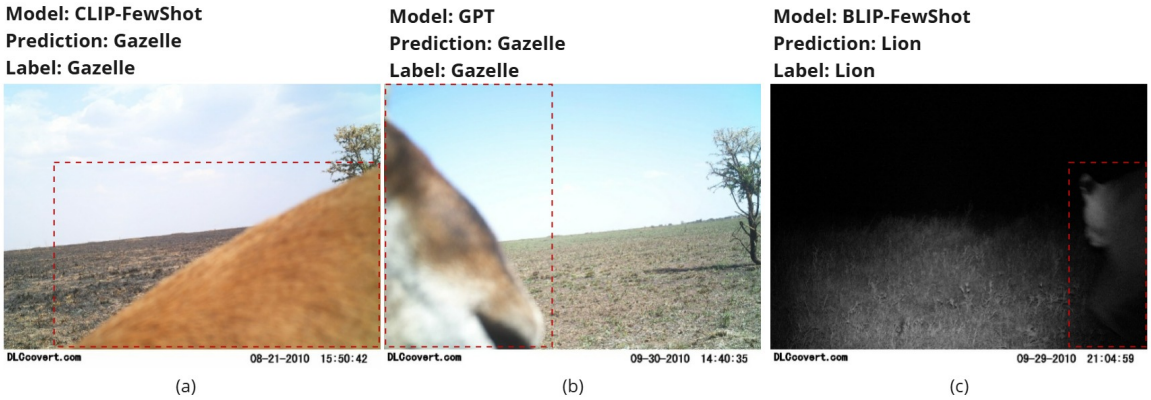

Figure 2. Examples of correctly species identification: (a) Only CLIP-FewShot correctly identified a gazelle; (b) Only GPT correctly identified a gazelle; (c) Only BLIP-FewShot correctly identified a lion (the bounding box in the image is just to help see where the animal is, due to the conditions of some images).

was notably weak on identifying bird (0.1038) and lion (0.1893).

- As a consequence of the reduced performance, the recall values of CLIP-FewShot dropped for almost every class, especially for gazelle (0.2502), zebra (0.2070), and giraffe (0.4760), indicating that the few-shot adaptation may not be optimal for this architecture.
- Gemini achieved highly balanced and robust recall across all species, with values ranging from 0.6123 (wildebeest) to 0.8821 (giraffe). Its consistent performance for elephant (0.8319), zebra (0.7795) and bird (0.6291) shows its potential for real-world multi-class retrieval tasks.
- GPT also performed well with high recall values, particularly for hyena (0.9442), elephant (0.7582), and giraffe (0.8368). While it showed slightly lower scores for bird (0.4037) and wildebeest (0.4991), in general its recall rates emphasize good generalization across species.

This evaluation of recall scores highlights the easiest species for classification. Giraffe was the species more often correctly classified across models, particularly BLIP-FewShot (0.8812), Gemini (0.8821), and GPT (0.8368). Models were also successful on classifying hyena, with GPT (0.9442), BLIP-FewShot (0.8575) and Gemini (0.7469) leading performance. The third more frequently well classified species is Elephant, with the highest scores from Gemini (0.8319), BLIP-FewShot (0.8058), and GPT (0.7582).

4.2.2 Discussion

In practical deployment scenarios, recall becomes a crucial metric — especially in safety-critical applications. In our context of biodiversity monitoring, missing a true positive (e.g., failing to detect an endangered species) can also carry significant consequences. Based on these findings, our results indicate the real-world potential of Gemini and GPT, since these models reached high recall, even in a zero-shot context, suggesting their robustness in species classification without retraining. BLIP's zero-shot performance was inadequate, reinforcing that it is an unreliable model unless few-shot learning is employed, in which case its recall improved dramatically. Conversely, CLIP's recall degradation in few-shot settings suggests challenges in adapting the model to small-scale training data, which could hinder its deployment in data-scarce environments.

