# Knowledge Augmented Instruction Tuning for Zero-shot Animal Species Recognition

**Zalan Fabian**[*†] **Zhongqi Miao**[*] **Chunyuan Li**[‡] **Yuanhan Zhang**[◦] **Ziwei Liu**[◦]

**Andrés Hernández**[*♦] **Andrés Montes-Rojas**[♦] **Rafael Escucha**[♦] **Laura Siabatto**[♦]

**Andrés Link**[♦] **Pablo Arbeláez**[♦] **Rahul Dodhia**[*] **Juan Lavista Ferres**[*]

## Abstract

Due to deteriorating environmental conditions and increasing human activity, conservation efforts directed towards wildlife is crucial. Motion-activated camera traps constitute an efficient tool for tracking and monitoring wildlife populations across the globe. Supervised learning techniques have been successfully deployed to analyze such imagery, however training such techniques requires annotations from experts. Reducing the reliance on costly labelled data therefore has immense potential in developing large-scale wildlife tracking solutions with markedly less human labor. In this work, we propose a novel zero-shot species classification framework that leverages multimodal foundation models. In particular, we instruction tune vision-language models to generate detailed visual descriptions of camera trap images using similar terminology to experts. Then, we match the generated caption to an external knowledge base of descriptions in order to determine the species in a zero-shot manner. We investigate techniques to build instruction tuning datasets for detailed animal description generation and propose a novel knowledge augmentation technique to enhance caption quality. We demonstrate the performance of our proposed method on a new camera trap dataset collected in the Magdalena Medio region of Colombia.

## 1 Introduction

Camera traps are motion-activated remote cameras that are used extensively to monitor wildlife populations. They are deployed worldwide for tasks such as density estimation of animal populations, species inventory and analysis of animal behavior [11, 16, 3]. Wildlife monitoring is more important than ever due to the devastating effects of increasing human activity and climate change on natural habitats. Camera traps offer a non-invasive and scalable solution, however the analysis of obtained imagery requires significant effort from wildlife experts [13].

Supervised machine learning approaches have been proposed and successfully deployed for wildlife detection and species classification in camera trap imagery [13, 17, 20, 12]. Even though these

[*]Microsoft AI for Good Lab, Redmond
[†]University of Southern California, Los Angeles
[‡]Microsoft Research, Redmond
[◦]Nanyang Technological University, Singapore
[♦]Universidad de Los Andes, Colombia
Corresponding authors: Zalan Fabian ( zfabian@usc.edu) and Zhongqi Miao (zhongqimiao@microsoft.com)

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

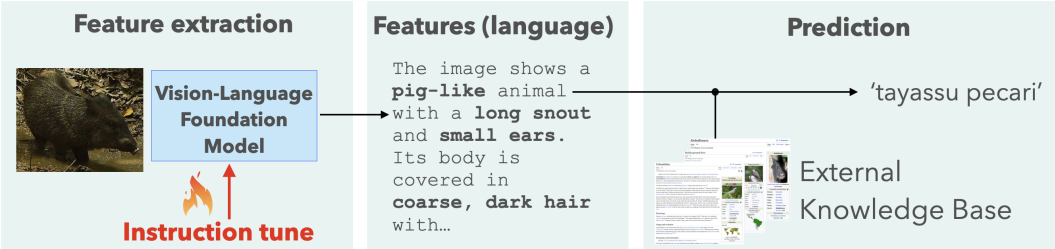


Figure 1: Overview of our proposed pipeline. We instruction tune a large multimodal model for detailed animal description generation. We extract visual features of camera trap imagery in the form of fine-grained captions. We leverage a knowledge base of animal descriptions to find the best match in order to identify the species.

techniques can help automate much of the visual recognition pipeline, they suffer from the well-known shortcomings of supervised techniques. First, massive amounts of annotated data is required to train the models. As the distribution of camera trap images exhibits strong domain variations (different environment, local species, camera setup), a new dataset needs to be collected for every region. Therefore supervised models are unfit for use-cases where such annotated dataset is not available yet. The problem of data collection is compounded by the fact that expert annotators with specialized knowledge of the local species are needed to label the images. Second, supervised models often lack robustness when deployed on data even slightly different from the training set [22]. Therefore, these models have difficulties when encountering low-quality and corrupted images, both very common in camera trap imagery due to motion blur, low-light conditions and occlusions.

