Received: 29 October 2024 | Accepted: 11 July 2025

DOI: 10.1111/2041-210x.70120

Methods in Ecology and Evolution

PERSPECTIVE

Conservation, Ecology and Artificial Intelligence: Advances and Symbiotic Solutions

New frontiers in artificial intelligence for biodiversity research and conservation with multimodal language models

Zhongqi Miao[1] | Yuanhan Zhang[2] | Zalan Fabian[1,3] | Andres Hernandez Celis[1,4] | Sara Beery[5] | Chunyuan Li[6] | Ziwei Liu[2] | Amrita Gupta[1] | Md Nasir[1] | Wanhua Li[7] | Jason Holmberg[8] | Meredith Palmer[9] | Kaitlyn Gaynor[10] | Pablo Arbelaez[4] | Pengce Wang[1] | Rahul Dodhia[1] | Juan Lavista Ferres[1]

[1]Microsoft AI for Good Lab, Redmond, Washington, USA; [2]Nanyang Technological University, Singapore; [3]University of Southern California, Los Angeles, California, USA; [4]Universidad De Los Andes, Bogotá, Colombia; [5]Massachusetts Institute of Technology, Boston, Massachusetts, USA; [6]Microsoft Research, Redmond, Washington, USA; [7]Harvard University, Cambridge, Massachusetts, USA; [8]Wild Me Labs, Conservation X Labs, Washington, District of Columbia, USA; [9]Yale University, New Haven, Connecticut, USA and [10]University of British Columbia, Vancouver, British Columbia, Canada

Correspondence
Zhongqi Miao
Email: zhongqimiao@microsoft.com

Handling Editor: Nicolas Lecomte

Abstract

1. The integration of artificial intelligence (AI) into biodiversity research and conservation is growing rapidly, demonstrating great potential in reducing the intensive human labour required for data preprocessing, thereby, facilitating larger data collections that offer ecological insights at unprecedented scales. However, most of these AI applications for biodiversity are still in the early stages of development, hindered by challenges inherent in real-world datasets and the limited accessibility of these technologies to practitioners without extensive programming knowledge.
2. The recent advent of multimodal language models, which can process and generate multiple data modalities, has significantly expanded the realm of possible AI applications in biodiversity research. These models have demonstrated the ability to classify species and recognize more complex concepts, such as animal postures and orientations, without prior exposure during training. Multimodal language models can also provide explanations for their predictions and interact with humans in natural language, thereby making them more transparent, intuitive and accessible to non-specialists. Despite these advancements, the use of multimodal language models for biodiversity still needs to overcome unique barriers to application, including high computational and financial demands, reliance on prompt engineering for consistent model performance on large datasets and insufficient open-source sharing of state-of-the-art methods.
3. This paper explores the transformative potential of multimodal language models for biodiversity research and discusses several possible applications in biodiversity research. We also discuss challenges to implement these models in real-world

Zhongqi Miao and Yuanhan Zhang contributed equally to this work.

conservation scenarios and propose directions for future research to overcome these hurdles.
4. Our goal is to encourage robust discussions and research into the integration of multimodal language models to advance AI for biodiversity research and conservation.

KEYWORDS
AI for biodiversity, AI for conservation, explainable AI, few-shot learning, generative AI, human machine interaction, multimodal language models, zero-shot learning

1 | INTRODUCTION

The field of artificial intelligence (AI) for biodiversity research and conservation is rapidly gaining traction within the ecological and biological sciences (Kwok, 2019; Tuia et al., 2022). An increasing body of research underscores the advantages of integrating AI techniques into biodiversity monitoring tasks, such as wildlife observation with automated animal recognition in both imagery/video (e.g. camera traps and aerial photos) (Ahumada et al., 2020; Beery et al., 2019; Kellenberger et al., 2021; Miao et al., 2021; Miao, Yu, et al., 2023) and audio data sources (e.g. bioacoustics) (Dodhia, 2024; Kahl et al., 2021; Rhinehart et al., 2020; Stowell et al., 2019; Zhong et al., 2021). These applications have demonstrated potential for reducing the substantial human labour traditionally required for data processing (Miao et al., 2021; Tuia et al., 2022), thus enabling the collection of more extensive datasets, providing ecological granularity at unprecedented spatial and temporal scales (Ahumada et al., 2020). This expansion paves the way for a more in-depth understanding of long-term patterns, drivers and consequences of global biodiversity changes.

While numerous efforts have been made to integrate AI into biodiversity data workflows, the majority remain in preliminary and proof-of-concept stages (i.e. unsuitable for practical implementation) due to various technical and data-related challenges. For example, model performance inconsistencies and the lack of reproducibility can be caused by serval factors such as severely imbalanced or long-tailed data distribution (Liu et al., 2019), differences in datasets and applications (i.e. multi-domain discrepancies) (Liu et al., 2020) and various issues arising from the complexity of open-world datasets (e.g. varying data quality and novel/unseen categories) (Miao et al., 2021). More importantly, the technical complexity of existing algorithms can often lead to inaccessibility for practitioners with limited programming and engineering knowledge.

The recent advent of multimodal language models—models that can process and generate both textual content and other data modalities (e.g. video and audio) (Alayrac et al., 2022; Gemini Team Google et al., 2023; Li, Zhang, Chen, Wang, Pu, et al., 2025; Li, Zhang, Chen, Wang, Yang, et al., 2023; Liu, Li, et al., 2023; OpenAI, 2023a; Radford et al., 2021)—has markedly enhanced the versatility and possibilities of AI applications (Li, Gan, Yang, Yang, Li, et al., 2023). This advancement has garnered considerable interest across disciplines—including the biodiversity and conservation community—as it overcomes many challenges that inhibit AI deployment into real-world applications. For instance, an off-the-shelf multimodal language model like GPT-4V (OpenAI, 2023a) can recognize animals without having seen them during training by, for example, using textual descriptions to infer their visual features (i.e. zero-shot transfer, Radford et al., 2021). Some models can even achieve supervised learning performance under zero-shot setting (Alayrac et al., 2022; Miao, Elizalde, et al., 2025). This holds promise for improving model robustness to variations in geographical location of the open dataset collection and distribution of species. Our experiments in this paper have also demonstrated GPT-4v's ability to distinguish more complex concepts, such as animal orientations and postures, without dedicated model fine-tuning on these tasks. Additionally, because multimodal language models closely integrate natural language processing with other data modalities (e.g. image and audio), these models can provide direct explanations for their predictions in natural language, enabling better explainable AI and allowing practitioners to better understand why and how these models make predictions. All of these capabilities are guided by human language inputs (i.e. text prompts). In other words, most interactions between humans and multimodal language models now become natural language-based, eliminating the need for complex computer programming procedures. This can significantly improve the accessibility of AI techniques for practitioners with limited engineering and programming experience.

Despite the promise of multimodal language models, their application faces unique challenges compared to traditional AI techniques. These include a significantly higher demand for computational and financial resources (Alayrac et al., 2022), a strong reliance on developing the correct text prompts (i.e. prompt engineering) for model performance (Zhou et al., 2022a, 2022b), limited open-source sharing of advanced multimodal language model methods (Alayrac et al., 2022; OpenAI et al., 2023; Radford et al., 2021) and a series of systematic failures of these methods (Thrush et al., 2022; Tong et al., 2023; Tong et al., 2024; Yuksekgonul et al., 2022), such as failing to differentiate sentences with quantifiers and numbers. Therefore, in this paper, we aim to explore the transformative impact multimodal language models can have on the future of biodiversity research and conservation and fully discuss the challenges of such novel techniques within

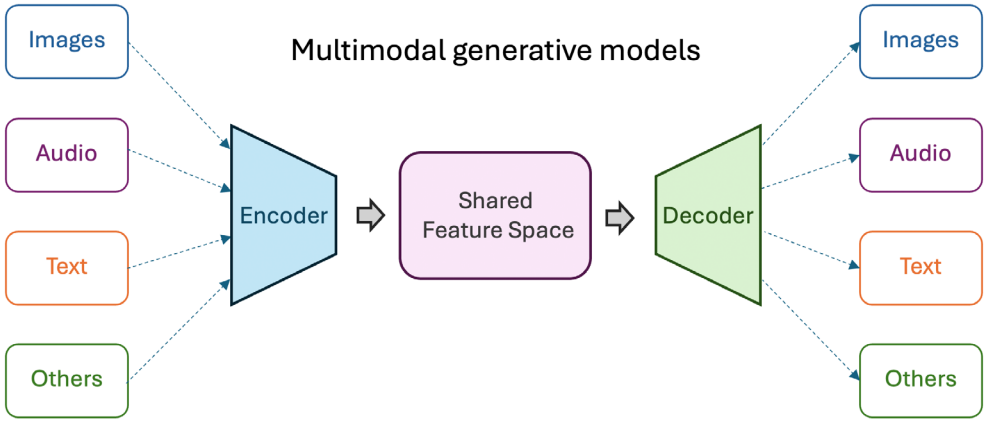


FIGURE 1 Illustration of multimodal generative models. In multimodal generative models, the encoders focus on constructing a shared feature space, aligning features from various input modalities such as images, audio, text and other modalities. Then the decoders utilize this shared feature space to generate any output modalities from any input modalities of data, such as creating images from a language description and vice versa.

these contexts. We begin with a general overview of the mechanisms of multimodal language models (Section 2) and explore how these mechanisms engender new applications of multimodal language techniques in biodiversity research and conservation (Sections 3 and 4). Then, we discuss the challenges and limitations we have identified for successfully implementing multimodal language models in real-world scenarios and propose potential future research directions to overcome these challenges (Section 5). Our objective is to foster robust discussion and research into the sustainable and equitable integration of multimodal language models, which could significantly advance the field of AI for biodiversity understanding and conservation.

2 | MULTIMODAL LANGUAGE MODELS

2.1 | Multimodal generative models

The exploration of multimodal models has gained considerable attention in recent years, especially for the line of work on multimodal generative models like GPT-4v (OpenAI, 2023a) and Stable Diffusion (Rombach et al., 2022). This is largely due to their unique capability to simultaneously process and generate any mixture of output modalities (i.e. data types) (Sun et al., 2023; Wu, Fei, et al., 2024; Yu et al., 2023) from various combinations of input modalities. For instance, just as the popular Stable Diffusion model (Rombach et al., 2022) can generate images from textual inputs, models like Flamingo (Alayrac et al., 2022) or GPT-4v (OpenAI, 2023a) can directly generate the textual output 'This is a flamingo' when presented with a picture of a flamingo bird and a corresponding natural language question (i.e. text prompt), such as 'What is this animal?' This capability relaxed the need for categories predefined or post-defined by humans as the model does not know what categories to focus on before inference; rather, it generates the predictions like the sentence 'This is a flamingo' directly, which increases the flexibility of the models for different use cases.

