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

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

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

Abstract

Artificial Intelligence (AI) for biodiversity and conservation is growing rapidly, demonstrating great potential in reducing the intensive human labor 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.

The recent advent of multimodal language (MML) models has significantly expanded the realm of possible AI applications in biodiversity research. These models have demonstrated the ability to recognize animals and complex concepts, such as animal postures and orientations, without prior exposure during training. MML 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, there are unique barriers to the use of MML models for biodiversity applications, including high computational and financial demands, reliance on prompt engineering for consistent performance on large datasets, and insufficient open-source sharing of state-of-the-art methods.

This paper explores the transformative potential of MML models for biodiversity research, compared with traditional machine learning methods, and discusses several potential applications in biodiversity research. We also discuss challenges to implementing these models in real-world biodiversity scenarios and propose directions for future research to overcome these hurdles. Our goal is to encourage robust discussions and research into the integration of MML models to advance AI for biodiversity research and conservation.

1 Introduction

The field of Artificial Intelligence (AI) for biodiversity research and conservation is rapidly gaining traction within the ecological and biological sciences [1, 2]. 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) [3, 4, 5, 6, 7] and audio data sources (*e.g.*, bioacoustics) [8, 9, 10, 11, 12]. These applications have demonstrated potential for reducing the substantial human labor traditionally required for data processing [1, 4], thus enabling the collection of more extensive datasets, providing ecological granularity at unprecedented spatial and temporal scales [7]. 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. These include, but are not limited to, model performance inconsistencies caused by severely imbalanced or long-tailed data distribution [13] and differences in datasets and applications (*i.e.*, multi-domain discrepancies) [14], various issues from the complexity of open-world datasets (*e.g.*, varying data quality and novel/unseen categories) [4], and most importantly, the inaccessibility of existing algorithms to practitioners with limited programming and engineering knowledge.

The recent advent of multimodal language (MML) models—models that can process and generate both textual content and other data modalities (*e.g.*, video and audio) [15, 16, 17, 18, 19, 20, 21]—has markedly enhanced the versatility and possibilities of AI applications [22]. 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 MML model like GPT-4v can recognize animals, even those that closely resemble each other, without seeing them during training (known as

zero-shot recognition [23]), holding 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 training. Additionally, because MML 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 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 MML models now become natural language-based, eliminating the need for complex programming procedures. This can significantly improve the accessibility of AI techniques for practitioners with limited engineering and programming experience.

Despite the promise of MML models, their application faces unique challenges compared to traditional AI techniques. These include a significantly higher demand for computational and financial resources [17], a strong reliance on developing the correct text prompts (*i.e.*, prompt engineering) for model performance [24, 25], limited open-source sharing of advanced MML model methods [17, 26, 15], and a series of systematic failures of these methods [27, 28, 29, 30], such as failing to differentiate sentences with quantifiers and numbers. Therefore, in this paper, we aim to explore the transformative impact MML models can have on the future of biodiversity research and conservation and fully discuss the challenges of such novel techniques within these contexts. We begin with a detailed comparison of the fundamental differences between MML models and conventional machine learning methods (Section 2) and explore how these differences engender new applications of MML techniques in biodiversity research and conservation (Section 3 and 4). Then, we discuss the challenges and limitations we have identified for successfully implementing MML models in real-world scenarios and propose 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 MML models, which could significantly advance the field of AI for biodiversity understanding and conservation.

2 Multimodal Language Models

2.1 Multimodal models

The exploration of multimodal models has gained considerable attention in recent years, largely due to their unique capability to simultaneously process and generate a variety of data types or data modalities, including visual, audio, and language.

In general, there are two types of multimodal models: multimodal contrastive models and multimodal generative models. The former focuses on creating a shared multi-dimensional embedding space (*i.e.*, feature space) across different data modalities, while the latter uses such a space to generate different modalities of data (Figure 1):

- **Multimodal contrastive models:**

  These models encode information from a variety of data types (modalities), such as language/text, imagery, and audio data, into a shared **feature space**. This feature space is where each data sample is encoded or represented as a multi-dimensional vector using feature encoders like Convolutional Neural Networks (CNNs) [31] or Transformers [32]. Feature representations from semantically related data (*e.g.*, images and audio clips from the same animal species), irrespective of their modalities, are aligned in this shared feature space through a technique known as Contrastive Learning—a type of machine learning technique that aims to maximize the similarities among sample features. This feature space is crucial for the effectiveness of downstream applications that utilize the outputs of multimodal contrastive models [33].

  For example, CLIP (Contrastive Language-Image Pretraining) [15], a Vision-Language Model, learns an aligned image-text feature space by training with 400 million image-text pairs. This process creates an association between images and texts, enabling recognition from categories defined after the model is trained through similarity calculation between texts of post-defined categories and input images (rather than relying on predefined categories as would be required for training traditional machine learning models). Subsequently, AudioCLIP [34] and Wav2CLIP [35] extend CLIP to the audio modality, embedding audio into a shared feature space with images and text, allowing audio data to be directly associated with natural language. Recently, Meta AI presented ImageBind [33], a multimodal model that specializes in integrating six modalities—text, image/video, audio, depth, thermal, and inertial measurement units (IMU)—into a shared feature space where all modalities of data can be interchangeably associated with each other.

- **Multimodal generative models:**

  Multimodal generative models move one step further than contrastive models, possessing the capacity to generate any mixture of output modalities [36, 37, 38] from various combinations of input modalities. For instance, just as the popular Stable Diffusion model [39] can generate images from textual inputs, models like Flamingo [17] or GPT-4v [16] can directly generate the textual output "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 eliminates the need to convert model outputs into categories, as is common in multimodal contrastive models and traditional machine learning protocols. In other words, the outputs are not limited to categories predefined or post-defined by humans;

rather, they can directly generate the outputs based on the inputs to the models.