4.3 Behavior Identification

The zero-shot BLIP model presented the same pattern observed in the two previous tasks, i.e. it reached the lowest performance, with accuracy of just 0.0011, precision of 0.0312, recall of 0.0002 , and F1-score of 0.0004. BLIP-Seq, which used sequences of images, reached low improvement, with accuracy of 0.0879 and F1-score of 0.0361, suggesting minimal benefit from sequence data when fine-tuning is not performed. On the other hand, again, BLIP's performance

Table 4. Recall values for the species identification task.

Model	Giraffe	Lion	Bird	Zebra	Hyena	Buffalo	Wildebeest	Gazelle	Elephant
BLIP	0.0018	0.0027	0.0000	0.0009	0.0000	0.0000	0.0000	0.0000	0.0000
BLIP-FewShot	0.8812	0.8000	0.6690	0.6535	0.8575	0.5669	0.7460	0.7991	0.8058
CLIP	0.5249	0.1893	0.1038	0.4797	0.6018	0.2518	0.7273	0.7051	0.5378
CLIP-FewShot	0.4760	0.3547	0.1233	0.2070	0.5283	0.2579	0.3556	0.2502	0.3277
Gemini	**0.8821**	**0.8107**	0.6291	**0.7795**	0.7469	**0.7685**	**0.6123**	**0.7758**	**0.8319**
GPT	0.8368	0.5289	**0.9442**	0.7525	**0.9442**	0.7130	0.4991	0.7535	0.7582

improved significantly with BLIP-FewShot, which led to increasing accuracy to 0.7014, precision to 0.7041, recall to 0.7011, and F1-score to 0.7011. A higher performance improvement was observed with BLIP-FewShot-Seq, whose accuracy increased to 0.7557 and F1-score to 0.7567, indicating that temporal context can aid behavior classification in a few-shot scenario.

Even though CLIP in zero-shot provided higher rates compared to BLIP, it was not a high-performing method: accuracy of 0.3425 and F1-score of 0.3321. Moreover, CLIP-Seq tested in zero-shot settings with sequences performed similarly to its single-image version, achieving accuracy of 0.3506 and F1-score of 0.3397. Here, however, unlike in previous tasks, few-shot improved CLIP's performance to accuracy of 0.4817 and F1-score of 0.4766. The same improvement is observed for CLIP-FewShot-Seq, whose accuracy improved to 0.5006 and F1-score to 0.5016, highlighting the added value of temporal information when few-shot is employed.

In its turn, Gemini performed reasonable, considering the zero-shot learning scenario, with accuracy of 0.5766, precision of.4533, recall of 0.4312, and F1-score of 0.3940. When sequences of images were introduced in Gemini-Seq, its performance improved to accuracy of 0.6171 and F1-score of 0.4336, indicating that sequential data also helped Gemini to better conduct behavior recognition. An opposite behavior was shown by GPT. It achieved accuracy of 0.4819, precision of 0.6339, recall of 0.4813, and F1-score of 0.4195. However, when sequences were incorporated in GPT-Seq, its accuracy dropped to 0.4058 and its F1-score to 0.3351. This suggests that the temporal context does not universally enhance model performance and may even reduce its performance in some cases.

4.3.1 Error distribution across classes

It is also important to consider the ability of the models to accurately identify animal behaviors, especially in ambiguous or challenging scenarios. Figure 3 illustrates examples in which only one model was able to assign the correct label, reinforcing the prediction variance of the different models. In Figure 3(a), BLIP-FewShot-Seq was the only model to correctly identify the resting behavior, maybe due to its attention to subtle posture cues. Conversely, CLIP-FewShot-Seq accurately detected the moving behavior in Figure 3(b), suggesting robustness to motion blur or posture ambiguity. In Figures 3(c) and 3(d), only GPT-Seq and Gemini-Seq, respectively, succeeded in classifying the behavior as moving, even when the movement was partially occluded or visually subtle. Lastly, Figure 3(e) shows a case where CLIP-Seq was the sole model to correctly identify the resting behavior. These examples emphasize the complementary strengths of the models in complex behavior recognition scenarios.