A multimodal generative models usually consists two major components: a multimodal encoder and a multimodal decoder (Figure 1). The encoder is responsible for encoding the input data into a shared feature (i.e. embedding space), where feature representations from semantically related data (e.g. images and audio clips from the same animal species) are aligned irrespective of their modalities. The decoder, on the other hand, generates output data in all sorts of modalities from the shared feature space. The training of multimodal generative models is usually based on large-scale datasets with paired data from different modalities, such as image–text pairs (Alayrac et al., 2022) and video–text pairs (Wu, Fei, et al., 2024).

2.2 | Multimodal language models

Among the many combinations of multiple modalities (Li et al., 2021; Lv et al., 2021; Stafylakis & Tzimiropoulos, 2017), multimodal language models have emerged as one of the most widely and actively researched areas in the field of multimodal models, spurred by advancements in single-modality large language models (LLMs) (OpenAI et al., 2023; Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023). For example, Vision–Language Models (VLMs) are multimodal models between vision and language modalities (Radford et al., 2021). The learning process of multimodal language models typically follows a contrastive learning protocol (Radford et al., 2021), which aims to maximize similarities between features of language concepts and features of perceptual data (i.e. feature alignment). Language concepts are pre-tokenized in the language encoders of the models (Vaswani et al., 2017)—breaking down text into smaller units called tokens, which can be words, subwords, characters or other meaningful segments—which are analogous to predefined labels or categories in conventional

categorical machine learning, but the scale of language tokens and their possible combinations is significantly larger. In other words, multimodal language models can be thought of as an extension of conventional categorical machine learning models, by aligning data (e.g. images and audio) to a vast array of words and language concepts (Li, Zhang, Chen, Wang, Yang, et al., 2023; Liu, Li, et al., 2023; Radford et al., 2021), instead of to a limited number of categories during training. To achieve this feature alignment, most multimodal language models utilize large-scale online datasets comprising perception-to-language pairs, such as image-to-language (Radford et al., 2021) or audio-to-language pairs (Elizalde et al., 2022). This pairing format ensures that each perceptual input is associated with unique language descriptions, providing comprehensive perception-to-language alignment. Figure 2b provides an example of how such image-to-language pairs may appear in the context of wildlife imagery.

Additionally, as the language concept features learned by LLMs usually encode extensive descriptive attributes and semantic relationships between words, the aligned perceptual features inherit these semantic relationships (Radford et al., 2021). This enables the recognition and prediction of subtle similarities and differences between complex concepts, such as implicit metaconcepts like animal families, shapes, orientations and postures (detailed in Section 3). For example, VLMs can position the features of a Koala and an Antechinus closer in the shared vision–language feature space, even before generalization based on visual similarities (Figure 2). This is because both words are implicitly connected to the meta-concept *Marsupial* in the language model, which typically lacks a strong semantic connection to the word *Mouse*. These semantic relationships, inherently encoded in VLMs through LLMs trained on large-scale online textual resources such as Wikipedia and Visipedia (Perona, 2010), serve as a form of supervision to regulate the shared feature space, provided these relationships are covered by the language resources (see Section 5.2 on how language sources can also limit model performance). Therefore, such supervision further enhances the semantic relevance between categories or animals.

3 | MULTIMODAL LANGUAGE MODELS AND ZERO-SHOT RECOGNITION

The feature alignment between data and vast amount of language concepts and semantics of multimodal language models introduces a considerable degree of flexibility to various AI tasks (Alayrac et al., 2022; Brown et al., 2020; Li et al., 2022; Misra et al., 2017; Wu, Fei, et al., 2024). Among these tasks, the capability to conduct zero-shot recognition—recognition of categories and concepts without specifically training on them—is particularly noteworthy.

Zero-shot recognition with multimodal language models—usually referred to as *Zero-shot transfer* (Radford et al., 2021) to differentiate from traditional zero-shot learning—depends largely on the alignment between perceptual information (e.g. visual features) and natural languages. It operates during inference by inferring visual features (or any other modality) from textual descriptions. For example, as shown in Figure 3, GPT-4v is able to associate visible morphological characteristics of animals with their textual descriptions (i.e. vision–language alignment), which can be used to differentiate similar-looking animals without dedicated training or fine-tuning on the task.

Figure 4a also shows an example of how VLMs work in zero-shot categorical recognition without another image to compare to. The language prompts included a context section where the VLM was instructed to emulate a professional microbiologist with experience in microscopic fungi identification. The model was then asked to describe the morphological characteristics from the input microscopy image. Subsequently, we prompted the model to generate a categorical prediction based on the visual descriptions. In our test, the model not only provided precise descriptions of the visual traits but also successfully classified the genus of the microscopic fungi.

These examples demonstrate not only the recognition potential of multimodal language models, but their generative capability to provide natural language explanations for a better interpretation of the results. This generative capability is further utilized in studies like (Fabian et al., 2023) to conduct zero-shot animal recognition without the need for human text inputs by matching generated descriptions of animal appearances from input images with online resources such as Wikipedia.

The alignment of perceptual features with language features and the flexibility of language features can further facilitate unprecedented zero-shot tasks, such as open-vocabulary segmentation and detection (Figure 4b) (Liang et al., 2023; Wu, Zhu, et al., 2023), a technique that allows a model to detect and segment objects in images or scenes using a flexible vocabulary that is not limited to a fixed set of categories. In addition, recognition tasks that go beyond rigid categorical recognition are also made possible because natural language is not confined to categorical concepts. Figure 4c illustrates the potential of VLMs in recognizing animal orientations, even when the animal is in a relatively complex posture, such as lying on the ground. This capability could be highly beneficial for downstream tasks such as animal re-identification (re-id) (Jiao et al., 2024), which heavily depends on accurately matching animal body markings to the correct sides of animals. More importantly, all these different language-based tasks can be realized with a single multimodal language model (GPT-4v in this case), instead of using independent models for each task as in traditional machine learning. Managing multiple models can become increasingly complex as their numbers grow and unrealistic to prepare sufficient training data for each one. In contrast, a unified model can leverage shared knowledge across tasks, potentially leading to improved performance. (Li, Zhang, Chen, Wang, Pu, et al., 2025).

However, as shown in both Figures 3 and 4, zero-shot recognition can heavily rely on text prompts and human inputs. In Sections 5.1 and 5.3, we discuss these limitations of the reliance on prompt inputs and other systematic failures that might occur in the applications of multimodal language models in detail.

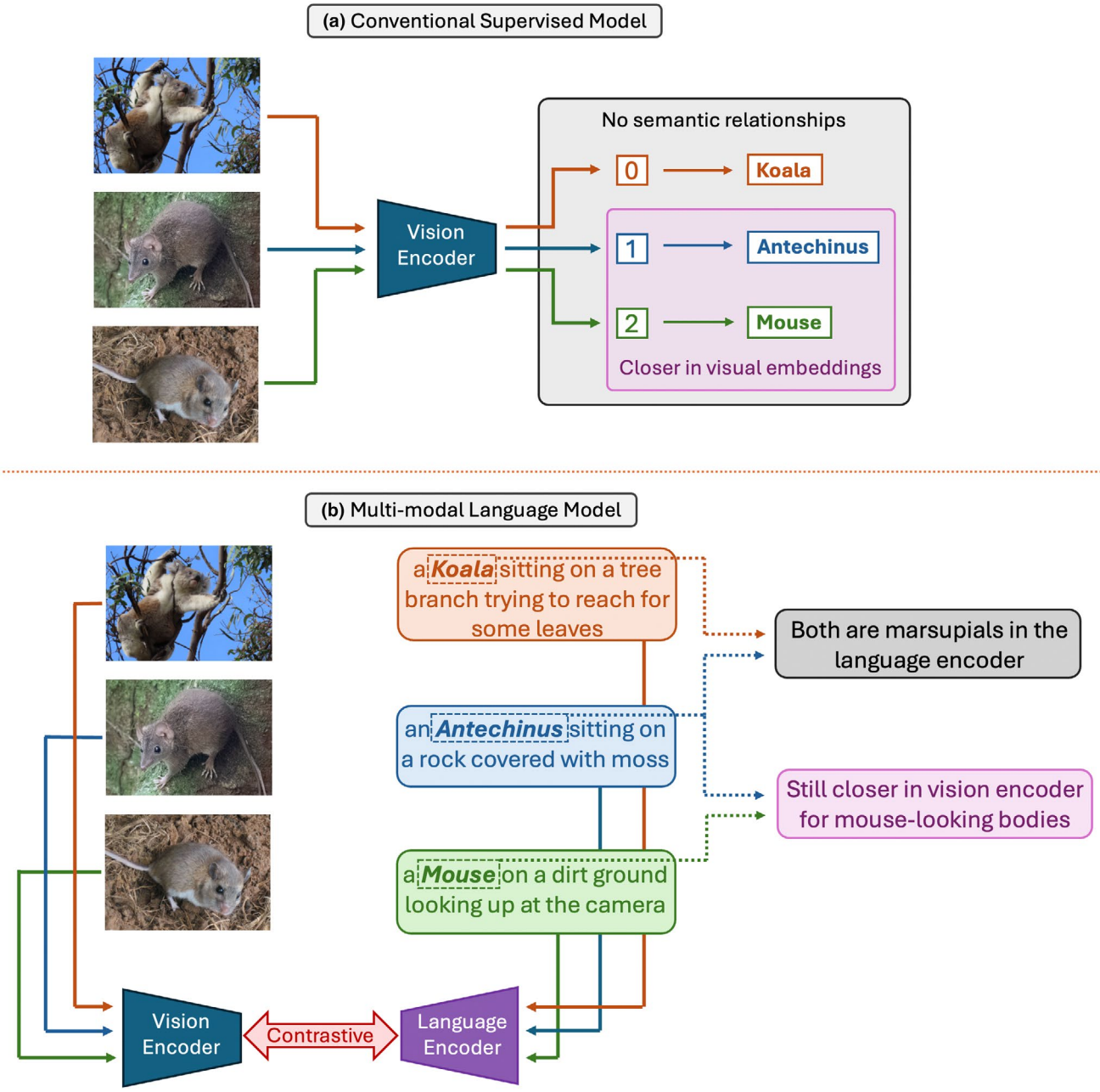


FIGURE 2 Illustration of categorical supervised learning (a) and multimodal language models (b). The training of multimodal language models is aimed at aligning other input modalities with the features of a vast array of language instead of to a relatively smaller number of categories. In addition, semantic relationships are naturally encoded and expressed in the shared feature space of multimodal language models. For instance, even though Koala and Antechinus look distinctly different from each other, these images can still have connections to each other because Koala and Antechinus are both Marsupials, and these two words have direct connections in the language feature space. Through the feature alignment process, the image feature structure will also follow the corresponding language feature structure. All images of animals are from iNaturalist (Van Horn et al., 2018).