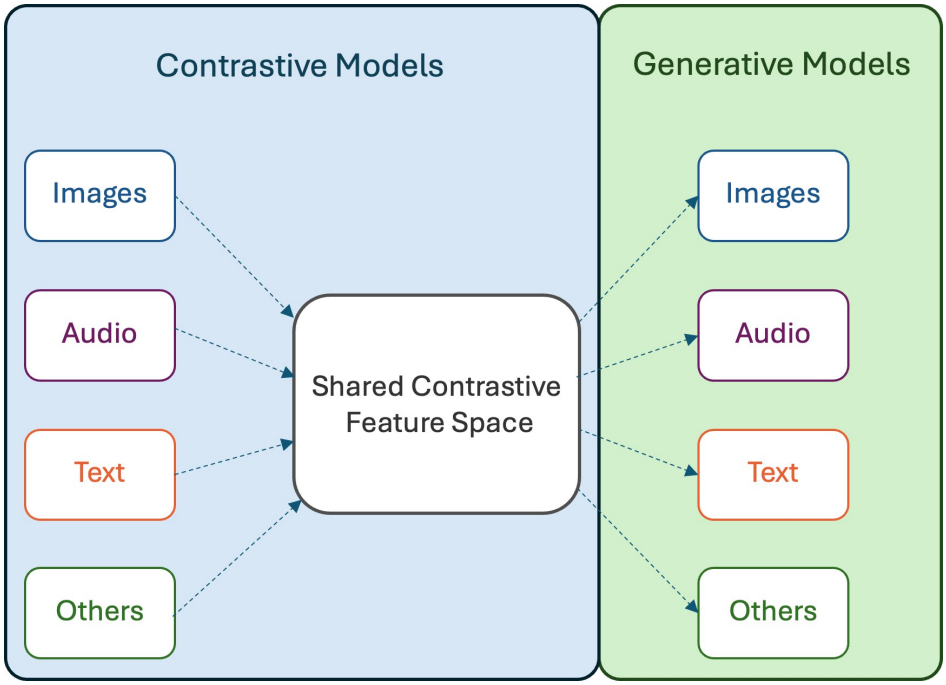

Figure 1: **Illustration of multimodal contrastive and multimodal generative models.** While multimodal contrastive models focus on constructing a shared feature space, aligning features from various input modalities such as images, audio, text, and other modalities, multimodal generative models take a step further. They utilize this shared feature space to generate any output modalities from any input modalities of data, such as creating images from a language description.

2.2 Multimodal language models and conventional machine learning

Among the many combinations of multiple modalities [40, 41, 42], multimodal language (MML) models have emerged as one of the most widely and actively researched areas, spurred by the advancement of single-modality large language models (LLMs) [43, 44, 26]. One of the fundamental differences between MML models and conventional machine learning (supervised or unsupervised) lies in the focus of MML models on aligning language concepts and semantics with other data modalities, primarily perceptual ones such as images and audio. Moreover, these language concepts and their semantics are not confined to predefined language categories, such as a predetermined set of animal species. Instead, they can encompass a wide array of words and phrases representing ideas, including descriptive attributes of animals and the relationships between these ideas.

In traditional machine learning frameworks, perceptual data typically lack direct connections to language concepts or semantics. For example, in categorical supervised learning, data are often mapped to discrete numerical labels [31], representing the language categories humans predefined (for example, 1 = dogs, 2 = cats). However, these categories are often overly simplified abstractions of the ultimate language concepts they aim to represent. For example, in an animal species classification dataset, each species is usually labeled with numerical numbers, with no relationships between each other (Figure 2 a). However, animal species often have taxonomic and morphological relationships, which can provide useful information for model training [45]. A species classification model trained with traditional supervised machine learning frameworks would be limited by the absence of encoded semantic relationships between these discrete numerical labels, inhibiting the explicit or implicit learning of semantic relationships through discrete categorical supervision. Take iWildCam2019 [46], a typical categorical wildlife recognition dataset, as an example. Despite containing 7 antelope species, these categories are nothing more than independent and discrete digits, indistinguishable from categories such as *Raccoon* or *Black Bear*. Even though similar looking categories are usually closer to each other in a trained feature space in terms of embedding distance [47], the semantic relationships between categories are not naturally expressed under label-based supervision.

MML models are not constrained by the sample-to-label mapping typical of supervised learning approaches. Instead, they directly align the features (or "embeddings") of perceptual data directly with language concept features (Figure 2 b). In other words, the learning process of MML models aims to maximize similarities between language concept features and features of perceptual data, rather than using predefined labels to dictate/supervise where the features of input data should be positioned in the feature space. To achieve

this feature alignment, most MML models utilize large-scale online datasets comprised of perception-to-language pairs, such as image-to-language pairs [15] and audio-to-language pairs [48]. This pairing format ensures that each perceptual input can be associated with unique language descriptions, thereby not only providing a comprehensive breadth of perception-to-language alignment but also effectively eliminating the need for predefined labels. Figure 2 b gives an example of how such image-to-language pairs may look like when it comes to wildlife imagery.

Most importantly, since the language concept features learned by large language models (LLMs) are continuous instead of categorical and directly encode semantic relationships within the feature space [49], the aligned perceptual features also inherit these continuous semantic relationships. Such a continuous feature space with intrinsic semantic relationships allow the recognition of subtle similarities and differences between categories with complex concepts such as animal orientations and postures (detailed in Section 3), that is not feasible in conventional, predefined machine learning protocols.

For instance, in Figure 2, *Koala*, *Antechinus*, and *Mouse* are encoded as discrete and independent digit labels in categorical supervised learning. Conventional supervised learning models would only generalize basic visual similarities between *Antechinus* and *Mouse* while completely ignoring the phylogenetic relationship between the two marsupials, *Koala* and *Antechinus* [47]. In contrast, models developed through the alignment of visual and language features, particularly Vision-Language Models (VLMs) [15, 20, 19], can compel the features of the *Koala* and *Antechinus* to be closer in the shared visual-language feature space even before generalizing on visual similarities, recognizing both categories as *Marsupials*, which are semantically distinct from Mouse. These semantic relationships, inherently encoded in Vision-Language Models (VLMs) through Large Language Models (LLMs) learned from large-scale online textual resources such as Wikipedia, serve as a form of supervision to regulate the shared feature space. Therefore, such supervision can further enhance the semantic relevance between categories or animals, rather than producing only visual categorical similarities, as is common in supervised models. On the other hand, a trained VLM may still consider the *Antechinus* and *Mouse* visually closer to each other because both resemble mice, allowing for the recognition of both visual similarities and semantic relationships.