Multimodal foundation models, such as Multimodal GPT-4 [14], Flamingo [2], LLaVA [10], InstructBLIP [4] and Otter [7] have been the driving force behind the revolution of artificial intelligence recently (see [8] for a comprehensive survey). Large multimodal models (LMMs) ground images to the natural language domain and demonstrate strong capabilities in image understanding and reasoning. Multimodal models such as CLIP [15] have proven to have strong zero-shot classification performance and are able to generalize to novel concepts and categories not directly seen in the training set. Moreover, recent work [6] serves as a strong indicator that in-context language descriptions can guide LMMs to differentiate between fine-grained categories, a crucial requirement in animal species classification.

In this work, we propose a pipeline for fine-grained zero-shot classification for animal species classification leveraging vision-language foundation models (see overview in Figure 1). We extract visual features in the natural language domain from camera trap images in the form of detailed image descriptions. Then, we compare the extracted features to a pre-compiled external knowledge base and output the category with the closest match, a technique we call *description matching*. We observe that out-of-the-box LMMs are incapable of generating detailed enough image descriptions and the captions are often ridden by irrelevant information and hallucinations. To tackle this, we propose an instruction tuning pipeline for detailed animal description generation by injecting common-sense and expert knowledge into vanilla LMM-generated image descriptions. We demonstrate the performance of our pipeline on a new camera trap dataset collected from a novel region in Colombia. Our method shows promising performance without the need for any in-domain training data.

## 2 Method

In this work, we propose a zero-shot animal species classification pipeline that leverages natural language descriptions rather than vision representations directly. The key component of traditional supervised vision models is a feature extractor that learns the relevant visual features in images from a training set. The extracted vision feature vectors are then passed through a classifier head that maps the vision features to discrete class labels. In stark contrast, our extracted features are *in natural language* in the form of image captions obtained from a LMM. The extracted natural language description is then compared with descriptions in an external knowledge base (summarized from online sources) and the entry with the closest match is selected as our final prediction. We introduce our matching technique in Section 2.4. The efficiency of the proposed pipeline therefore

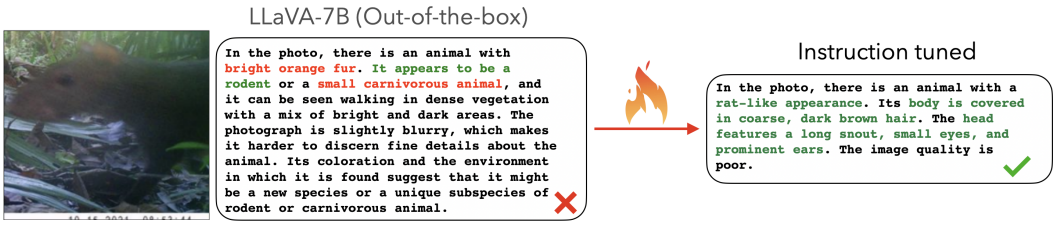


Figure 2: The general-domain LMMs trained on web data are unable to generate captions suitable for species identification out-of-the-box: the captions often miss relevant details, add excessive irrelevant information and often contain hallucinations. Domain-specific instruction tuned LMMs generate more relevant and succinct captions.

depends on the two key components of our framework: 1) the quality of the knowledge base and 2) the quality of captions obtained from the LMM.

### 2.1 Building a knowledge base

We obtain a list of $668$ species appearing in the publicly available LILA BC Camera Traps [1], a collection of $18$ camera trap datasets with species annotations. We scrape the Wikipedia article of each species and summarize the appearance and visual details of the animals using GPT-4. Details on prompts for summarization and text processing can be found in Appendix A.

### 2.2 Shortcomings of out-of-the-box LMMs

The performance of our pipeline strongly depends on the quality of image captions obtained from the LMM. However, we observe that most currently available LMMs (e.g. LLaVA, InstructBLIP or Otter) out-of-the-box are not suitable for extracting relevant details for species identification (Figure 2, left). First, LMM captions often miss relevant and characteristic features of animals that are necessary for correctly identifying the species. Second, out-of-the-box LMMs tend to generate excessive irrelevant information (e.g. speculations about the image or comments on the beauty of the scene) that has no use to or even mislead the matching algorithm. Lastly, hallucinations are very common in LMM captions [9], such as hallucinated colors, body parts, other animals or entities appearing in the image.