4 | OTHER TASKS MADE POSSIBLE BY MULTIMODAL LANGUAGE MODELS

Beyond zero-shot recognition, multimodal language models also enable a range of application tasks that are relatively challenging for conventional machine learning techniques. In this section, we list some examples that have been made possible by the potential and flexibility of multimodal language models.

4.1 | Learning from very few samples

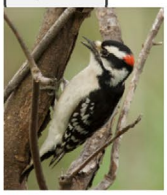
One of the tasks that multimodal language models have demonstrated particular success in is few shot learning—learning from very few (e.g. five or ten) training samples. This success is attributed to the surprising ability of LLMs to adapt to new tasks with high performance from few examples without extensive training or model fine-tuning (Tsimpoukelli et al., 2021). As presented in Flamingo (Alayrac

Context

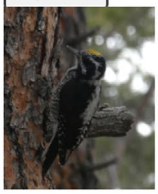
Assume you are not familiar with animals.

(a)

Input #1



Input #2



Input #3

Prompts

*There are three animal species, **Species 1**, **Species 2**, and **Species 3**.*

***Species 1** have black and white stripes throughout the body and yellow crowns.*

***Species 2** have a "ladder-like" black and white pattern in its back, and it has a red patch on their crowns.*

***Species 3** also some red on their crowns but covers a much smaller area on their head. They have white spots on their black wings.*

Simply by looking at the animal visible features and using the descriptions provided, which animal in the uploaded three images is a Species 1, Species 2, and Species 3? And explain why.

Predictions

Based on the provided descriptions and the visible features in the images:

The first image shows a bird with:
- *White spots on its black wings.*
- *A small red patch on its crown.*
- ***This matches the description of Species 3**, which has white spots on black wings and a small red patch on the crown.*

The second image shows a bird with:
- *Black and white stripes throughout its body.*
- *A yellow crown.*
- ***This matches the description of Species 1**, which has black and white stripes throughout the body and a yellow crown.*

The third image shows a bird with:
- *A "ladder-like" black and white pattern on its back.*
- *A red patch on its crown.*
- ***This matches the description of Species 2**, which has a ladder-like black and white pattern on its back and a red patch on the crown.*

Conclusion:
- ***First image: Species 3, because it has white spots on black wings and a small red patch on the crown.***
- ***Second image: Species 1, because it has black and white stripes and a yellow crown.***
- ***Third image: Species 2, because it has a ladder-like black and white pattern and a red patch on the crown.***

(b)

Input #1

Input #2

Input #3

Prompts

*There are three animal species, **Species 1**, **Species 2**, and **Species 3**.*

***Species 1** tend to have larger, more circular spots on their body with a dot in the center of these circular spots and have a stockier body shape.*

***Species 2** have an overall greyer body color, a white belly, and no central spots in the rosettes on their body.*

***Species 3** have smaller, more densely packed rosettes on their body without central spots.*

Simply by looking at the animal visible features and using the descriptions provided, which animal in the uploaded three images is a Species 1, Species 2, and Species 3? And explain why.

Predictions

Based on the provided descriptions and the visible features in the images:

The first image shows an animal with:
- *Larger, more circular spots with a dot in the center.*
- *A stockier body shape.*
- ***This matches the description of Species 1.***

The second image shows an animal with:
- *Smaller, more densely packed rosettes without central spots.*
- ***This matches the description of Species 3.***

The third image shows an animal with:
- *A greyer body color.*
- *A white belly.*
- *No central spots in the rosettes.*
- ***This matches the description of Species 2.***

Conclusion:
- ***First image: Species 1, because it has large circular spots with a central dot and a stockier body.***
- ***Second image: Species 3, because it has smaller, more densely packed rosettes without a central dot.***
- ***Third image: Species 2, because it has a greyer body, a white belly, and no central spots in the rosettes.***

FIGURE 3 Vision–Language Models understand morphological characteristics. We use GPT-4v to differentiate between two sets of animals by providing morphological descriptions to the model with a contextual prompt, 'Assume you are not familiar with animals'. We have also intentionally masked the animal names with Species 1, 2 and 3 to avoid potential model biases with animal species names. The model not only correctly differentiates these similar-looking animals based on the provided descriptions—implying its understanding of animal morphological characteristics—but also provides some reasoning for its predictions. All red text in (a, b) are real results generated by GPT-4v with the input prompts. All images are from the iNaturalist dataset (Van Horn et al., 2018).

et al., 2022), simply giving their model as few as four task-specific examples, the model is able to produce comparable if not superior performance than methods fine-tuned on thousands of examples from the same categories (as is shown in Alayrac et al., 2022). Few-shot learning is a task that is relevant to many AI for biodiversity and conservation application scenarios, such as endangered species monitoring (Sherman et al., 2020), where such tasks typically involve target categories/animals with limited available data. The advancement of few-shot learning with multimodal language models has the potential to improve the practical feasibility of these applications, but it has yet to be studied and examined in real-world.

4.2 | Generalization across varied data distributions and domains

Data distribution variation poses a major challenge in real-world applications of AI for biodiversity and conservation, particularly in animal recognition (Miao et al., 2021; Tuia et al., 2022). For example, models trained with conventional categorical supervised learning methods may not generalize well across different sites—even for the same animal species—due to regional variations in different environments, backgrounds, seasons, animal appearances (e.g. trait variation among subspecies) and setups of data collection devices (Miao et al., 2019; Miao et al., 2021). These differences in datasets are referred to as domain discrepancies (Kay et al., 2024).

Multimodal language models, on the other hand, often have a higher capacity for generalization across various data distribution/domains, primarily due to the alignment/similarity-based mechanism between perceptual and language concepts and the scale of training data. As mentioned in Section 2, the shared feature space of multimodal language models is based on how similar the input objects/concepts are to the existing language features in the feature space. This process does not need to be as precise as conventional categorical classification. In other words, any objects that look like a bird can be associated to the language concept 'bird', regardless of what environment these objects are in, thus generalizing across different data domains (Huh et al., 2024). Moreover, as most multimodal language models are trained on an extensive scale of generic online data—often magnitudes larger than the scale of training data for conventional, task-specific machine learning models—the feature space of these models is often robust enough to cover large variations of data as well (Radford et al., 2021). For example, in Miao, Elizalde, et al. (2025), the authors have demonstrated that the same audio-language model—trained on 2.1 million audio–text pairs from general purpose acoustic data—can generalize across eight different bioacoustics datasets (i.e. eight different data distributions that have substantially different sound attributes and qualities than the training data) recognizing the animal sounds from different datasets without dedicated model fine-tuning and still achieve supervised level performance. Despite the potential, the domain generalization capability of current multimodal language models is still limited by the scale of domain discrepancies—scale of differences between datasets collected from different domains (Li et al., 2024; Trinh et al., 2024). When the discrepancy is too large (e.g. the differences between general online imagery and real-world wildlife camera trap imagery), performance may be impaired. In Sections 5.2, 5.4, and 5.5, we review these limitations in detail.

4.3 | Enhanced model interpretability (explainable AI)

While traditional models function as 'black boxes', where researchers are unable to trace what features and mechanisms the models are using to make predictions, multimodal language models provide an unprecedented level of interpretability by the direct alignment between perceptual and language features in the shared feature space. Features extracted by conventional deep learning models are typically not interpretable by humans and therefore practitioners are often reluctant to trust the predictions obtained from such models, irrespective of the performance (Miao et al., 2019). In contrast, the shared perceptual-language feature space of multimodal models can provide venues for interpreting outputs directly in natural language and ultimately lead to a degree of insight into the inner workings of the model. For example, Figure 3 shows a model providing explanations on why it makes its classifications based on the input images and human text prompts. Similarly, the predictions in Figure 4a also provide reasoning about why the model thinks the input image is a *Aspergillus* species. In Figure 4c, the model even provides additional information on why it predicts the cheetah's left side is facing the camera, even though this is not a conventional recognition task. These insights can substantially increase the interpretability and explainability of AI techniques for practitioners.

Moreover, interpretability offers the added benefit of facilitating a better understanding of failure cases. In traditional categorical supervised learning with discrete class labels, prediction accuracy and similar metrics are the sole indicators used to evaluate model performance; this makes it challenging to anticipate the scenarios in which the model may fail. However, with natural language-based model interpretability, especially when the model explanations are wrong and irrelevant to input data, we can gain a clearer understanding of why models fail in certain cases—be it due to algorithmic failure or poor data quality (see Section 5.1; Figure 6 for more details).