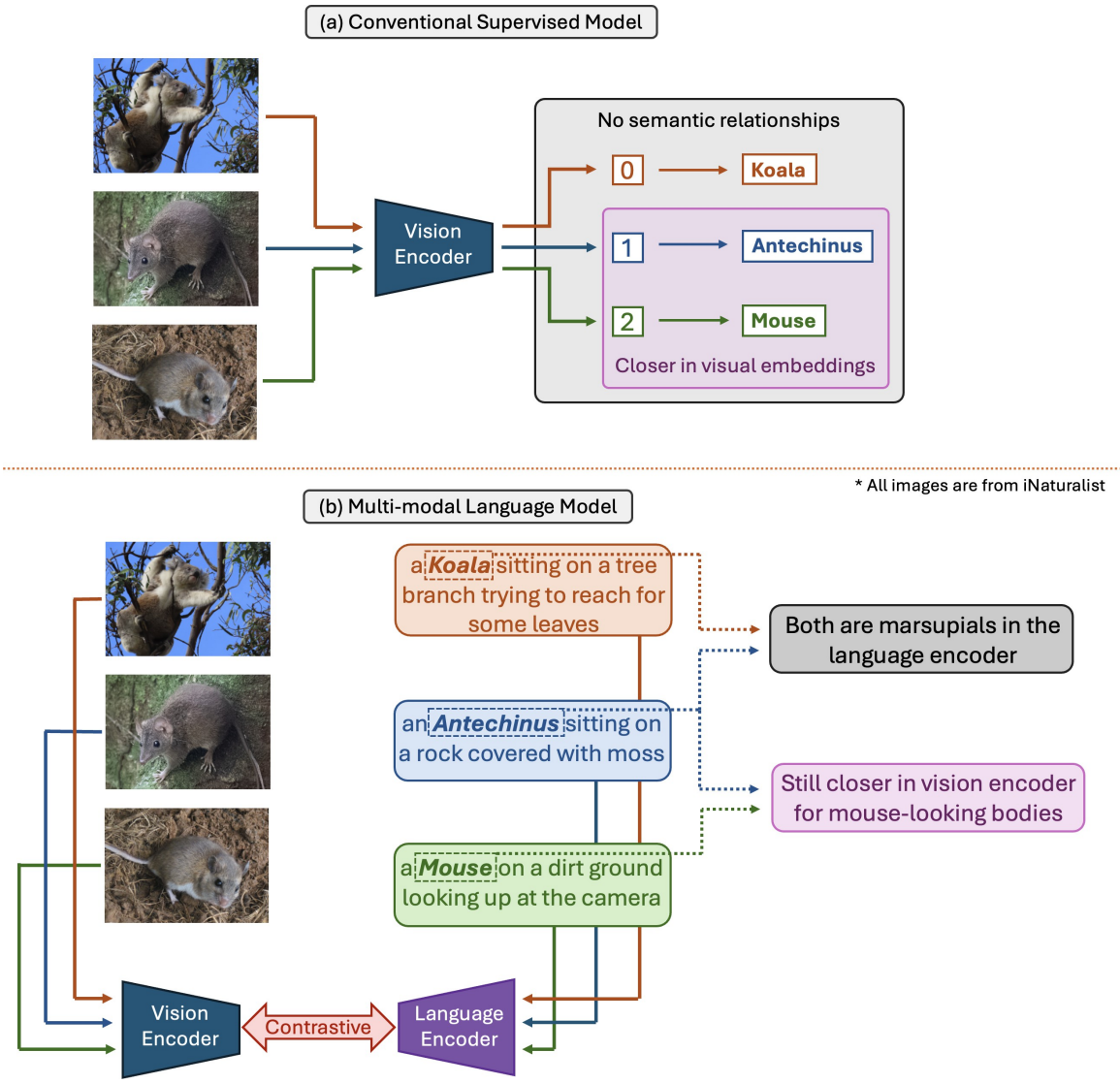

Figure 2: **Comparison between conventional supervised learning and MML models on semantic relationships.** In the conventional supervised learning framework, samples are typically encoded into discrete digital labels for recognition tasks. These labels do not possess intrinsic semantic relationships, even when the categories they represent are closely related, whether visually or semantically. Conversely, the training of MML models is aimed at aligning the features of language with other input modalities. These models do not have categorical limitations, and the semantic relationships are naturally encoded and expressed in the shared feature space. For instance, even though *Koala* and *Antechinus* look distinctly different from each other, these images can still have connections to each other in the shared feature space because both *Koala* and *Antechinus* are encoded as marsupials in the language encoder.

3 Multimodal language models and zero-shot recognition

MML models have revolutionized machine learning by transitioning from conventional sample-to-label mapping to sample-to-semantic-association mapping. This paradigm shift introduces a considerable degree of flexibility to various AI tasks [36, 50, 17, 51, 52]. 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 MML models depends largely on the alignment between perceptual information (*e.g.*, visual features) and natural languages. 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 it can use to differentiate pairs of similar-looking animals without dedicated training.

Figure 4 (a) 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 MML 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 [53] 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 em- beddings can further facilitate unprecedented zero-shot tasks, such as open-vocabulary segmentation and detection (Figure 4 b) [55, 56], 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, unconventional recognition tasks that go beyond rigid categorical recognition are also made possible because natural language is not confined to categorical concepts. Figure 4 (c) 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) [57], which heavily depends on accurately matching animal body markings to the correct side of the animal. More importantly, all these different language-based tasks can be realized with a single MML model (GPT-4v in this case), instead of using independent models for each task as in traditional machine learning.

However, as shown in both Figure 3 and 4, zero-shot recognition can heavily rely on text prompts and human inputs. In Section 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 MML models in detail.