LMMs have been pretrained on massive amounts of multimodal data scraped from the internet, and as such they are imbued with a rich understanding of visual concepts. However, as these models have been mostly trained on generic 'internet data', they are not suitable for the specific task of generating detailed descriptions of animals with the focus on species identification. Therefore, we propose adapting LMMs for detailed animal description generation via instruction tuning. Our goal with instruction tuning is to guide the model to generate captions that leverage language used by experts in describing animals.

### 2.3 Instruction tuning for detailed visual descriptions

In order to instruction tune the model, we need paired data of wildlife imagery with detailed enough captions for species identification. Even though some datasets exist with species annotations (ImageNet [5], iNaturalist [18]), bounding boxes (most LILA Camera Traps [1]) or even short captions (Caltech-UCSD Birds [19]), there is no available large-scale dataset with detailed visual descriptions of animals. In fact, collecting such a dataset requires significant effort from experts far exceeding simple species annotations. We tackle the challenge of instruction tuning data collection from two directions.

First, we collect a small dataset of human-captioned wildlife imagery. In order to circumvent the need for biologists for annotations, we extract visual features of species from their corresponding Wikipedia articles and ask human annotators to select the features that are visible in the image. We collect approximately $1.5k$ manually annotated samples from volunteers. Human-annotated data collection is detailed in Appendix F.

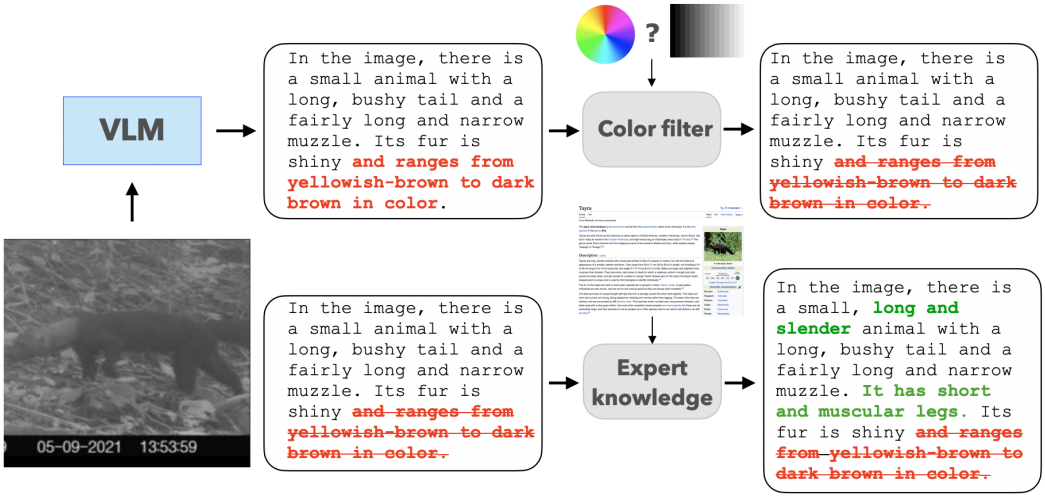


Figure 3: To improve caption quality of LMMs on wildlife images, we perform two processing steps on the vanilla LMM captions: 1) we filter out color related information if we detect that colors are not discernible in the image and 2) we inject expert knowledge from the species' Wikipedia article pertaining to visual characteristics of the animal that might be visible in the image. We leverage the resulting augmented captions to instruction tune a LMM for detailed animal description generation.

Second, we propose a scalable and automatic method of generating captions for species identification via augmenting vanilla out-of-the-box LMM captions with common-sense and expert knowledge. In particular, we process the vanilla LMM captions the following ways:

- **Color filtering –** LMMs often hallucinate colors even on grayscale or night time imagery that may entirely derail species identification. To tackle this, we filter out any color related information if we detect that it should not be identifiable from the image.
- **Expert knowledge –** We inject information from the species' Wikipedia article pertaining to visual characteristics of the animal that might be visible in the image. Furthermore, we remove details that directly contradict the knowledge base. This step enhances the captions with relevant expert terminology used to describe animal species.