4.4 | Learning with context

Multimodal language models have facilitated rapid progress in the field of compositional zero-shot learning (CZSL) (Misra et al., 2017)—a task that generalizes AI models to unseen compositions of perceptual attributes such as visual features/objects. For instance, CZSL models may be able to identify a *lying dog* after seeing a *lying horse* and a *dog* (Bao et al., 2023). CZSL can be further generalized to an area called *in-context learning*. In-context learning (Brown et al., 2020; Radford et al., 2019), originally introduced for language

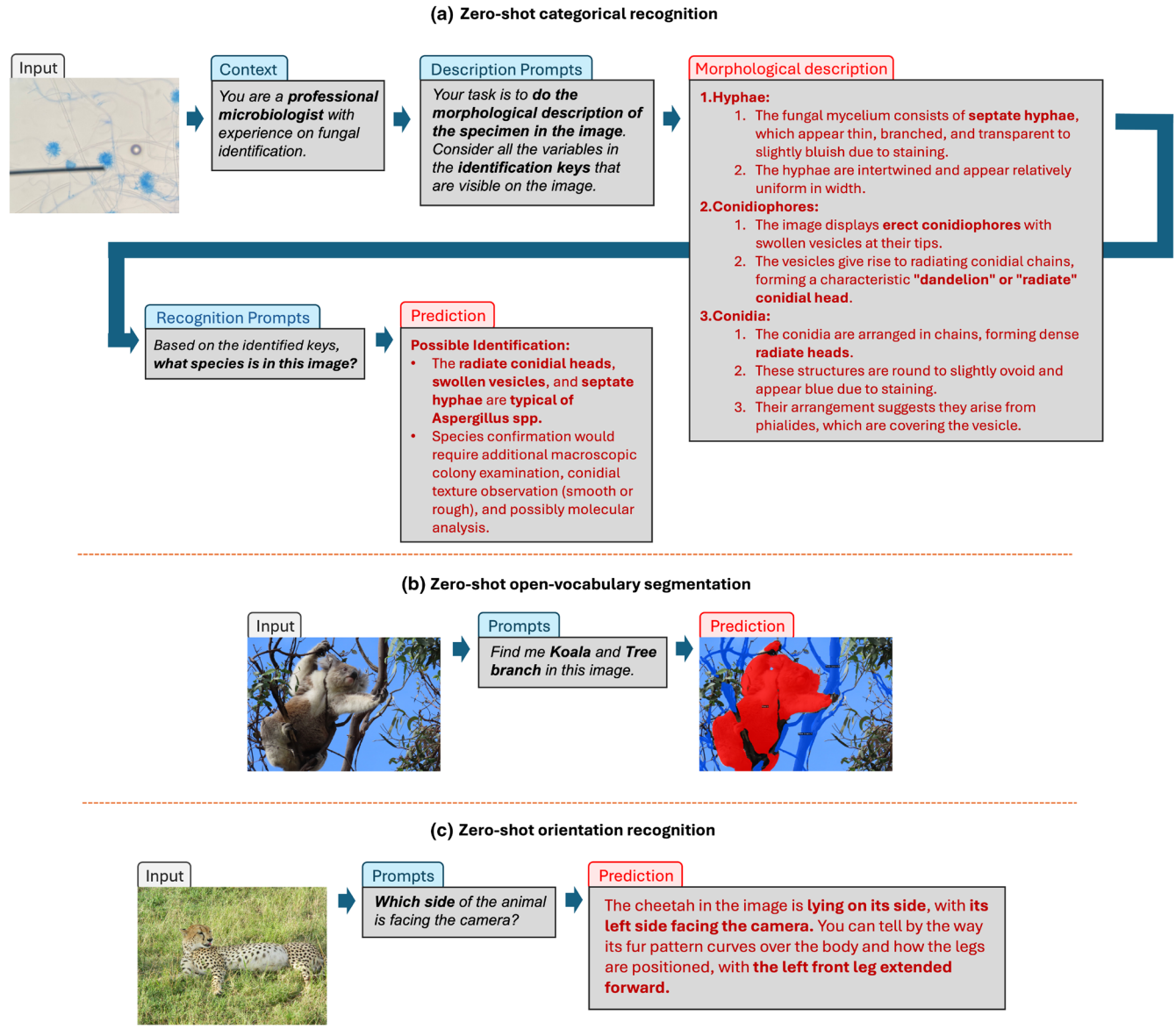


FIGURE 4 Zero-shot task examples with Vision–Language Models. (a) In mycology, mycologists typically utilize identification keys (i.e. a series of macroscopic and microscopic morphological descriptions formatted as dichotomous keys) to systematically conduct taxonomic classification (To-Anun et al., 2011). Multimodal language models can mimic these attribute-based classification processes and conduct zero-shot recognition without having previously seen such species (Menon & Vondrick, 2023). The recognition can be based on online information, such as that from Wikipedia, to which the multimodal language models have access (Fabian et al., 2023). (b) Multimodal recognition is not limited to full image and object recognition. Open-vocabulary segmentation and detection are direct extensions of the recognition capabilities of multimodal language models (Liang et al., 2023), in which categories are not predefined, and the number of categories is not fixed. Such methods can be further utilized for tasks such as zero-shot foreground/background separation in wildlife image datasets. (c) Moreover, recognition using multimodal language models can go beyond rigid categorical identification to more flexible recognition tasks, such as orientation recognition. Animal orientation can be complex, varying with animal postures (e.g. lying down or standing); therefore, this requires the recognition to be flexible as well. However, conventional categorical supervised models struggle with such tasks because they rely on predefined, rigid categories, leaving no room for nuance between these categories (e.g. direction between left and front). All red text in (a–c) are real results generated by GPT-4v with the input prompts. The Aspergillus image is provided by Andrea Katherine Alvarez Osorio from Universidad De Los Andes, Colombia. The Koala and Cheetah images are from the iNaturalist dataset (Van Horn et al., 2018).

tasks, is a technique for adapting a pre-trained model to novel and unseen tasks such as focusing on additional background information of an animal, even when it was not initially trained to do so. Such adaptation is realized by simply adding a few contextual examples to the input in order to guide the model to the appropriate context, like the context prompts in Figures 3 and 4a.

The interaction of in-context learning can also include external information, such as determining whether a visible animal is an

invasive species in a certain region to enhance the human–machine interaction (Figure 5). This service is made possible through the linkage of comprehensive internet knowledge sources with multimodal models like ChatGPT (OpenAI, 2023b) and Google Gemini (Gemini Team Google et al., 2023).

Of particular note is that this in-context learning process usually does not necessitate supplementary parameter updates or model fine-tuning, as long as the requisite contextual knowledge is either pre-encoded or can be extracted from an external knowledge base, such as readily accessible online materials like Wikipedia. In Sections 5.5 and 5.6, we discuss the requirements and challenges achieving a practical state of multimodal language model for biodiversity research and conservation in detail.

FIGURE 5 A conceptual example of how we envision an AI assistant might behave in the context of AI for biodiversity and conservation. Multimodal language models have the ability to adapt to specific domains, providing diverse outputs based on the user-provided context through in-context learning. This process is not confined to contexts directly related to input samples and can be extended to various other scenarios. For example, it is possible for a multimodal model to offer external information, such as determining 'if the visible animal is an invasive species in a certain region' for an AI assistant service. This service is made possible through the linkage of comprehensive internet knowledge sources with multimodal models like ChatGPT (OpenAI, 2023b), Microsoft Copilot and Google Gemini (Gemini Team Google et al., 2023). The Margay image is provided the Department of Biological Sciences at Universidad De Los Andes, Colombia. The Guina image is from the iNaturalist dataset (Van Horn et al., 2018).

4.5 | Natural language interaction

Since the advancement of multimodal language models (Gemini Team Google, 2024; OpenAI, 2023b), the interaction between humans and machines has become a prominent topic, especially in applied fields where practitioners often lack an engineering and computer programming background. With the language interface, practitioners and researchers do not need to go through programming and engineering workflows to obtain model predictions. All interactions between humans and machines can now be based solely on natural language, including human instruction, model prediction and model explanations, as shown in Figures 3 and 4. Moreover, as mentioned in Section 3, a single well-trained multimodal language model can handle many different tasks, potentially across different domains as well, eliminating the need for practitioners to train their own models, project by project, which would also require an engineering background.

However, none of these potential functionalities have been realized yet, and preliminary research in.

AI assistants has begun to focus on building powerful AI chatbots capable of fluently responding to human instructions and contexts with multimodalities to further enhance the usability based on natural language interactions (Gemini Team Google, 2024; OpenAI, 2023a). These studies aim to extend the capabilities of AI models to a broader range of tasks such as problem-solving and reasoning (Chen et al., 2021), complex image and video question answering (Mangalam et al., 2024; Yue et al., 2024) and translation (Wang et al., 2020).

Figure 5 is a conceptual example of how we envision an AI assistant might behave in the context of AI for biodiversity and conservation, illustrating how machines may gradually become more adaptive to users' needs through human–machine interactions.

5 | CHALLENGES AND DEVELOPMENTAL DIRECTIONS

Despite the flexibility and potential for new tasks enabled by multimodal language models, several limitations still exist that prevent their practical deployment and application in real-world conservation scenarios. In this section, we list some of the critical challenges and potential development directions for the use of multimodal language models for biodiversity monitoring and conservation.

5.1 | Prompt engineering and consistent model performance

A distinct challenge inherent in multimodal language models lies in the need for manual prompt engineering—manual refinement of input text prompts to generate optimal predictions—for consistent model performance on certain downstream tasks, such as large-scale categorical recognition and captioning data with language descriptions, where we cannot prompt the models sample by sample for optimal performance (Zhou et al., 2022a, 2022b). The choice of text prompts can significantly impact the generated outcomes (Zhou et al., 2022b); some prompts may enhance the performance in target tasks like in-context learning (Zhang et al., 2024b) (Section 4.4), while others could potentially derail task performance entirely (Figure 6). At present, manual prompt engineering is considered the most reliable technique for producing high-quality text prompts (i.e. text prompts that produce high-quality predictions) (Zhou, Sun, et al., 2024). For example, as shown in Figure 6, there is no effective way to prevent the model from generating the idea of a 'crab-like animal' from the rat image without manually tuning the input text prompt. Such requirement may result in added costs in terms of human labour and time for animal recognition using multimodal language models. Currently, the practical applicability of this technique is therefore limited in real-world applications due to the lack of clear guidelines on generating high-quality prompts for optimal results, context by context.

A number of approaches have sought to circumvent the need for manual prompts in categorical recognition tasks by utilizing captions and descriptions of input samples generated by language models (Fabian et al., 2023; Kim et al., 2023; Menon & Vondrick, 2023; Parkhi et al., 2012; Roth et al., 2023; Yan et al., 2023; Yang et al., 2023). For example, in (Fabian et al., 2023), they use a VLM to generate text descriptions of visible features of animals in camera trap images. These descriptions are then matched with species descriptions sourced from online databases to classify the animals. This recognition process operates without the need for active manual prompt engineering for every input image. On the other hand, these methods usually carry their own sets of limitations, such as the dependence on manually predefined visual attributes of objects (Menon & Vondrick, 2023) or inferior recognition performance compared to fully supervised models (Fabian et al., 2023). Despite being preliminary, these methods show promise in reducing the need for manual prompt engineering in multimodal language models. However, further research is needed to improve the performance of these methods and make them more accessible to practitioners in biodiversity and conservation to.