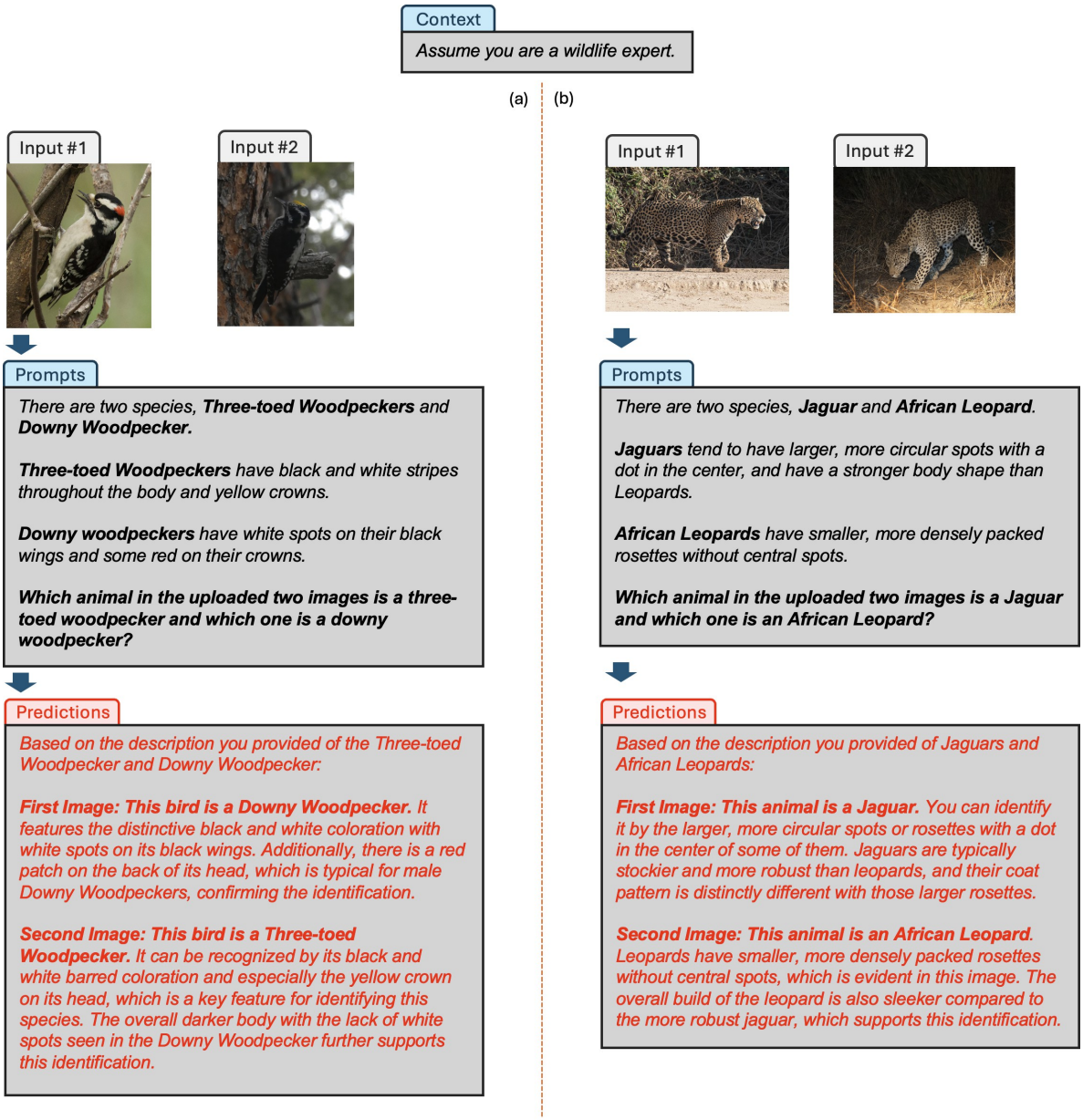

Figure 3: **Vision-Language models understand morphological characteristics.** We use GPT-4v to differentiate between two sets of animal pairs by providing morphological descriptions to the model with a contextual prompt, "Assume you are a wildlife expert". The model not only correctly differentiates these similar-looking animals based on the provided descriptions—implying its understanding of animal morphological characteristics—but also provides detailed reasoning for its predictions.

4 Other tasks made possible by multimodal language models

Beyond zero-shot recognition, MML 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 MML models.

4.1 Learning from very few samples

One of the tasks that MML 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 [58]. As presented in [17], simply prompting VLMs with as few as four task-specific examples (*i.e.*, train the model to recognize target categories with as few as four training samples), such models are able to produce comparable if not superior performance than methods fine-tuned on thousands of examples from the same categories. Few-shot learning is a task that is relevant to many AI for biodiversity and conservation application scenarios, such as endangered species monitoring [59], where such tasks typically involve target categories/animals with limited available data. The advancement of few-shot learning with MML 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 [1, 4]. For example, models trained with conventional supervised learning methods may not generalize 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 [4, 47]. These differences in datasets are referred to as domain discrepancies.

MML models, on the other hand, often have a higher capacity for generalization across various data distribution/domains, primarily due to the alignment mechanism between perceptual and language concepts. As mentioned in Section 2, the shared feature space of MML models is continuous and does not have hard boundaries (*e.g.*, decision boundaries) that define and confine categories; recognition is fully based on how similar the input objects/concepts are to the existing language features in the feature space. In other words, any objects that look like a bird can be associated to the language concept "bird", regardless what environment these objects are in, thus generalizing across different data domains. Moreover, since most MML 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. For example, in [60], 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) recognizing the animals sounds from different datasets without dedicated training and still achieve supervised level performance. Despite the potential, the domain generalization capability of current MML models is still limited by the scale of domain discrepancies—scale of differences between datasets collected from different domains [61, 62]. 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 Section 5.2, 5.4, and 5.5, we review these limitations in detail.