We perform these steps using GPT-4 prompting (details in Appendix B). We refer to the obtained descriptions as *pseudo-captions* (analogous to pseudo-labels in semi-supervised learning). In order to build the instruction tuning dataset, we generate single-turn conversations from the animal descriptions, where the instruction is sampled from various prompts asking to describe visual characteristics of the animal in the photo (details in Appendix C).

We apply an iterative instruction tuning scheme, where we train multiple generations of models. The first generation is instruction tuned on pseudo-captions generated by out-of-the-box LMMs, or the human-annotated samples. Subsequent generations are then instruction tuned on pseudo-captions from the previous generation's LMM captions.

### 2.4 Animal species classification via description matching

The key idea of our proposed pipeline is to caption camera trap images of animals via LMMs and subsequently find the closest matching species in a knowledge base. In particular, we sample a detailed description from our instruction tuned LMM and compare the generated caption to the description of each species in the knowledge base via a large language model (in this work GPT-4). The LLM is instructed to answer with the name of the species which best matches the LMM description. As we provide the knowledge base to the LLM before eliciting an answer, the matching algorithm can be thought of as an example of in-context learning. Details of the prompt used to instruct the LLM to perform the matching can be found in Appendix G.

| Model | Pseudo-captioning LMM | Instruction tuning data |
|---|---|---|
| *LLaVA-7B-M-gen1* | - | Manual-1.5k |
| *LLaVA-7B-M-gen2* | LLaVA-7B-M-gen1 | LILA-40k |
| *LLaVA-7B-P-gen1* | LLaVA-7B | LILA-10k |
| *LLaVA-7B-P-gen2* | LLaVA-7B-P-gen1 | LILA-40k |

Table 1: Summary of models we instruction tune for animal description generation.

| | Magdalena Camera Traps | |
|---|---|---|
| **Method** | Micro acc. | Macro acc. |
| Supervised | 93.94% | 85.50% |
| CLIP (category name) | 35.85 % | 33.13 % |
| CLIP (description) | 44.28 % | 36.16 % |
| LLaVA-7B | 54.96% | 50.76% |
| LLaVA-7B-M-gen1 | 62.00% | 59.28% |
| LLaVA-7B-M-gen2 | **70.12%** | **64.75%** |
| LLaVA-7B-P-gen1 | 65.23% | 61.57% |
| LLaVA-7B-P-gen2 | 69.94% | 64.36% |

Table 2: Results on the Magdalena Camera Traps dataset.

We further improve the performance of our method through a technique analogous to self-consistency in chain-of-thought reasoning [21]. Self-consistency appeals to the intuition that the correct answer can be obtained from multiple valid reasoning paths. In particular, we sample $N$ independent captions from the instruction tuned LMM and match each descriptions individually. We obtain the final prediction via majority voting between the $N$ resulting predictions.

## 3 Experiments

**Model –** We leverage the LLaVA framework for our LMM and instruction tune LLaVA-7B models on our animal description instruction tuning datasets.

**Instruction tuning data –** We create two splits from LILA BC Camera Traps datasets: 1) *LILA-10k* consists of image crops of the approx. $10k$ highest confidence bounding boxes from MegaDetector and all human-annotated bounding boxes where we limit the maximum number of crops per species to 25; 2) *LILA-40k* is created similarly but with maximum number of crops per species set to 100. Moreover, we create a human-annotated dataset of approx. $1.5k$ images, where we sample image crops of the 2 highest confidence bounding boxes for each species appearing in the LILA BC datasets, and also add the Wikipedia image from their corresponding article. We refer to this dataset as *Manual-1.5k*.

**Evaluation data –** We evaluate our technique's performance on a new camera trap dataset collected in the Magdalena Medio region in Colombia. As the dataset has not yet been publicly released, the LMM has not seen our validation dataset neither during pretraining nor during instruction tuning. The dataset consists of 41904 samples (33569 train, 8335 validation) with 36 discrete labels with genus level annotation. The image resolution is $256 \times 256$. For benchmarking, we sample 20 classes with imbalanced data distribution to simulate a realistic camera trap setting. We refer to this dataset as *Magdalena Camera Traps*. We use micro and macro classification accuracy as evaluation metrics, where micro accuracy refers to accuracy in the traditional sense and macro accuracy is the average of individual class accuracies across all classes.