5.2 | Language and terminology bias

The languages generated or used to train existing multimodal language models are often different from domain-specific language and terminologies required for ecological and conservation-related prompts. This disconnect creates an artificial domain and knowledge gap between pretrained models and real-world applications. For example, ornithologists use terms such as *caruncles*, *tectrices* or *pileum* when describing the appearance of various body parts of birds, which rarely occur in the general domain training sets of multimodal language models. However, accurately understanding such terminology and their connection to visual features in the image

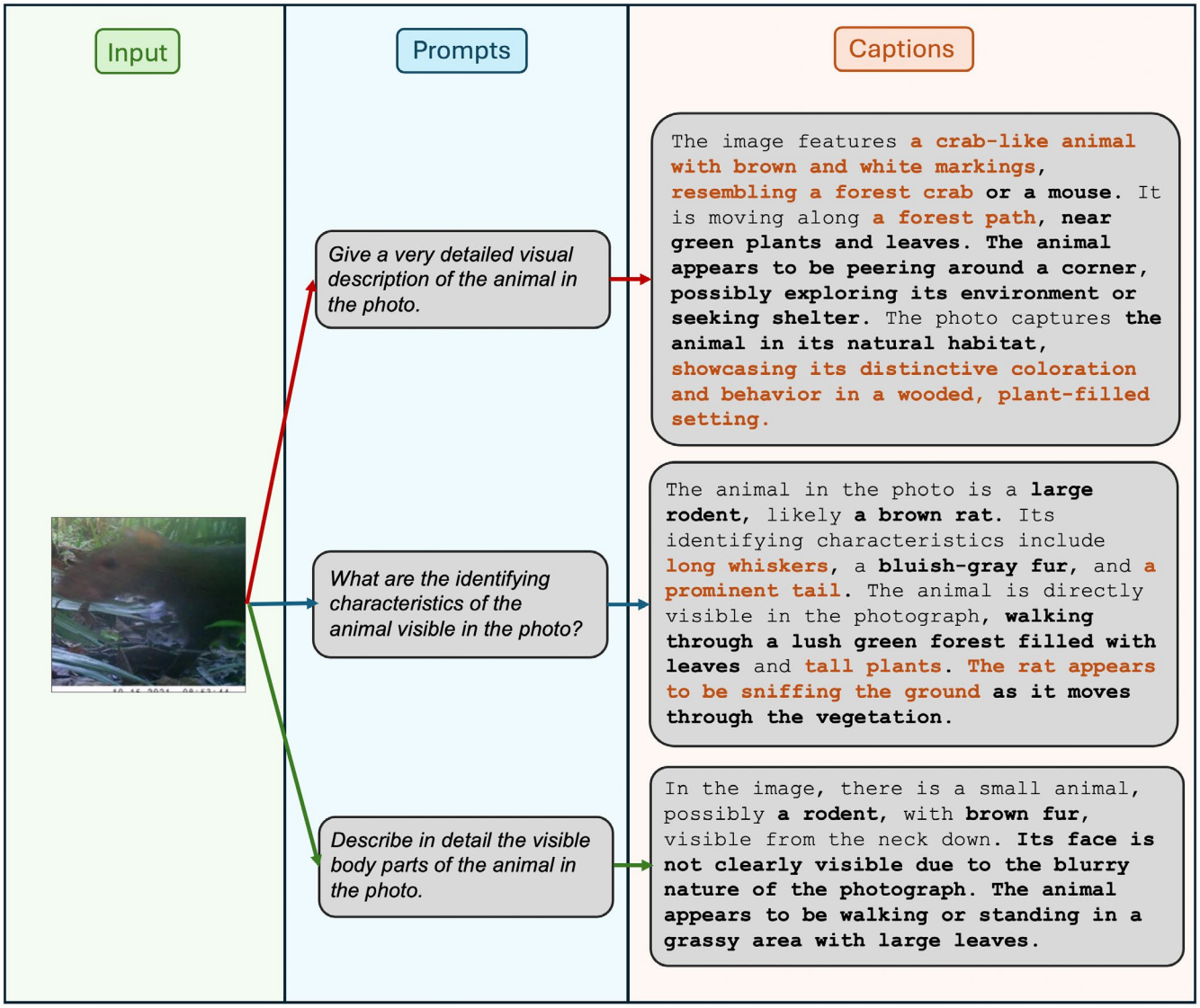


FIGURE 6 Prompt engineering and model hallucination. Prompts play a key role in the performance of multimodal language models. For example, a standard Vision–Language Model (LLaVA in this example Liu, Li, et al., 2023) produced three distinct captions for the same image using three slightly different prompts. Interestingly, the first caption identified a 'crab-like animal' that did not match the actual image content—a small rodent partially visible in the image. Unfortunately, there is no metric to evaluate the quality of prompts apart from comparing the generated outputs. However, this manual prompt engineering method may not be scalable for real-world applications like AI for biodiversity and conservation. (Bold texts are information that we think is relevant to describing the animals. Red texts are either wrong or hallucinated information by the model). The rat image is provided the Department of Biological Sciences at Universidad De Los Andes, Colombia.

can be essential in recognizing bird species. In addition, since multimodal language models heavily rely on the language feature space for alignment, their understanding of features is limited to the existing language scale. For example, if the term 'marsupials' is absent from the language features, the model cannot establish a closer relationship between features like Koala and Antechinus. In other words, language alignment does not always reflect real biological relationships unless explicitly defined. Besides terminology gaps, existing multimodal language models are largely trained in English (OpenAI, 2023a; Touvron, Lavril, et al., 2023), which may further lead to an imbalanced language representation causing challenges to practitioners from non-English speaking areas.

Instruction tuning is a technique that can address this terminology gap by inputting a relatively small amount of additional knowledge/data—compared to the scale of training data—into pretrained models for better performance on domain-specific tasks (Li et al., 2024). It is a specific type of fine-tuning technique for language-based models that focuses on task generalization (Ouyang et al., 2022) and helping the model better follow specific instructions, rather than solely improving performance on the same tasks (e.g. recognition), as in conventional transfer learning that usually fine-tunes models for similar or related tasks (Pan & Yang, 2009). For instance, (Fabian et al., 2023) successfully instruction-tuned a pretrained VLM to generate captions and descriptions with animal-specific terminology for

animal imagery from sources like camera traps and manual wildlife photographs. However, the resulting caption quality was inconsistent, with some captions offering better and more detailed descriptions of animals from input images while others only offered bare minimum descriptions (e.g. 'this is a monkey-looking animal'), primarily due to the inconsistent quality of annotations used for instruction tuning. Even though the requirement for the amount of training data and the financial and computational cost for instruction tuning is considered relatively low compared to training multimodal language models from scratch or even the traditional transfer learning with supervised approaches, the quality and variety within these annotations are critical to ensuring the performance of instruction tuning (Zhou, Liu, et al., 2023). Therefore, determining how to effectively prepare data of sufficient quality and variety for conservation and ecology tasks represents a promising research direction for exploring instruction tuning and its applications in the field.

5.3 | Systematic failures and hallucinations

Additionally, multimodal language models can exhibit systematic failures (Tong et al., 2023), which may greatly impact downstream applications. Systematic failures are errors in the model prediction triggered under specific conditions. For instance, some models may miss negative context, that is, a negation in a description, resulting in near equivalent representations for the text with and without negation (e.g. 'tree without leaves' and 'tree with leaves' being represented the same way). This can potentially lead to a flawed understanding of visual scenes. Moreover, models may fail to distinguish sentences that use quantifiers, such as *some* and *many* or specific numbers, leading to incorrect understanding of quantities of objects in images, such as the number of animals in a camera trap image. Figure 7 shows a VLM model can yield totally opposite results when the input prompts include numbers compared to when they do not.

Uncovering and addressing systematic failures in multimodal feature representations is an active area of research that defines the fundamental limitations of any practical deployment of such models (Thrush et al., 2022; Tong et al., 2023; Tong et al., 2024; Yuksekgonul et al., 2022). When it comes to biodiversity and conservation applications, such as querying data to assess whether a dataset contains invasive or endangered species, errors (either false positives or false negatives) can carry associated risks to downstream tasks such as decision and policy making. Understanding the potential pitfalls of different methods with such systematic failures is crucial when recommending such techniques to the ecological community.

When it comes to generative tasks, including image captioning, the algorithm can eventually cause model hallucination (e.g. models perceive non-existent content as existing due to various algorithmic idiosyncrasies). Hallucinations often stem from a mismatch between different data modalities (i.e. feature confusion). For instance, a multimodal language model might incorrectly respond with 'yes' to queries like 'Is X present in this image?' where X represents any animal or other objects, just like the 'crab-like animal' the model predicts in Figure 6. Such hallucination can be detrimental to real-world applications, especially in tasks that require high precision, such as animal movement habitat monitoring. There is currently no effective way to directly control hallucinations except for manual oversight. However, indirect detection of hallucinations through prediction uncertainty and consistency (Khan & Fu, 2024; Whitehead et al., 2022; Zhou, Cui, et al., 2023) are being actively studied. Such methods can usually produce quantifiable metrics on the existence hallucinations. In the applications of biodiversity, (Fabian et al., 2023) has shown potential of using instruction tuning methods and caption confidence scores to limit the caption hallucinations; however, how similar techniques can be effectively applied in the real world remains to be studied.

5.4 | The cost of model tuning

Known instances where the efficacy of multimodal language models is not guaranteed—terminology gaps, systemic failures, hallucinations—largely result from the models being trained on generic internet data rather than domain specific data (Radford et al., 2021; Schuhmann et al., 2021; Taori et al., 2023). While multimodal language models generally are better at generalizing to different data domains and distributions compared to conventional machine learning methods, substantial domain discrepancies (i.e. data differences between domains) may still cause inconsistent performance and errors (Li et al., 2024; Trinh et al., 2024). Given that applications of AI in biodiversity and conservation often encompass domain-specific tasks that may exhibit large domain differences compared to the generic internet training data (e.g. the difference between well-framed and well-lit internet images of animals and real-world noisy, obfuscated wildlife camera trap imagery) and necessitate generalization across diverse regions, time periods, sensor types and projects focusing on specific animals (Miao et al., 2021; Tuia et al., 2022), it becomes imperative to adapt existing multimodal models to cater to these distinctive requirements. While machine learning practitioners often rely on fine-tuning strategies to bridge the domain discrepancies between training and real-world inference data (Liu, Son, et al., 2023; Zhang et al., 2024a), the high cost in terms of money, time, carbon footprint, computational resources and data volume associated with multimodal models makes full model fine-tuning impractical within constrained budgets.