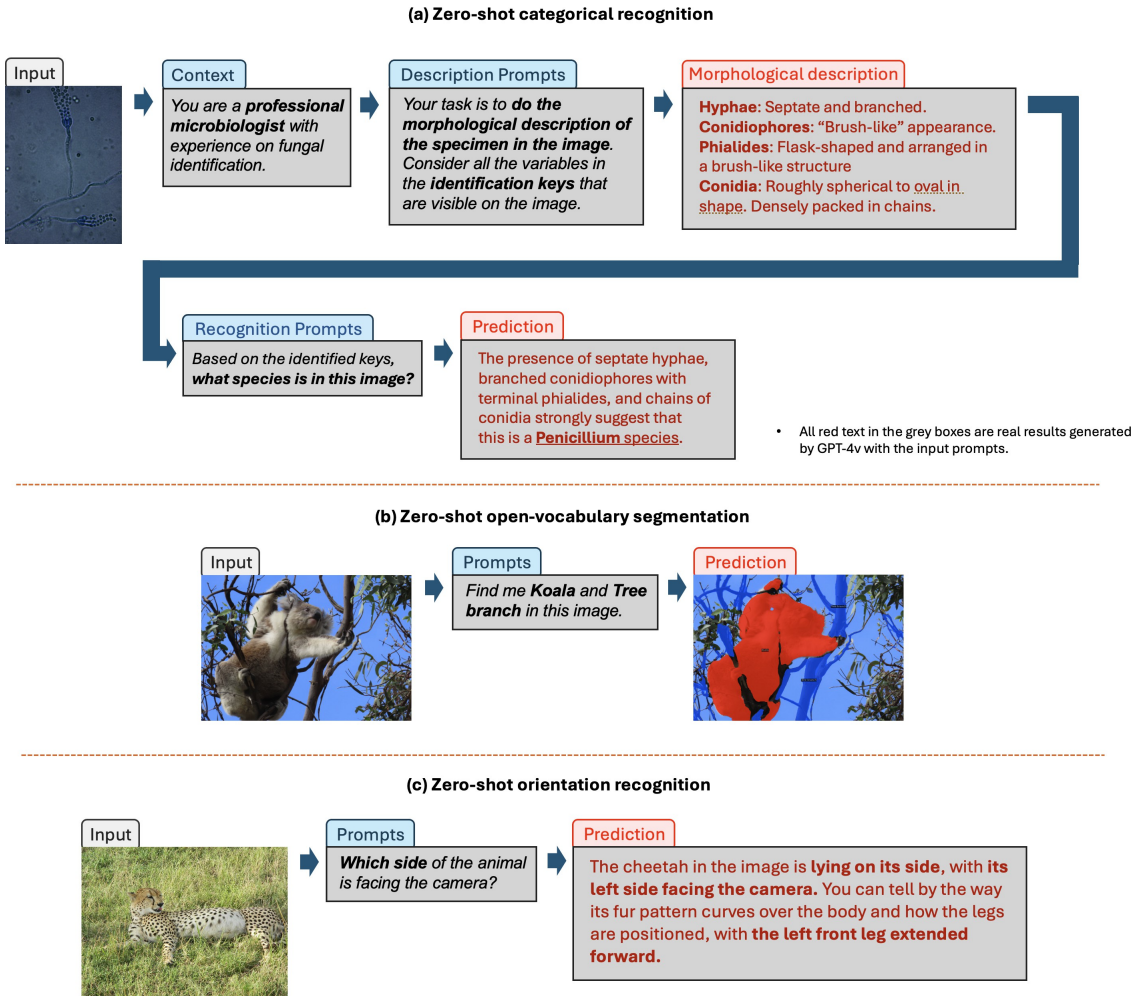

Figure 4: **Zero-shot task examples with Vision-Language models.** (a) In mycology, mycolo- gists typically utilize identification keys (*i.e.*, a series of macroscopic and microscopic morphological descriptions formatted as dichotomous keys) to systematically conduct taxonomic classification [54]. MML models can mimic these attribute-based classification processes and conduct zero- shot recognition without having previously seen such species. The recognition can be based on online information, such as that from Wikipedia, to which the MML models have access [53].
(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 MML models [55], 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 MML 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 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), (b), and (c) are generated by GPT-4v.

4.3 Enhanced model interpretability

While traditional models function as "black boxes", where researchers are unable to trace what features and mechanisms the models are using to make predictions, MML 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 [47]. In contrast, the shared perceptual-language feature space of multimodal models can provide venues for interpreting outputs directly in natural language and ultimately leading to a degree of insight into

the inner workings of the model. For example, Figure 3 shows a model providing detailed explanations on why it makes its classifications based on the input images and human text prompts. Similarly, the predictions in Figure 4 (a) also provide reasoning about why the model thinks the input image is a *Penicillium* species. In Figure 4 (c), 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 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, we can gain a clearer understanding of why models fail in certain cases—be it due to algorithmic failure or poor data quality. (Figure 6).

4.4 Learning with context

MML models have facilitated rapid progress in the field of compositional zero-shot learning (CZSL) [50]—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* in the training data [63]. CZSL can be further generalized to an area called *in-context learning*. In-context learning [52, 64], originally introduced for language tasks, is a technique for adapting a pre-trained model to novel and unseen tasks such as recognizing certain novel animal species without updating the model weights. 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 Figure 3 and 4 (a). Such contextual-based prediction and identification is not easily achievable through traditional machine learning methods, where the prediction outputs of models are usually within a predefined space. Traditional models usually need model update and fine-tuning to adapt to any new tasks.

In-context learning has the potential to be expanded to include broader scenarios and a wider range of information, thereby making the interactions between humans and AI more complex. Figure 5 provides a conceptual example of the application of in-context learning using MML models. In general, a multimodal model tends to provide generalized predictions to start, but with additional context provided, the model can adapt to the actual recognition tasks users want it to perform. In addition, if a model encounters difficulty or uncertainty during recognition, the introduction of additional context, such as habitat, can help refine the range of potential options, thereby augmenting the probability of accurate recognition. Of particular note is that this in-context learning process usually doesn't necessitate supplementary training or 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. To realize such complicated interactions between human and machines, a dedicated MML model might be necessary. In Section 5.5 and 5.6 we discuss the requirements and challenges achieving a practical state of MML model for biodiversity research and conservation in detail.

4.5 Natural language interaction

Since the advancement of MML models [65, 66], the interaction between humans and machines has become a prominent topic, especially in applied fields where practitioners often lack an engineering 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 Figure 3 and 4. Moreover, as mentioned in Section 3, a single well-trained MML model can handle most of the 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 [16, 66]. These studies aim to extend the capabilities of AI models to a broader range of tasks such as problem-solving and reasoning [67], complex image and video question answering [68, 69], and translation [70].

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 MML 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 MML models for biodiversity monitoring and conservation.