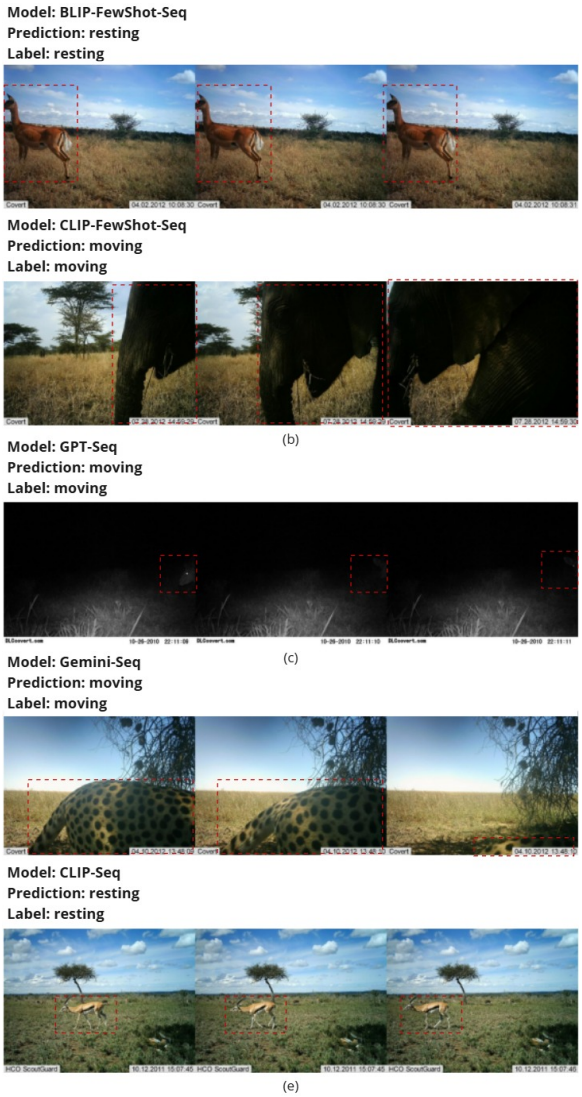

Figure 3. Examples of the behavior identification task where only one model produced the correct prediction while all others were incorrect: (a) model BLIP-FewShot-Seq and class resting; (b) model CLIPFewShot-Seq and class moving; (c) model GPT-Seq and class moving; (d) model Gemini-Seq and class moving; (e) model CLIP-Seq and class resting (the bounding box in the image is just to help see where the animal is, due to the conditions of some images).

Examining the class-wise recall scores in Table 5, it becomes clear that the resting class posed significant challenges for most of the models. While BLIP-FewShot-Seq achieved the highest recall scores for moving (0.7392) and eating (0.8040), its performance for resting is slightly lower (0.7232). Similarly, GPT showed high recall for moving (0.9026) but

has a considerably lower recall for resting (0.0731). This pattern is also detected for CLIP and BLIP, whose recall for the resting class is notably weak—often close to zero (0.3497 and 0.0000 respectively). Gemini and its sequence variant, despite demonstrating top-tier recall for eating (0.8252 and 0.8629) and moving (0.7464 and 0.7596), also had difficulty in recognizing the resting behavior (0.1532 and 0.2239 respectively). These differences highlight that distinguishing the resting class remains challenging for all evaluated models.

Table 5. Recall values for the behavior identification task.

Model	Resting	Eating	Moving
BLIP	0.0000	0.0000	0.0033
BLIP-FewShot	0.6706	0.7707	0.6620
BLIP-FewShot-Seq	**0.7232**	0.8040	0.7392
BLIP-Seq	0.0000	0.2614	0.0000
CLIP	0.3497	0.1909	0.4886
CLIP-FewShot	0.5748	0.3289	0.5436
CLIP-FewShot-Seq	0.4584	0.4442	0.5995
CLIP-Seq	0.3500	0.1974	0.5060
Gemini	0.1532	0.8252	0.7464
Gemini-Seq	0.2239	**0.8629**	0.7596
GPT	0.0731	0.4680	**0.9026**
GPT-Seq	0.0496	0.3265	0.8398

4.3.2 Discussion

Overall, GPT and Gemini performed well, particularly when identifying the moving behavior with recall values as high as 0.9026 and 0.7596, respectively. BLIP and CLIP demonstrated notable gains in few-shot learning scenarios, especially when image sequences were incorporated. Among all configurations, BLIP-FewShot-Seq exhibited the most balanced recall across behaviors, notably achieving 0.7232 for resting, 0.8040 for eating, 0.7392 for moving and the highest performance in general. These results underscore the value of combining few-shot learning with sequential inputs to enhance model sensitivity, particularly for difficult classes such as resting.