The substantial demand for computational resources is one of the key constraints of training and fine-tuning multimodal language models. For instance, the Flamingo (Alayrac et al., 2022) model used 1536 TPU chips, along with a substantial training period of 15 days, which is far beyond the scale of accessible resources for most academic and conservation groups. This requirement sometimes even extends to model inference (i.e. using the models for predictions) (Gemini Team Google, 2024) particularly for models that use LLMs as their language encoders like Flamingo (Alayrac et al., 2022) and GPT-4v (OpenAI, 2023a). For example, according to the pricing page

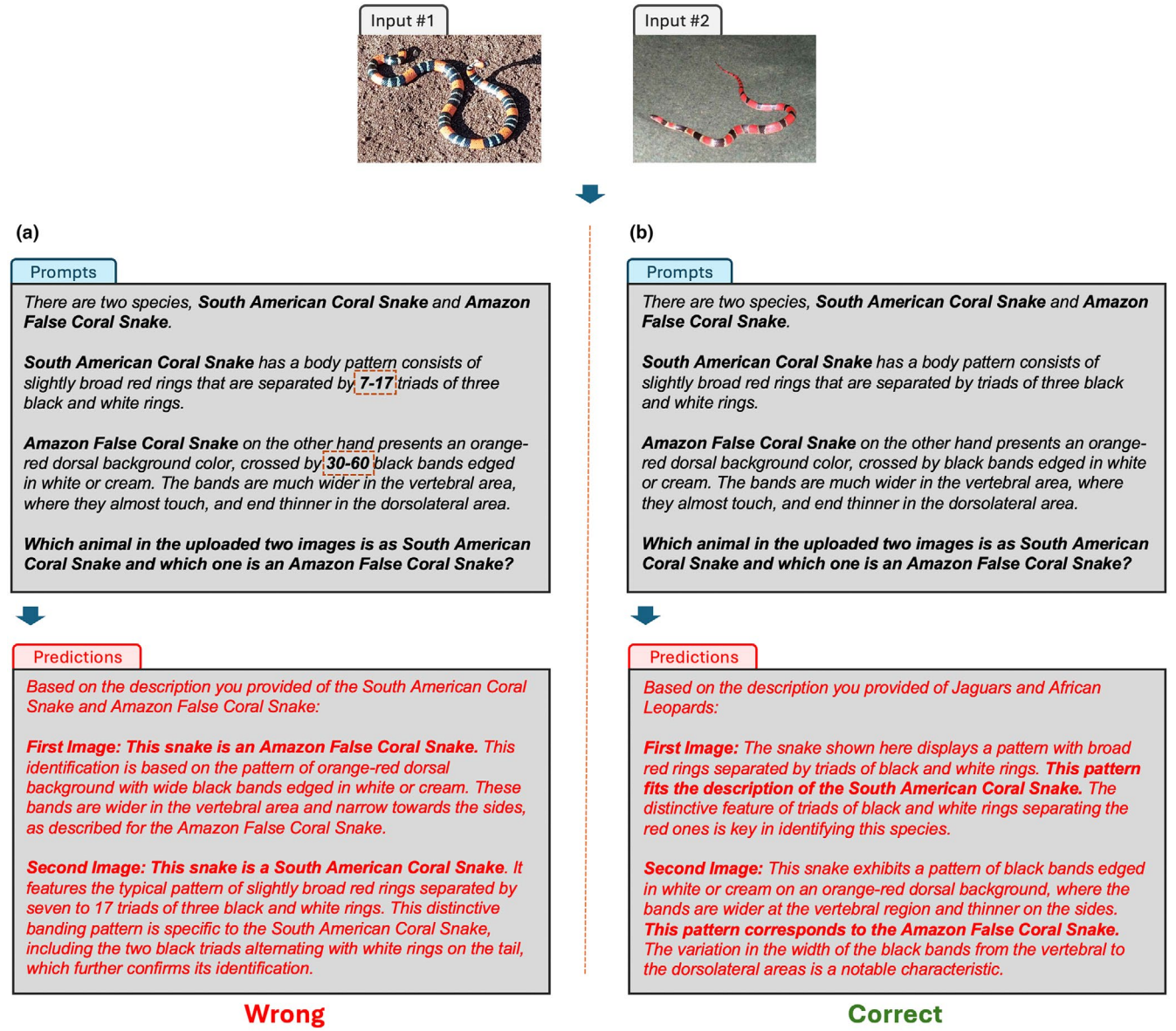


FIGURE 7 Systematic failures with numbers. (a) is an example of a failed case (b) is an example of a correct case on the same set of images. The model can make totally opposite predictions when numbers are included in the input prompts, especially when these numbers are uncertain, such as the ranges provided in the prompts (7–17 and 30–60). The snake images are from the iNaturalist dataset (Van Horn et al., 2018). All red text in (a, b) are real results generated by GPT-4v with the input prompts.

of OpenAI (https://openai.com/pricing) at the time this paper is being written, GPT-4v costs $0.04 per 1000 tokens—the basic unit of text/image processing in multimodal language models (Sennrich, 2015)—which is roughly one 224×224 image plus a paragraph of three to four hundred words. Such intensive resource requirements result in limited distribution and deployment of large-scale multimodal models in domain/task-specific and resource-limited areas such as biodiversity research and conservation that usually have large-scale datasets.

There exist methodical approaches and research directions specifically aimed at mitigating the costs associated with fine-tuning large-scale models for downstream tasks and domains (particularly in terms of time and money), thereby facilitating their adoption in real-world AI for biodiversity and conservation applications. Techniques such as model adaptors (Houlsby et al., 2019), parameter efficient tuning (Jia et al., 2022; Zhang et al., 2024a), model distillation (Zhu et al., 2023), and model compression (Kuznedelev et al., 2023) are examples that reduce the financial and computational costs of model adaptation, fine-tuning and subsequent inference. These techniques work either by introducing newly added smaller-scale trainable parameters (Jia et al., 2022; Zhang et al., 2024a) or by compressing and distilling smaller-scale models from the original large-scale models to cater downstream tasks (i.e. tasks that further make use of the outputs of these models) (Gou et al., 2021; Kuznedelev et al., 2023). The DeepSeek-R1 (DeepSeek-AI et al., 2025) is one of the best

examples, having successfully reduced the overall training cost of large models through model distillation while open sourcing much of its algorithm. It also highlights the potential of training models for fields like biodiversity research and conservation.

From Figures 3 and 4, we can also see that a generalized multimodal language model can already function within a wildlife context; therefore, a dedicated multimodal language model trained for ecology and conservation from the ground up may not be necessary. In addition, once a large model is trained—considering the task generalization ability discussed in previous sections—there may be no need to train smaller models, and the large model would enter a maintenance phase. In other words, we might only need a single multimodal model for the field of AI for conservation. While training such a model would be resource-intensive, it could ultimately save the community from having to train separate models for every individual project and task, assuming the multimodal model is properly trained.

However, at this moment, it remains unclear whether the resource trade-off between training large and small models has a net positive or negative impact, and whether there will be competition to train the best large models that will exacerbate the situation. In addition, research into the cost-efficient multimodal language model updating techniques is still in its preliminary stages with respect to real-world applications. This area, therefore, requires further research and exploration.

5.5 | Biodiversity datasets

Beyond the cost of model fine-tuning, the lack of wildlife multimodal datasets also hinders the development of such models in the applications of AI to biodiversity monitoring and conservation. Presently, the most notable contributions to the development of large-scale ecology datasets for AI/ML are embodied by LILA (https://lila.science/) and iNaturalist (Van Horn et al., 2018). These datasets, however, are predominantly designed for traditional sample-to-label based machine learning tasks. Datasets for multimodal language tasks need to have different modalities that are directly associated with each other and at least one language description for each of the imagery or audio samples. BioCLIP (Stevens et al., 2024) is a recent effort to create a multimodal dataset for biodiversity and conservation. However, the language aspect of the dataset is mainly based on direct information from the Tree of Life, rather than on image-specific contextual and descriptive information such as the ones in Figure 2. Descriptive information is crucial for the semantic alignment of multimodal language models as it helps define and incorporate ecological and conservation context information into multimodal language models. However, creating such datasets is no trivial task; even though multimodal models accept almost all types of data, making it easier to collect large-scale training datasets through Internet sources, the quality of the data remains crucial for model performance. Expert verification and annotation are required, a process that can be both time-consuming and resource-intensive.

As mentioned in Section 5.2, techniques such as instruction tuning can significantly reduce the requirement for training data to update existing multimodal language models for biodiversity and conservation-specific tasks. How to effectively and efficiently collect such data, or to augment existing biodiversity datasets with additional modalities—perhaps through collective approaches such as citizen science—remains an open question, particularly when it comes to ensuring the quality required for techniques like instruction tuning.

5.6 | Closed-source models and open-source efforts

The landscape of state-of-the-art multimodal language models is largely dominated by closed-source algorithms and datasets (Alayrac et al., 2022; Gemini Team Google, 2024; OpenAI et al., 2023; Radford et al., 2021). This approach significantly hampers the advancement of multimodal language models and poses a barrier to scientific progress in various fields, especially when modifications to existing multimodal language models are typically exclusive to contracted partners and paid services (OpenAI, 2023a), making them less accessible to practitioners. Furthermore, these closed-source strategies inhibit researchers from fully grasping the underlying mechanisms of these algorithms, even though paid services, thereby curtailing the potential for specific model modifications for different projects, tasks, and applications in real-world settings. For instance, the absence of transparency makes it impossible to understand the training process, data volume, and details of model design, such as in GPT-4v (OpenAI, 2023a), let alone to make any structural and algorithmic modifications to the models. This lack of accessibility is one of the main reasons why studies such as Fabian et al. (2023) and Stevens et al. (2024) can only use less well-developed multimodal language models to produce their wildlife models, as models like GPT cannot easily be modified by general researchers.