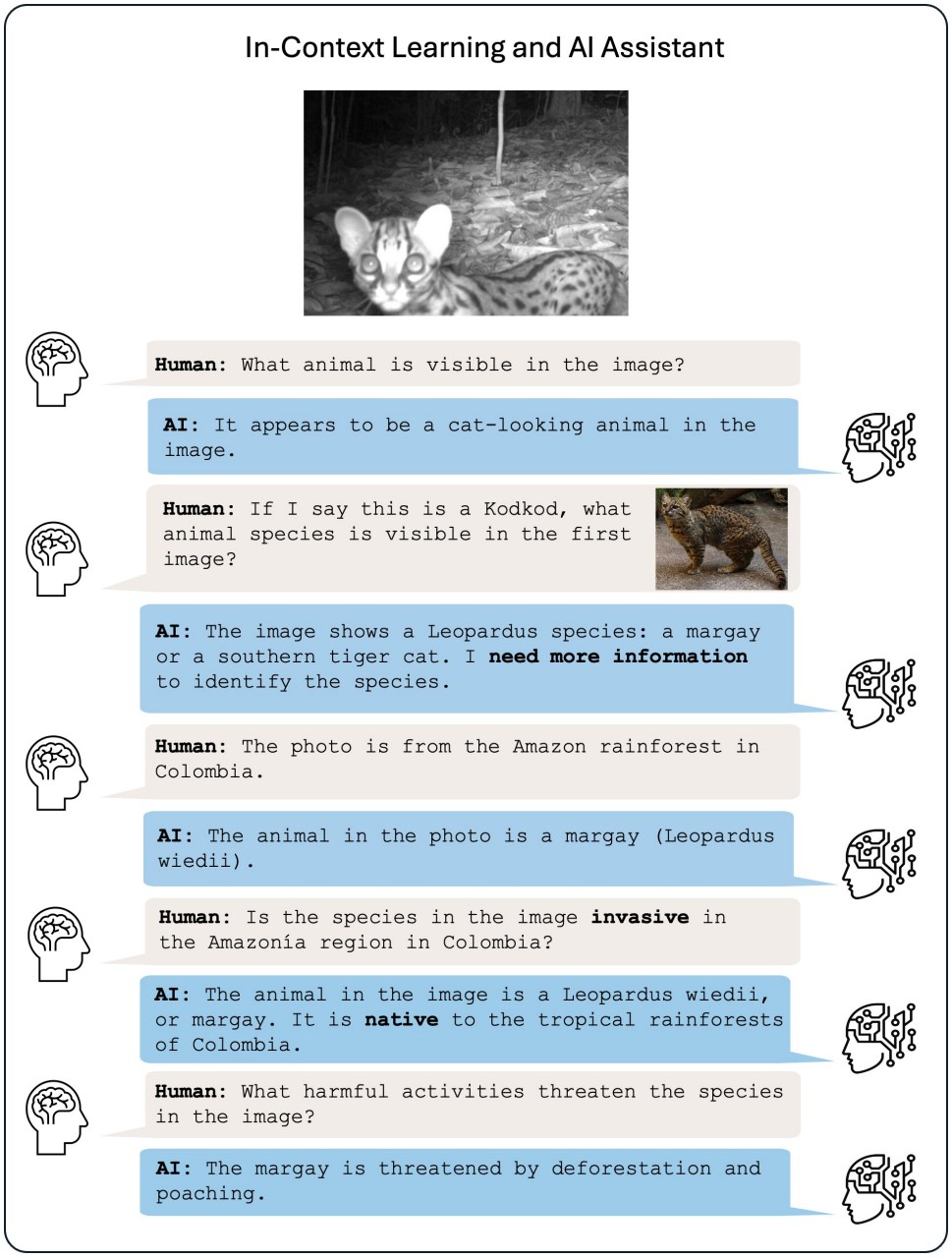

Figure 5: **MML models have the ability to adapt to specific domains, providing diverse outputs based on the user-provided context through in-context learning.** This process is not confined to contexts directly related to input samples and can be extended to various other scenarios. For example, it 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 [65], Microsoft Copilot, and Google Gemini [18].

5.1 Prompt engineering and consistent model performance

A distinct challenge inherent in MML models lies in the need for input data dependent on 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 [24, 25]. The choice of text prompts can significantly impact the generated outcomes [25]; some prompts may enhance the performance in target tasks like in-context learning [71] (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 produces high-quality predictions) [72]. 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 labor and time for animal recognition using MML 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 [73, 74, 75, 76, 77, 78, 53]. However, these methods are either in preliminary stages or carry their own set of limitations, such as the dependence on manually predefined visual attributes of objects [74] or inferior recognition performance compared to fully supervised models [53].

5.2 Language and terminology bias

The languages generated or used to train existing MML 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 MML models. However, accurately understanding such terminology and their connection to visual features in the image can be essential in recognizing bird species. Besides domain specific terminologies, existing MML models are largely trained in English [79, 16], 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 injecting 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 [61]. For instance, [53] 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 MML 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 [80]. Therefore, we still need to figure out how to effectively prepare data of sufficient quality and variety for conservation and ecology tasks.

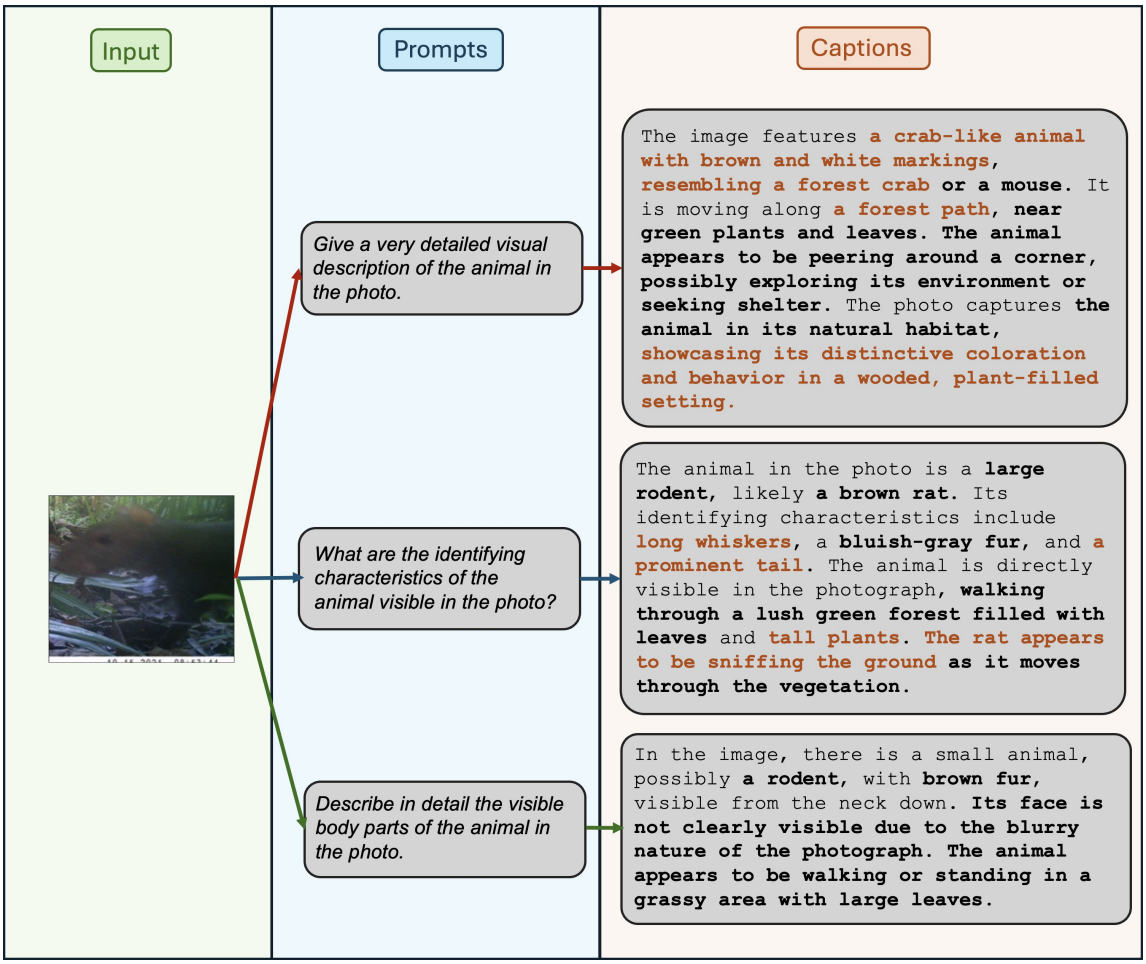

Figure 6: **Prompt engineering and model hallucination.** Prompts play a key role in the performance of MML models. For example, a standard LLaVA model produced three distinct captions for the same image using three slightly different prompts. Interestingly, the first caption identified a "crab-like animal" that didn't 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).

5.3 Systematic failures and hallucinations

Additionally, MML models can exhibit systematic failures [28], 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 [27, 28, 29, 30]. 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.