**Instruction tuned models –** We investigate the value of both the number of samples used in instruction tuning as well as the quality of descriptions. To this end, we instruction tune various models on both pseudo-captioned and human-annotated data (see Table 1). We build a knowledge base of all species occuring in the LILA datasets for expert knowledge augmentation. We use $4\times$ A100 GPUs for instruction tuning.

**Inference setting –** We build separate knowledge bases for the Magdalena Camera Traps dataset, consisting of summaries of each genus appearing in the dataset. We leverage GPT-4 for description matching and we utilize self-consistency (SC) with majority voting to obtain the final predictions. We use $N = 5$ independent samples in the experiments unless mentioned otherwise.

**Main results –** An overview of our main results is depicted in Table 2. We train a ResNet-50 classifier on the full dataset to provide a supervised baseline. For a zero-shot baseline, we use CLIP (ViT L/14) where the text embeddings are either category names (genus in this case) or the description of the category from the knowledge base. We observe significant improvement over the naive CLIP baseline of close to $+25\%$ in a 20-way classification problem using our best instruction tuned model. Furthermore, generation 2 models consistently outperform their predecessors indicating the utility of additional instruction tuning data and improved data quality. However, we observe only a small gap

| | Magdalena Camera Traps | | Qualitative | |
|---|---|---|---|---|
| **Instruction tuning data** | Micro acc. | Macro acc. | Relevance (↑) | Hallucination (↑) |
| None | 54.96 % | 50.76 % | 3.98 (1.92) | 4.30 (1.83) |
| Pseudo-captions | 57.20 % | 57.10 % | 4.25 (1.89) | 4.41 (1.77) |
| Manual captions | **62.00%** | **59.28%** | **5.10** (1.98) | **5.33** (1.85) |

Table 3: Experiments on the effect of instruction tuning data quality on performance. We evaluate relevance and hallucination scores via GPT-4 prompting (higher the better). Standard deviation of scores in parentheses.

between generation 2 models that have been trained on either pseudo-captions or human annotations in the first generation.

**Impact of instruction tuning data quality –** As collecting human annotated wildlife imagery is costly, it is crucial to investigate how our automatic captioning technique with knowledge augmentation compares to human annotated captions. To this end, we create train and test splits from *Manual-1.5k* and instruction tune two models: one on human-annotated captions of the train split and a second on pseudo-captions of the same split. Pseudo-captions are generated via out-of-the-box LLaVA-7B and our knowledge augmentation method. We evaluate the generated captions quantitatively via classification accuracies when used in our pipeline, and qualitatively via GPT scoring. In particular, we prompt GPT-4 to score the generated caption based on how close they are to the corresponding human annotated ground truth caption. We define two scores: 1) relevance score, which measures how much of the information in the ground truth caption is captured in the generated caption and 2) hallucination score, which assigns higher scores to captions that have less excess details compared to the ground truth. More details and the specific prompts are in Appendix E. We evaluate these qualitative scores on the test split of *Manual-1.5k*. The results are summarized in Table 3. Overall, we observe consistent improvement in all metrics when using human-annotated data compared to automatically generated pseudo-captions for instruction tuning. Our experiments indicate that collecting large-scale human captioned wildlife image datasets may enable further improvements in the zero-shot performance of our proposed pipeline.

## 4 Conclusion

In this work, we have proposed a novel zero-shot classification framework for wildlife species recognition leveraging multimodal foundation models. We instruction tune vision-language models for detailed animal description generation and utilize a large language model to match the generated description to a pre-compiled external knowledge base. As datasets with detailed enough descriptions of animals are lacking, we propose a novel pipeline for leveraging a combination of human-annotated data and automatically generated (knowledge augmented) captions for instruction tuning. We evaluate the quality of generated descriptions both via their utility in our zero-shot classification framework, and via GPT scoring methods. Our instruction tuned models far surpass the performance of naive zero-shot baselines and out-of-the-box models in animal species classification. The main limitation of the current framework is its high compute requirement compared to supervised models, which we address in future work.