While open-source initiatives—mainly driven by the academia—like Open-Flamingo (Awadalla et al., 2023), Open-CLIP (Ilharco et al., 2021), LION (Schuhmann et al., 2021), BioCLIP (Stevens et al., 2024) and DeepSeek (DeepSeek-AI et al., 2025) are commendable efforts to mitigate this challenge and afford developers more accessible methods, the majority of them unfortunately fall short of achieving the performance standards set by their closed-source counterparts at the time this paper is written (Ilharco et al., 2021). Even if the cost of model training and accessibility of biodiversity and conservation-focused datasets were not a concern, the lack of technological transparency still makes training models for AI for biodiversity and conservation a challenging task. This is also one of the reasons why existing open-source efforts often have subpar performance. One point worth noting is that, despite DeepSeek-R1 (DeepSeek-AI et al., 2025) being one of the few exceptions with performance comparable to closed-source models, it is a LLM rather than a multimodal model. However, given the success of DeepSeek,

we can anticipate the release of more high-performance open-source models in the future.

5.7 | Ethical concerns and responsible AI

Ever since the introduction of AI into the field of biodiversity and conservation, ethical concerns have been raised regarding the potential misuse of AI techniques (Wearn et al., 2019). The introduction of multimodal language models into this field is no exception. For instance, besides the potential biases in the language representations of multimodal language models discussed in Section 5.2, these models can also generate biases based on imbalanced academic and social representation. Studies such as (Urzedo et al., 2024) and (Sworna et al., 2024) have shown that outputs from LLMs reflect biases present in their training data because they are heavily influenced by studies from certain countries, regions or demographic groups. While it may seem that simply adding more training data from underrepresented groups and countries could solve this problem, we argue that the issue is more complex and not unique to multimodal language models or AI. As the majority of large-scale language-based models are trained on internet data, which reflect the imbalanced nature of academic research and publications, meaningful solutions require addressing these disparities within the broader context of the academic research and publication system.

On the other hand, certain concerns about AI may seem more pressing with the advancement of large models and multimodal language models, such as their potential impact on existing staff and community members in the field of biodiversity and conservation, misinterpretation of model predictions in decision-making and unwarranted techno-optimism that may distracts effort and potential funding away from conventional solutions (Sandbrook, 2025; Wearn et al., 2019). As with any technology, there is always a risk of misuse or unintended consequences. Therefore, it is crucial to establish ethical guidelines and frameworks for the responsible use of AI techniques in biodiversity and conservation (Nandutu et al., 2023; Ullah et al., 2025; Wearn et al., 2019), an effort that requires close collaboration between technological and practitioner communities. This includes ensuring transparency in model predictions, addressing biases in training data and promoting equitable access to AI technologies across different regions and communities. By fostering a culture of responsible AI development and deployment, we can mitigate the risks associated with the misuse of AI techniques while maximizing their potential benefits for biodiversity research and conservation.

It is also important to emphasize that the development of AI techniques should not be seen as a replacement for existing human knowledge and experience but rather as a tool to assist and augment the work of practitioners in the field. This is especially true in biodiversity and conservation, where human expertise is crucial for understanding the complexities of ecosystems and species interactions.

The introduction of multimodal language models should be viewed as an opportunity to enhance the capabilities of existing staff and community members. As suggested by (Miao et al., 2021), there must be a symbiotic relationship between AI and human expertise, where AI and humans enhance each other's strengths reciprocally. This approach can not only help avoid misinterpretations and misuses of AI but also ensure that human expertise remains at the core of decision-making processes in biodiversity and conservation. Moreover, with multimodal language models, the ability to provide explanations for their predictions in natural language can further enhance the understanding and interpretability of model outputs, making it easier for practitioners to integrate AI techniques into their workflows, leading to more robust outcomes. How this new technology can be effectively incorporated into decision-making and policy-making in biodiversity and conservation remains an open question, requiring further research and exploration.

6 | PERSPECTIVES AND CONCLUSION

Considering the potential and limitations of the current development of multimodal language models, several direct applications could serve as effective starting points for utilizing these models in biodiversity research and conservation, without requiring extensive model training or fine-tuning. For instance, Figure 3 demonstrates that GPT-4v can recognize morphological characteristics of animals and leverage these traits for tasks such as animal species differentiation. Conversely, the model can also perform image retrieval tasks, functioning as a search engine for wildlife images based on natural language queries. This capability has already been implemented on the iNaturalist website (https://www.inaturalist.org/) and discussed by (Gabeff et al., 2024). Although the performance may not be flawless, it provides a strong example of how multimodal language models can be applied to real-world scenarios at their current stage of development. There are also pioneering works that have demonstrated the potential of multimodal language models in animal behaviour studies (Brookes et al., 2024; Dussert et al., 2025) as the obscure nature of animal behaviour and language descriptions can be a natural fit.

Moreover, with the robust language understanding capabilities and access to internet resources, existing multimodal language models can be directly integrated into research workflows for tasks such as initial mass data collection, data analysis and research summarization. Additionally, as open-source solutions like DeepSeek (DeepSeek-AI et al., 2025) become more prevalent, customized multimodal language models tailored for biodiversity research and conservation will likely become more accessible to the public in the future.

In conclusion, multimodal language models represent a revolutionary advancement in AI, with the potential to serve as AI assistants that can transform biodiversity research and conservation through natural language-based human–machine interactions, enabling a wide range of tasks (Sections 3 and 4). However, the deployment of these multimodal models in ecological research and conservation practice, in particular, still faces several challenges

that must be addressed when moving forward (Section 5). The development of multimodal language models will require interdisciplinary collaboration across different communities, including computer science and biodiversity experts, as well as efforts from both industry and academia. We hope this discussion will inspire further research and interdisciplinary collaboration to fully realize the potential of multimodal language models in biodiversity research and conservation.

AUTHOR CONTRIBUTIONS

This project was conceived by Zhongqi Miao, Yuanhan Zhang, Zalan Fabian, Sara Beery, Rahul Dodhia and Juan Lavista Ferres. Experiments were done by Zhongqi Miao, Yuanhan Zhang and Andres Hernandez Celis. Main text was written and reviewed by all of the authors.

ACKNOWLEDGEMENTS

We want to thank Andres Montes-Rojas, Rafael Escucha, Laura Siabatto, and Andres Link from the Department of Biological Sciences at Universidad de los Andes, Colombia, for giving us permission to use some of their wildlife images in our paper. We want to thank Andrea Katherine Alvarez Osorio from Universidad de Los Andes, Colombia, for giving us permission to use her image of Aspergillus in our paper. We would also like to express our gratitude to everyone who contributed to this project. This project is partially funded by AI for Biodiversity Change Global Centre (NSF Award No. 2330423 and NSERC Award No. 585136).

CONFLICT OF INTEREST STATEMENT

Several authors are affiliated with Microsoft, which is active in AI and large model development. However, we confirm that nothing mentioned in this paper involves any direct financial or other relationships with our authors. This paper is intended as a purely academic discussion.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70120.

DATA AVAILABILITY STATEMENT

This paper did not use large-scale datasets to produce experiment results. Most of the images used in this project are from publicly available and published sources like iNaturalist. There are two images used in the figures that are provided by individuals from the Department of Biological Sciences at Universidad de los Andes, Colombia. All the images used in this paper and prompts used for actual VLM experiments are available at https://zenodo.org/records/15794521 with https://doi.org/10.5281/zenodo.15794520 (Miao, Zhang, et al., 2025). There is no code developed for this project as most of the results are directly generated from ChatGPT. However, we strongly encourage readers to explore the multimodal language models and their codebase discussed in this paper such as LLaVA and BioCLIP as a starting point for their own biodiversity research and conservation applications. These models and codebase are usually available through GitHub and Hugging Face.

ORCID

Zhongqi Miao https://orcid.org/0000-0002-0439-8592

REFERENCES

Ahumada, J. A., Fegraus, E., Birch, T., Flores, N., Kays, R., O'Brien, T. G., Palmer, J., Schuttler, S., Zhao, J. Y., Jetz, W., Kinnaird, M., Kulkarni, S., Lyet, A., Thau, D., Duong, M., Oliver, R., & Dancer, A. (2020). Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, *47*(1), 1–6.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J. L., Borgeaud, S., … Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, *35*, 23716–23736.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P. W., Ilharco, G., Wortsman, M., & Schmidt, L. (2023). Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Bao, W., Chen, L., Huang, H., & Kong, Y. (2023). Prompting language-informed distribution for compositional zero-shot learning. *arXiv preprint arXiv:2305.14428*.

Beery, S., Morris, D., & Yang, S. (2019). Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*.

Brookes, O., Mirmehdi, M., Kuhl, H., & Burghardt, T. (2024). Chimpvlm: Ethogram-enhanced chimpanzee behaviour recognition. *arXiv preprint arXiv:2404.08937*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Jeffrey, W., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhou, S. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dodhia, R. (2024). *AI for social good: Using artificial intelligence to save the world*. Wiley.

Dussert, G., Miele, V., Van Reeth, C., Delestrade, A., Dray, S., & Chamaillé-Jammes, S. (2025). Zero-shot animal behaviour classification with vision-language foundation models. *Methods in Ecology and Evolution*, *16*(7), 1460–1472.

Elizalde, B., Deshmukh, S., Ismail, M. A., & Wang, H. (2022). Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*.

Fabian, Z., Miao, Z., Li, C., Zhang, Y., Liu, Z., Hernández, A., Montes-Rojas, A., Escucha, R., Siabatto, L., Link, A., Arbeláez, P., Dodhia, R., & Ferres, J. L. (2023). Multimodal foundation models for zero-shot animal species recognition in camera trap images. *arXiv preprint arXiv:2311.01064*.

Gabeff, V., Rußwurm, M., Tuia, D., & 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*, 1–3786.

Gemini Team Google. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*.

Gemini Team Google, 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*.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*, 1789–1819.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *International conference on machine learning* (pp. 2790–2799). PMLR.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. (2021). OpenCLIP. *Zenodo*. https://doi.org/10.5281/zenodo.5143773

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S.-N. (2022). Visual prompt tuning. In *European conference on computer vision* (pp. 709–727). Springer Nature Switzerland.

Jiao, B., Liu, L., Gao, L., Wu, R., Lin, G., Wang, P., & Zhang, Y. (2024). Toward re-identifying any animal. *Advances in Neural Information Processing Systems*, *36*, 40042–40053.