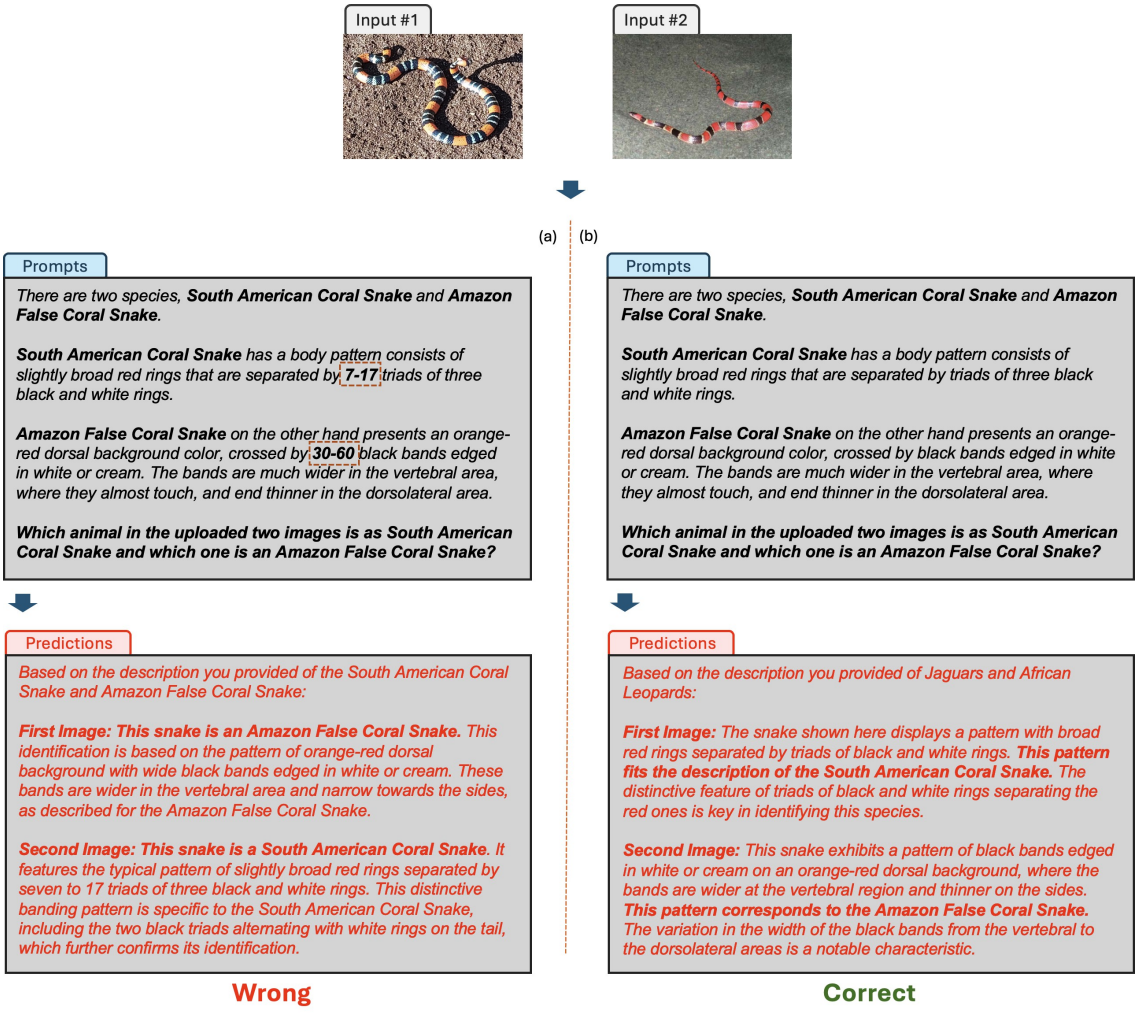

Figure 7: **Systematic failures with numbers.** The model can make totally opposite predictions when numbers are included in the input prompts, especially when these numbers are uncertain, such as the ranges provided in the prompts (7-17 and 30-60).

When it comes to generative tasks, including image captioning, it can eventually cause model halluci- nation (*e.g.*, models perceive non-existent content as existing due to various algorithmic idiosyncrasies). There is currently no effective way to control hallucinations except for manual oversight. Despite an increase in recent studies addressing the issue of hallucinations in large language models [81], research on hallucinations in MML models remains in the preliminary stages. Hallucinations often stem from a mismatch between different data modalities (*i.e.*, **embedding confusion**). For instance, a MML model might incorrectly respond with "yes" to queries like "Is X present in this image?" where X represents any animals or other objects, just like the "crab-like animal" the model predicts in Figure 6. Such hallucination can be massively detrimental to real-world applications, especially in tasks that require high precision, such as animal movement and habitat monitoring. [53] has shown potential of using instruction tuning methods and caption confidence scores to limit the caption hallucinations; however, how similar techniques can be applied in the real world remains to be studied.

5.4 The cost of model tuning

Known instances where the efficacy of MML 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 [82, 83, 15]. While MML models generally are better at generalizing to different data domains and distributions compared to conventional machine learning methods, pronounced domain discrepancies (*i.e.*, data

differences between domains) may still cause inconsistent performance and errors [61, 62]. Given that applications of AI in biodiversity and conservation often encompass domain-specific tasks that may exhibit substantial 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 [1, 4], 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 [84, 85], 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 computational budgets.

The substantial demand for computational resources is one of the key constraints of training and fine-tuning MML models. For instance, the Flamingo [17] model used 1,536 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) [66] particularly for models that uses LLMs as their language encoders like Flamingo [17] and GPT-4v [16]. For example, according to the pricing page of OpenAI (https://openai.com/pricing), GPT-4v costs $0.04 per 1000 tokens, 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 [86], parameter efficient tuning [85, 87], model distillation [88], and model compression [89] are such 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 [85, 87] 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) [90, 89].

From Figure 3 and 4, we can also see that a generalized MML model can already function within a wildlife context; therefore, a dedicated MML model trained for ecology and conservation from the ground up may not be necessary. However, research into the cost-efficient MML 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

Despite 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 [91]. These datasets, however, are predominantly designed for traditional sample-to-label based machine learning tasks. Datasets for MML 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 [92] 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. The descriptive information is crucial for the semantic alignment of MML models as it helps define and incorporate ecological and conservation context information into MML models. However, creating such datasets is no trivial task; it requires the collection and annotation of data across multiple modalities, 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 MML models for biodiversity and conservation- specific tasks and data. 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 MML models is largely dominated by closed-source algorithms and datasets [17, 26, 15, 66]. This approach significantly hampers the advancement of MML models and poses a barrier to scientific progress in various fields, especially when modifications to existing MML models are typically exclusive to contracted partners and paid services [16], making them less accessible to practitioners. Furthermore, these closed-source strategies inhibit researchers from fully grasping the underlying mechanisms of these algorithms, even through 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 [16], let alone make any structural and algorithmic modifications to the models. This lack of accessibility is one of the main reasons why studies such as [53] and [92] can only use less well-developed

MML 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 [93], Open- CLIP [94], LION [82] and BioCLIP [92] are commendable efforts to mitigate this challenge and afford developers more accessible methods, they unfortunately fall short of achieving the performance standards set by their closed-source counterparts [94]. Even if the cost of model training and accessibility of biodiveristy 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. However, more research is needed to provide additional evidence on whether the performance differences between open-source efforts and their closed-source counterparts matter in real-world fields like conservation AI.