For real-world contexts, these results indicate that this combination may be incorporated in a practical solution for behavior classification. However, the high performance difference between methods and classes indicates that errors remain a challenge, and model reliability can vary depending on the behavior being classified. The recall values, especially for eating, reinforce that some misclassifications could be a serious problem in applications where high accuracy is essential.

4.4 Cost and Deployment Considerations

One very important point to consider is that, while our experiments demonstrate that MLLMs can automate key camera-trap tasks, continuous 24/7 monitoring raises questions related to computational cost. The models evaluated in this work require GPU inference (BLIP and CLIP), while Gemini and GPT demand additional API charges. Compared to conventional CNN pipelines that involve full retraining, our zero-/few-shot approach reduces training costs but still entails non-trivial inference expenses if every frame is processed centrally. Practical deployments may adopt a tiered pipeline—e.g., a lightweight detector to discard obvious empty frames before invoking an MLLM—or apply model-compression and distillation techniques to run on edge devices. Exploring these engineering trade-offs is an important direction for future work.

5 Conclusions

This study demonstrated the potential of using MLLMs in advancing biodiversity monitoring through the camera trap workflow. Key findings revealed that few-shot learning significantly improved model performance, particularly with BLIP, which reached 0.9100 accuracy in filtering empty images compared to just 0.0761 under zero-shot conditions. Gemini's stable zero-shot performance in species identification (0.7589 accuracy) and the significant enhancement in behavior classification through sequence-based processing (with BLIP reaching 0.7557 accuracy) further reinforce the capabilities of MLLMs when better contexts are provided to the models. These results highlight the importance of minimal supervision and temporal data integration in optimizing wildlife monitoring systems.

The implications of this research extend beyond technical performance metrics, directly addressing the research questions posed in this study. In response to RQ1, our results demonstrate that large-scale MLLMs can effectively enhance biodiversity monitoring tasks, particularly by automating species identification, behavior classification, and empty image filtering. This automation not only improves efficiency but also enhances the accuracy and consistency of biodiversity assessments, thereby supporting more informed conservation strategies. Moreover, the methodology employed in this study can be adapted to various ecological contexts, extending their applicability to other environmental monitoring systems.

Regarding RQ2, our results highlight key comparative advantages and limitations of zero-shot and few-shot learning approaches in biodiversity monitoring. While zero-shot learning proves effective for simpler tasks such as empty image filtering and species classification, it struggles with more complex analyses like behavior identification, emphasizing the need for adaptive learning strategies. Additionally, closed-source models like GPT and Gemini, which are not open for fine-tuning by the general public, exhibited inconsistent performance across different tasks, particularly when handling sequential data. Another challenge lies in the generalization of these models to diverse ecological settings, as performance may vary depending on species, environments, and camera trap conditions. Finally, a limitation of this work is that the Snapshot Serengeti dataset investigated in our experiments provides no labels for challenging scenarios beyond time-of-day, preventing us from reporting metrics related to factors such as silhouetting or animal distance.

Future research should focus on developing more adaptable and robust MLLMs capable of generalizing across diverse ecosystems without the need for retraining. Enhancing the models' ability to recognize rare or partially occluded species is another critical avenue for exploration. Hybrid learning strategies that combine zero-shot, few-shot, and supervised

learning may offer a balanced approach to optimizing model performance. Further investigation into the role of temporal context in various tasks could help refine model architectures, particularly to address inconsistencies observed in this study with sequence-based inputs. Finally, we also intend to explore whether more complex or descriptive prompts can further improve performance, particularly for larger models.