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236.

Kay, J., Haucke, T., Stathatos, S., Deng, S., Young, E., Perona, P., Beery, S., & Van Horn, G. (2024). Align and distill: Unifying and improving domain adaptive object detection. *arXiv preprint arXiv:2403.12029*.

Kellenberger, B., Veen, T., Folmer, E., & Tuia, D. (2021). 21 000 birds in 4.5 h: Efficient large-scale seabird detection with machine learning. *Remote Sensing in Ecology and Conservation*, *7*(3), 445–460.

Khan, Z., & Fu, Y. (2024). Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10854–10863).

Kim, J. M., Koepke, A., Schmid, C., & Akata, Z. (2023). Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2584–2594).

Kuznedelev, D., Tabesh, S., Noorbakhsh, K., Frantar, E., Beery, S., Kurtic, E., & Alistarh, D. (2023). Vision models can be efficiently specialized via few-shot task-aware compression. *arXiv preprint arXiv:2303.14409*.

Kwok, R. (2019). Ai empowers conservation biology. *Nature*, *567*(7746), 133–134.

Li, B., Shen, Y., Yang, J., Wang, Y., Ren, J., Che, T., Zhang, J., & Liu, Z. (2022). Sparse mixture-of-experts are domain generalizable learners. *arXiv preprint arXiv:2206.04046*.

Li, B., Zhang, Y., Chen, L., Wang, J., Fanyi, P., Cahyono, J. A., Yang, J., Li, C., & Liu, Z. (2025). Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.48550/arXiv.2305.03726

Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., & Liu, Z. (2023). Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., & Gao, J. (2023). Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020,1(2):2*.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2024). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, *36*, 28541–28564.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, *34*, 9694–9705.

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., & Marculescu, D. (2023). Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7061–7070).

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, *36*, 34892–34916.

Liu, H., Son, K., Yang, J., Liu, C., Gao, J., Lee, Y. J., & Li, C. (2023). Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15148–15158).

Liu, Z., Miao, Z., Pan, X., Zhan, X., Lin, D., Yu, S. X., & Gong, B. (2020). Open compound domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S. X. (2019). Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2537–2546).

Lv, F., Chen, X., Huang, Y., Duan, L., & Lin, G. (2021). Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2554–2562).

Mangalam, K., Akshulakov, R., & Malik, J. (2024). Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, *36*, 46212–46244.

Menon, S., & Vondrick, C. (2023). *Visual classification via description from large language models*. ICLR.

Miao, Z., Elizalde, B., Deshmukh, S., Kitzes, J., Wang, H., Dodhia, R., & Ferres, J. L. (2025). Multi-modal Language models in bioacoustics with zero-shot transfer: A case study. *Scientific Reports*, *15*, 7242. https://doi.org/10.1038/s41598-025-89153-3

Miao, Z., Gaynor, K. M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M. S., McInturff, A., Bowie, R. C., Nathan, R., Yu, S. X., Bowie, R. C. K., & Getz, W. M. (2019). Insights and approaches using deep learning to classify wildlife. *Scientific Reports*, *9*(1), 8137.

Miao, Z., Liu, Z., Gaynor, K. M., Palmer, M. S., Yu, S. X., & Getz, W. M. (2021). Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, *3*(10), 885–895.

Miao, Z., Yu, S. X., Landolt, K. L., Koneff, M. D., White, T. P., Fara, L. J., Hlavacek, E. J., Pickens, B. A., Harrison, T. J., & Getz, W. M. (2023). Challenges and solutions for automated avian recognition in aerial imagery. *Remote Sensing in Ecology and Conservation*, *9*(4), 439–453.

Miao, Z., Zhang, Y., Fabian, Z., & Celis, A. H. (2025). Prompts and images for VLM conservation perspective (Version v1). *Zenodo*. https://doi.org/10.5281/zenodo.15794520

Misra, I., Gupta, A., & Hebert, M. (2017). From red wine to red tomato: Composition with context. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1792–1801).

Nandutu, I., Atemkeng, M., & Okouma, P. (2023). Integrating ai ethics in wildlife conservation ai systems in South Africa: A review, challenges, and future research agenda. *AI & Society*, *38*, 1–13.

OpenAI. (2023a). *Gpt-4v(ision) system card*.

OpenAI. (2023b). *Introducing chatgpt*.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., & Avila, R. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

Ouyang, L., Jeffrey, W., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano,

P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Parkhi, O. M., Vedaldi, A., Zisserman, A., & Jawahar, C. (2012). Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 3498–3505). IEEE.

Perona, P. (2010). Vision of a visipedia. *Proceedings of the IEEE*, *98*(8), 1526–1534.

Radford, A., Jeffrey, W., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PmLR.

Rhinehart, T. A., Chronister, L. M., Devlin, T., & Kitzes, J. (2020). Acoustic localization of terrestrial wildlife: Current practices and future opportunities. *Ecology and Evolution*, *10*(13), 6794–6818.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695).

Roth, K., Kim, J. M., Koepke, A. S., Vinyals, O., Schmid, C., & Akata, Z. (2023). Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 15746–15757).

Sandbrook, C. (2025). Beyond the hype: Navigating the conservation implications of artificial intelligence. *Conservation Letters*, *18*(1), e13076.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., & Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Sennrich, R. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sherman, J., Ancrenaz, M., & Meijaard, E. (2020). Shifting apes: Conservation and welfare outcomes of Bornean orangutan rescue and release in Kalimantan, Indonesia. *Journal for Nature Conservation*, *55*, 125807.

Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*.

Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., Dong, L., Dahdul, W. M., Stewart, C., Berger-Wolf, T., Chao, W.-L., & Su, Y. (2024). Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19412–19424).

Stowell, D., Wood, M. D., Pamu la, H., Stylianou, Y., & Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution*, *10*(3), 368–380.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., & Wang, X. (2023). Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Sworna, Z. T., Urzedo, D., Hoskins, A. J., & Robinson, C. J. (2024). The ethical implications of chatbot developments for conservation expertise. *AI and Ethics*, *4*(4), 917–926.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). *Stanford alpaca: An instruction-following llama model*. https://github.com/tatsu-lab/stanford_alpaca

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5238–5248).

To-Anun, C., Hidayat, I., & Meeboon, J. (2011). Genus Cercospora in Thailand: Taxonomy and phylogeny (with a dichotomous key to species). *Plant Pathology & Quarantine*, *1*(1), 11–87.

Tong, S., Jones, E., & Steinhardt, J. (2023). Mass-producing failures of multimodal systems with language models. *arXiv preprint arXiv:2306.12105*.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes wide shut? Exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi'ere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi'ere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Jeremy, F., Wenyin, F., … Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature*, *625*(7995), 476–482.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., & Hill, F. (2021). Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, *34*, 200–212.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., van Horn, G., Crofoot, M. C., Stewart, C. V., & Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications*, *13*(1), 792.

Ullah, F., Saqib, S., & Xiong, Y.-C. (2025). Integrating artificial intelligence in biodiversity conservation: Bridging classical and modern approaches. *Biodiversity and Conservation*, *34*(1), 45–65.

Urzedo, D., Sworna, Z. T., Hoskins, A. J., & Robinson, C. J. (2024). Ai chatbots contribute to global conservation injustices. *Humanities and Social Sciences Communications*, *11*(1), 1–8.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8769–8778).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* 30, Long Beach, CA, USA.

Wang, C., Wu, A., & Pino, J. (2020). Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Wearn, O. R., Freeman, R., & Jacoby, D. M. (2019). Responsible AI for conservation. *Nature Machine Intelligence*, *1*(2), 72–73.

Whitehead, S., Petryk, S., Shakib, V., Gonzalez, J., Darrell, T., Rohrbach, A., & Rohrbach, M. (2022). Reliable visual question answering: Abstain rather than answer incorrectly. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *European conference on computer vision* (pp. 148–166). Springer.

Wu, S., Fei, H., Leigang, Q., Ji, W., & Chua, T.-S. (2024). Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, Messe Wien Exhibition Congress Center, Vienna, Austria.

Wu, X., Zhu, F., Zhao, R., & Li, H. (2023). Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7031–7040).

Yan, A., Wang, Y., Zhong, Y., Dong, C., He, Z., Lu, Y., Wang, W. Y., Shang, J., & McAuley, J. (2023). Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3090–3100).

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., & Yatskar, M. (2023). Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19187–19197).

Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., Ross, C., Polyak, A., Howes, R., Sharma, V., Xu, P., Tamoyan, H., Ashual, O., Singer, U., Li, S.-W., … Aghajanyan, A. (2023). Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., … Chen, W. (2024). Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9556–9567).

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2022). When and why vision-language models behave like bags-of-words, and what to do about it? In *The eleventh international conference on learning representations*.

Zhang, Y., Zhou, K., & Liu, Z. (2024a). Neural prompt search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2024.3435939

Zhang, Y., Zhou, K., & Liu, Z. (2024b). What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, *36*, 17773–17794.

Zhong, M., Torterotot, M., Branch, T. A., Stafford, K. M., Royer, J.-Y., Dodhia, R., & Lavista Ferres, J. (2021). Detecting, classifying, and counting blue whale calls with Siamese neural networks. *The Journal of the Acoustical Society of America*, *149*(5), 3086–3094.

Zhou, C., Liu, P., Puxin, X., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Ping, Y., Lili, Y. U., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023). Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, *36*, 55006–55021.

Zhou, D.-W., Sun, H.-L., Ning, J., Ye, H.-J., & Zhan, D.-C. (2024). Continual learning with pre-trained models: A survey. *arXiv preprint arXiv:2401.16386*.

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a). Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16816–16825).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision*, *130*(9), 2337–2348.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., & Yao, H. (2023). Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Zhu, X., Li, J., Liu, Y., Ma, C., & Wang, W. (2023). A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*.

How to cite this article: Miao, Z., Zhang, Y., Fabian, Z., Hernandez Celis, A., Beery, S., Li, C., Liu, Z., Gupta, A., Nasir, M., Li, W., Holmberg, J., Palmer, M., Gaynor, K., Arbelaez, P., Wang, P., Dodhia, R., & Ferres, J. L. (2026). New frontiers in artificial intelligence for biodiversity research and conservation with multimodal language models. *Methods in Ecology and Evolution*, *17*, 238–256. https://doi.org/10.1111/2041-210x.70120