## References


[1] Labeled information library of alexandria: Biology and conservation. `https://lila.science/category/camera-traps/`. Category: Camera Traps.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[3] Anthony Caravaggi, Peter B Banks, A Cole Burton, Caroline MV Finlay, Peter M Haswell, Matt W Hayward, Marcus J Rowcliffe, and Mike D Wood. A review of camera trapping for conservation behaviour research. *Remote Sensing in Ecology and Conservation*, 3(3):109–122, 2017.

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

[7] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[8] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023.

[9] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[11] Tim CD Lucas, Elizabeth A Moorcroft, Robin Freeman, J Marcus Rowcliffe, and Kate E Jones. A generalised random encounter model for estimating animal density with remote sensor data. *Methods in Ecology and Evolution*, 6(5):500–509, 2015.

[12] Zhongqi Miao, Kaitlyn M Gaynor, Jiayun Wang, Ziwei Liu, Oliver Muellerklein, Mohammad Sadegh Norouzzadeh, Alex McInturff, Rauri CK Bowie, Ran Nathan, Stella X Yu, et al. Insights and approaches using deep learning to classify wildlife. *Scientific reports*, 9(1):8137, 2019.

[13] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. 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, 2018.

[14] R OpenAI. Gpt-4 technical report (arxiv: 2303.08774), 2023.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[16] Dede Aulia Rahman, Georges Gonzalez, and Stéphane Aulagnier. Population size, distribution and status of the remote and critically endangered bawean deer axis kuhlii. *Oryx*, 51(4):665–672, 2017.

[17] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

[18] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[19] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[20] Jana Wäldchen and Patrick Mäder. Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11):2216–2225, 2018.

[21] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[22] Sicheng Zhao, Bo Li, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020.

## A Wikipedia article summarization

We build the knowledge base from Wikipedia articles of species that are present in either of the LILA BC Camera Trap datasets or the Amazon dataset used for evaluation. We extract the page summary and any sections where the section title contains any of the following words: *description, characteristics, appearance, anatomy*. We use GPT-4 to summarize features of the animal that may be visible in a photograph, but without referring to specific physical measurements of the animal as those are typically not possible to determine from an image (exact height, weight etc). The following prompt is used for the summarization:

**System message:** `You are an AI assistant specialized in biology and providing accurate and detailed descriptions of animal species.`

**Prompt:** `You are given the description of an animal species. Provide a very detailed description of the appearance of the species and describe each body part of the animal in detail. Only include details that can be directly visible in a photograph of the animal. Only include information related to the appearance of the animal and nothing else. Make sure to only include information that is present in the species description and is certainly true for the given species. Do not include any information related to the sound or smell of the animal. Do not include any numerical information related to measurements in the text in units: m, cm, in, inches, ft, feet, km/h, kg, lb, lbs. Remove any special characters such as unicode tags from the text. Return the answer as a single paragraph. Species description: <WIKI_ARTICLE> Answer:`

## B Knowledge augmentation

We use GPT-4 to enhance the quality of LMM-generated captions with external knowledge and common sense.

**Color filtering –** We detect low color variation in images via the condition

$$\max_{i \in C} \left[\max\left(|R_i - G_i|, |R_i - B_i|, |B_i - G_i|\right)\right] < \epsilon, \tag{1}$$

where $C$ denotes pixels of a center crop of an image and $R_i$, $G_i$, $B_i$ denote the R, G, B channel values of pixel $i$ correspondingly. We perform center cropping in order to discard various external markings on the camera trap image (date, brand etc) that may be in color. We set $\epsilon = 10$. If low color variation is detected, we consider color related information in captions a hallucination. We use the following prompt to remove color information in this case:

**Prompt:** `This is the description of an animal in a photograph: <LMM_CAPTION>. Remove any mentions of color other than black or white. Answer:`

**Expert knowledge –** We enhance LMM captions with knowledge from our external knowledge base by adding details that may be visible in the image and removing information that directly contradicts the expert knowledge (extracted from Wikipedia articles). We use the following prompt for augmentation:

Figure 4: A sample conversation from our instruction tuning pipeline.