Declarations

Authors' Contributions

Luiz Alencar performed the experiments and data processing. Fagner Cunha, Luiz Alencar, and Eulanda Miranda contributed to the conception of this study. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Finance Code 001. This work was partially supported by Amazonas State Research Support Foundation - FAPEAM - through the POSGRAD project 2024/2025.

Availability of data and materials

The materials (models, data, results) used for this study are available at: https://github.com/LuizAlencar17/multimodal-models-in-camera-traps

References

Alencar, L., Cunha, F., and dos Santos, E. M. (2023). A context-aware approach for filtering empty images in camera trap data using siamese network. In *2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 85–90. IEEE. DOI: 10.1109/sibgrapi59091.2023.10347159.

Alencar, L., Cunha, F., and Dos Santos, E. M. (2024). Zero and few-shot learning with modern mllms to filter empty images in camera trap data. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. IEEE. DOI: 10.1109/sibgrapi62404.2024.10716305.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473. DOI: 10.1007/978-3-030-01270-0$_2$8.

Binta Islam, S., Valles, D., Hibbitts, T. J., Ryberg, W. A., Walkup, D. K., and Forstner, M. R. (2023). Animal species recognition with deep convolutional neural networks from ecological camera trap images. *Animals*, 13(9):1526. DOI: 10.3390/ani13091526.

Choiński, M., Rogowski, M., Tynecki, P., Kuijper, D. P., Churski, M., and Bubnicki, J. W. (2021). A first step towards automated species recognition from camera trap images of mammals using ai in a european temperate forest. In *Computer Information Systems and Industrial Management: 20th International Conference, CISIM 2021, Ełk, Poland, September 24–26, 2021, Proceedings 20*, pages 299–310. Springer. DOI: 10.1007/978-3-030-84340-3$_2$4.

Cunha, F., dos Santos, E. M., Barreto, R., and Colonna, J. G. (2021). Filtering empty camera trap images in embedded systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2438–2446. DOI: 10.1109/cvprw53098.2021.00276.

Cunha, F., dos Santos, E. M., and Colonna, J. G. (2023). Bag of tricks for long-tail visual recognition of animal species in camera-trap images. *Ecological Informatics*, 76:102060. DOI: 10.1016/j.ecoinf.2023.102060.

Dorm, F., Millard, J., Purves, D., Harfoot, M., and Mac Aodha, O. (2025). Large language models possess some ecological knowledge, but how much? *bioRxiv*, pages 2025–02. DOI: 10.1016/j.ecoinf.2026.103699.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., *et al*. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*. DOI: 10.48550/arxiv.2303.03378.

Dussert, G., Miele, V., Van Reeth, C., Delestrade, A., Dray, S., and Chamaille-Jammes, S. (2024). Zero-shot animal behavior classification with image-text foundation models. *bioRxiv*, pages 2024–04. DOI: 10.1101/2024.04.05.588078.

Fabian, Z., Miao, Z., Li, C., Zhang, Y., Liu, Z., Hernandez, A., Arbelaez, P., Link, A., Montes-Rojas, A., Escucha, R., *et al*. (2023). Knowledge augmented instruction tuning for zero-shot animal species recognition. *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*. Available at:`https://openreview.net/forum?id=OQHckRYbpT`.

Fennell, M., Beirne, C., and Burton, A. C. (2022). Use of object detection in camera trap image identification: Assessing a method to rapidly and accurately classify human and animal detections for research and application in recreation ecology. *Global Ecology and Conservation*, 35:e02104. DOI: 10.1016/j.gecco.2022.e02104.

Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N. A., Ma, W.-C., and Krishna, R. (2024). Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer. DOI: 10.1007/978-3-031-73337-6$_9$.

Gabeff, V., Rußwurm, M., Tuia, D., and Mathis, A. (2024). Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *International Journal of Computer Vision*, 132(9):3770–3786. DOI: 10.1007/s11263-024-02026-6.

Guo, C., Miguel, A., and Maciejewski, A. A. (2024). Automatic identification of individual african leopards in unlabeled camera trap images. *IEEE Transactions on Automation Science and Engineering*. DOI: 10.1109/tase.2024.3379553.

Iannarilli, F., Erb, J., Arnold, T. W., and Fieberg, J. R. (2021). Evaluating species-specific responses to camera-trap survey designs. *Wildlife Biology*, 2021(1):1–12. DOI: 10.2981/wlb.00726.

Islam, R. and Moushi, O. M. (2024). Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*. DOI: 10.36227/techrxiv.171986596.65533294/v1.

Koh, J. Y., Fried, D., and Salakhutdinov, R. R. (2024). Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36. DOI: 10.52202/075280-0939.

Leorna, S. and Brinkman, T. (2022). Human vs. machine: Detecting wildlife in camera trap images. *Ecological Informatics*, 72:101876. DOI: 10.1016/j.ecoinf.2022.101876.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR. DOI: 10.48550/arXiv.2201.12086.

Liu, H., Li, C., Li, Y., and Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306. DOI: 10.1109/cvpr52733.2024.02484.

Ma, Y., Cao, Y., Sun, J., Pavone, M., and Xiao, C. (2024). Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pages 403–420. Springer. DOI: 10.1007/978-3-031-72995-9$_2$3.

Muhtar, D., Li, Z., Gu, F., Zhang, X., and Xiao, P. (2024). Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer. DOI: 10.1007/978-3-031-72904-1$_2$6.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725. DOI: 10.1073/pnas.1719367115.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al*. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR. DOI: 10.48550/arxiv.2103.00020.

Santamaria, J. D., Isaza, C., and Giraldo, J. H. (2024). Catalog: A camera trap language-guided contrastive learning model. *arXiv preprint arXiv:2412.10624*. DOI: 10.1109/wacv61041.2025.00124.

Schneider, S., Greenberg, S., Taylor, G. W., and Kremer, S. C. (2020). Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and evolution*, 10(7):3503–3517. DOI: 10.1002/ece3.6147.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., and Packer, C. (2015a). Data from: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*. DOI: doi:10.5061/dryad.5pt92.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., and Packer, C. (2015b). Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1):1–14. DOI: 10.1038/sdata.2015.26.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., *et al*. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590. DOI: 10.1111/2041-210x.13120.

Tan, M., Chao, W., Cheng, J.-K., Zhou, M., Ma, Y., Jiang, X., Ge, J., Yu, L., and Feng, L. (2022). Animal detection and classification from camera trap images using different mainstream object detection architectures. *Animals*, 12(15):1976. DOI: 10.3390/ani12151976.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., *et al*. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. DOI: 10.48550/arXiv.2312.11805.

Vecvanags, A., Aktas, K., Pavlovs, I., Avots, E., Filipovs, J., Brauns, A., Done, G., Jakovels, D., and Anbarjafari, G. (2022). Ungulate detection and species classification from camera trap images using retinanet and faster r-cnn. *Entropy*, 24(3):353. DOI: 10.3390/e24030353.

Vélez, J., McShea, W., Shamon, H., Castiblanco-Camacho, P. J., Tabak, M. A., Chalmers, C., Fergus, P., and Fieberg, J. (2023). An evaluation of platforms for processing camera-trap data using artificial intelligence. *Methods in Ecology and Evolution*, 14(2):459–477. DOI: 10.1111/2041-210X.14044.

Vyskočil, J. and Picek, L. (2024). Towards zero-shot camera trap image categorization. *arXiv preprint arXiv:2410.12769*. DOI: 10.48550/arXiv.2410.12769.

Wang, Z., Cai, S., Liu, A., Jin, Y., Hou, J., Zhang, B., Lin, H., He, Z., Zheng, Z., Yang, Y., *et al*. (2024). Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: 10.48550/arxiv.2311.05997.

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., and Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1):80–91. DOI: 10.1111/2041-210x.13099.

Wu, J., Gan, W., Chen, Z., Wan, S., and Philip, S. Y. (2023). Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE. DOI: 10.1109/bigdata59044.2023.10386743.

Yang, D.-Q., Li, T., Liu, M.-T., Li, X.-W., and Chen, B.-H. (2021). A systematic study of the class imbalance problem: Automatically identifying empty camera trap images using convolutional neural networks. *Ecological Informatics*, 64:101350. DOI: 10.1016/j.ecoinf.2021.101350.