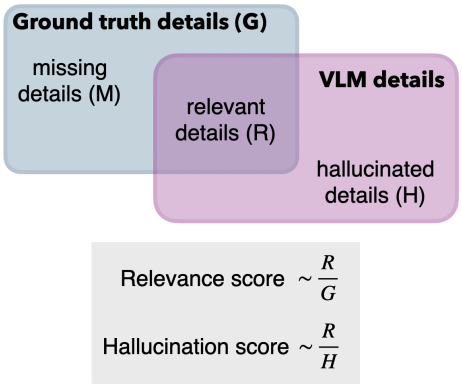


Figure 5: Overview of GPT scores used to evaluate caption quality.

**System message:** You are an AI assistant specialized in biology and providing accurate and detailed descriptions of animal species.

**Prompt:** This is an expert description of the appearance of an animal species: <EXPERT_DESCR>. This is an image description of the same species I can see in a photograph: <LMM_CAPTION>. Imagine that you can also see this photo and perform the following steps:
1. Rewrite the image description by adding details from the expert description of the species that are visible in the photo. Make sure you only add details about body parts of the animal already present in the image description.
2. Remove any information from the image description which directly contradicts the expert description.
3. Do not mention the species name in the description and do not try to guess the species.
Answer:

## C Instruction tuning data

We build instruction tuning datasets of single-turn conversations where the model is asked to describe the animal in the image in great detail. In particular, we randomly sample from the following instructions:

- Give a very detailed visual description of the animal in the photo.
- Describe in detail the visible body parts of the animal in the photo.
- What are the visual characteristics of the animal in the photo?
- Describe the appearance of the animal in the photo.
- What are the identifying characteristics of the animal visible in the photo?
- How would you describe the animal in the photo?
- What does the animal in the photo look like?

We randomly place the image either before or after the instruction. The expected response is either the human-annotated caption we collected (*Manual-1.5k*) or the automatically generated pseudo-caption corresponding to the image. A sample conversation is depicted in Figure 4.

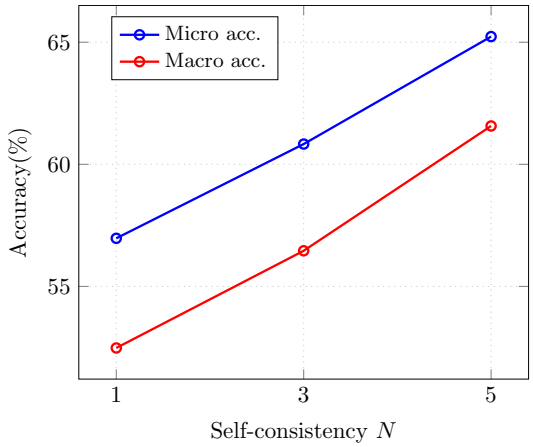


Figure 6: We observe consistent improvement as we increase the number of independent samples for inference (Amazon 20-class, LLaVA-7B-P-gen1).

## D Self-consistency ablations

We analyze the effect of the number of independent LMM captions drawn ($N$) for inference in our pipeline. Due to the stochasticity in LMM-generated text, independently sampled captions may contain complementary information. To investigate this effect, we run experiments on the Amazon 20-class benchmark with LLaVA-7B-P-gen1 and vary $N$ (Figure 6). We observe steady increase in classification accuracy in both metrics as $N$ increases, hinting at potential for further zero-shot performance improvement with large $N$. The current pipeline leverages LLMs for description matching during inference which results in high compute costs with increasing $N$. We investigate opportunities to reduce the cost and increase $N$ in future work.

## E GPT-4 scoring

We perform qualitative analysis of generated captions via GPT-4 scoring. Assume that the image contains a set of ground truth details $G$ that can be found in an image description provided by an expert. Our LMM generates an image caption that covers some details from $G$ and we call this intersection *relevant details* denoted by $R$. The rest of the details constitute the set of hallucinated details, denoted by $H$. An overview is depicted in Figure 5.

**Relevance score –** The relevance score evaluates the portion of ground truth details captured by the LMM caption. The more details are covered, the higher the score and no penalty is assigned for generating hallucinations. That is, the relevance score is proportional to $\frac{R}{G}$.The prompt used to evaluate the relevance score is as follows:

**Prompt:** `You are given two descriptions of an image: Description A and Description B. Description A is the correct and accurate description of the image. Your job is to score on a scale from 1 to 10 how well Description B describes the image. Follow these rules:`
`1. Only give the maximum score of 10, if Description B contains all the information in Description A.`
`2. Only give the score of 1 if Description B contains no information that is given in Description A.`
`3. Otherwise, assign scores from 2 to 9 to assess how much information from Description A is mentioned in Description B (the higher score the more information from Description A is present in Description B).`
`4. Disregard any information in Description B that is not mentioned in Description A in your scoring.`
`5. Your answer is a single score from 1 to 10 without accompanying`

```
explanation of the score.
Description A: <EXPERT_DESCR>
Description B: <LMM_CAPTION>
Your score:
```

**Hallucination score –** This score measures how much additional information is included in the generated caption that is not present in the ground truth description and thus it is likely hallucinated or irrelevant. A higher score is assigned when the ratio $\frac{R}{H}$ is higher. That is as opposed to the relevance score, the hallucination score penalizes generated content not in the ground truth description. We use the following prompt to evaluate the hallucination score:

**Prompt:** You are given two descriptions of an image: Description A and Description B. Description A is the correct and accurate description of the image. Definition of a hallucination: a hallucination is a detail in Description B that is not mentioned in Description A. Your job is to score on a scale from 1 to 10 how accurately Description B describes the image, assigning higher score to descriptions with less hallucinations. Follow these rules:

1. Only give the maximum score of 10, if Description B contains all information from Description A and Description B does not contain any hallucinations.
2. Only give the score of 1 if Description B contains no information that is given in Description A, but may contain any number of hallucinated details.
3. Otherwise, assign scores from 2 to 9 to assess how much hallucinated information is present in Description B: the higher the score the less hallucinations are present in Description B.
5. Your answer is a single score from 1 to 10 without accompanying explanation of the score.

Description A: <EXPERT_DESCR>
Description B: <LMM_CAPTION>
Your score:

## F Human-annotated data collection

In order to measure the quality of pseudo-captions, we also collect a small dataset of human-captioned wildlife images. The dataset contains 2 images of each species from the LILA camera traps and 1 corresponding 'clean' image from the species' Wikipedia article.

Instead of relying on expert annotators, we extract a list of visible features of each animal species in the dataset from Wikipedia using GPT-4. Then, the annotators are presented each of the features along with the image and are asked to select whether the feature is *fully visible*, *partially visible* or *not visible* in the image. Furthermore, we ask annotators whether colors are discernible in the image in order to filter out color information from captions if needed.

Finally, we use GPT-4 to combine the features annotated as visible for each image into a descriptive image caption, and we apply post-processing to remove color information if it shouldn't be discernible.

## G Description matching details

In our proposed classification framework, we leverage an LLM to match the generated image description to an entry in the knowledge base. The label of the matched entry is used as the predicted label. In particular, we use GPT-4 to perform the matching step with the following prompt:

**System message:** You are an AI expert in biology specialized in animal species identification.
**Prompt:** <KNOWLEDGE_BASE>
Question: You are given the following description of an animal: <LMM_CAPTION>. What is the most likely animal being described from the following list: <SPECIES_LIST>. Make sure your answer is a single word from the list <SPECIES_LIST>.
Answer:

In the above prompt, the knowledge base is given as a list of entries in the form of <SPECIES>:<DESCRIPTION>. We collect all <SPECIES> labels into a list <SPECIES_LIST> in order to encourage the model to select one of the valid categories.

**Hierarchical predictions –** The above method becomes infeasible with large knowledge bases and a large number of possible output labels. Therefore, instead of predicting the most fine-grained label directly (species or genus), we use hierarchical predictions to iteratively narrow down the possible fine-grained classes. As animal taxonomy is inherently hierarchical, species recognition lends itself to hierarchical classification. One may build separate knowledge bases for each taxonomic rank (class, order, family, genus, species) and perform the matching technique on a reduced set of categories. In this work, we use a two-step prediction scheme. First, we predict the *class* (in the sense of taxonomic rank) the animal belongs to, that is *Mammalia* (mammal) or *Aves* (bird). Then, in the fine-grained prediction step we only include categories that are under the super-class predicted in the previous step. A more extensive hierarchical prediction framework for species classification is going to be discussed in-depth in future work.