6 Conclusion

In conclusion, MML models stand as a revolutionary advancement in AI, providing a robust and adaptable approach poised to transform the realms of biodiversity research and conservation. This innovative technique facilitates a multitude of tasks, including zero-shot learning, few-shot learning, domain generalization, and enhanced model interpretability. Moreover, it significantly improves the accessibility of AI for practitioners through natural-language-based human-machine interactions. However, the deployment of these multimodal models in ecological research and conservation practice in particular remains challenged by several barriers, including the need for extensive computational resources, the requirement for prompt engineering for consistent performance on large datasets, systematic model failure and hallucination, and insufficient open-source sharing of state-of-the-art methods. Moving forward, it is imperative to address these challenges through continued research and development. Efforts should focus on enhancing the computational efficiency of multimodal models, reducing their cost, and increasing their transparency to facilitate wider adoption and innovation. Ultimately, by overcoming these obstacles, AI can play a crucial role in biodiversity and conservation efforts worldwide, providing tools that are not only powerful and efficient but also equitable and accessible. It is our hope that this discussion will spur further research and collaboration across disciplines to realize the full potential of MML models in biodiversity research and conservation.

References

[1] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt *et al.*, "Perspectives in machine learning for wildlife conservation," *Nature communications*, vol. 13, no. 1, p. 792, 2022. 2, 8, 16

[2] R. Kwok *et al.*, "Ai empowers conservation biology," *Nature*, vol. 567, no. 7746, pp. 133–134, 2019. 2

[3] S. Beery, D. Morris, and S. Yang, "Efficient pipeline for camera trap image review," *arXiv preprint arXiv:1907.06772*, 2019. 2

[4] Z. Miao, Z. Liu, K. M. Gaynor, M. S. Palmer, S. X. Yu, and W. M. Getz, "Iterative human and automated identification of wildlife images," *Nature Machine Intelligence*, vol. 3, no. 10, 2021. 2, 8, 16

[5] B. Kellenberger, T. Veen, E. Folmer, and D. Tuia, "21 000 birds in 4.5 h: efficient large-scale seabird detection with machine learning," *Remote Sensing in Ecology and Conservation*, vol. 7, no. 3, pp. 445–460, 2021. 2

[6] Z. Miao, S. X. Yu, K. L. Landolt, M. D. Koneff, T. P. White, L. J. Fara, E. J. Hlavacek, B. A. Pickens, T. J. Harrison, and W. M. Getz, "Challenges and solutions for automated avian recognition in aerial imagery," *Remote Sensing in Ecology and Conservation*, vol. 9, no. 4, pp. 439–453, 2023. 2

[7] J. A. Ahumada, E. Fegraus, T. Birch, N. Flores, R. Kays, T. G. O'Brien, J. Palmer, S. Schuttler, J. Y. Zhao, W. Jetz *et al.*, "Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet," *Environmental Conservation*, vol. 47, no. 1, pp. 1–6, 2020. 2

[8] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021. 2

[9] R. Dodhia, *AI for Social Good: Using Artificial Intelligence to Save the World*. Wiley, 2024. [Online]. Available: https://books.google.com/books?id=O8D3EAAAQBAJ 2

[10] T. A. Rhinehart, L. M. Chronister, T. Devlin, and J. Kitzes, "Acoustic localization of terrestrial wildlife: Current practices and future opportunities," *Ecology and Evolution*, vol. 10, no. 13, pp. 6794–6818, 2020. 2

[11] M. Zhong, M. Torterotot, T. A. Branch, K. M. Stafford, J.-Y. Royer, R. Dodhia, and J. Lav- ista Ferres, "Detecting, classifying, and counting blue whale calls with siamese neural networks," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3086–3094, 2021. 2

[12] D. Stowell, M. D. Wood, H. Pamul-a, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019. 2

[13] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2537–2546. 2

[14] Z. Liu, Z. Miao, X. Pan, X. Zhan, D. Lin, S. X. Yu, and B. Gong, "Open compound domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 2, 3, 6, 16, 17

[16] OpenAI, "Gpt-4v(ision) system card," 2023. 2, 3, 11, 13, 16, 17

[17] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022. 2, 3, 6, 8, 16, 17

[18] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023. 2, 12

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. 2, 6

[20] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023. 2, 6

[21] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," 2023. 2

[22] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao, "Multimodal foundation models: From specialists to general-purpose assistants," *arXiv preprint arXiv:2309.10020*, vol. 1, no. 2, p. 2, 2023. 2

[23] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018. 2

[24] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 816–16 825. 2, 13

[25] ——, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022. 2, 13

[26] OpenAI, :, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Bal- tescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. But- ton, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. M´ely, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph, "Gpt-4 technical report," 2023. 2, 4, 17

[27] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," *arXiv preprint arXiv:2401.06209*, 2024. 2, 15

[28] S. Tong, E. Jones, and J. Steinhardt, "Mass-producing failures of multimodal systems with language models," *arXiv preprint arXiv:2306.12105*, 2023. 2, 14, 15

[29] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "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*, 2022. 2, 15

[30] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, "Winoground: Probing vision and language models for visio-linguistic compositionality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5238–5248. 2, 15

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 3, 4

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 3

[33] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190. 3

[34] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980. 3

[35] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4563–4567. 3

[36] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," 2023. 3, 6

[37] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang, "Generative pretraining in multimodality," *arXiv preprint arXiv:2307.05222*, 2023. 3

[38] L. Yu, B. Shi, R. Pasunuru, B. Muller, O. Golovneva, T. Wang, A. Babu, B. Tang, B. Karrer, S. Sheynin *et al.*, "Scaling autoregressive multi-modal models: Pretraining and instruction tuning," *arXiv preprint arXiv:2309.02591*, 2023. 3

[39] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695. 3

[40] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "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*, 2021, pp. 2554–2562. 4

[41] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021. 4

[42] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," *arXiv preprint arXiv:1703.04105*, 2017. 4

[43] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozi`ere, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 4

[44] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023. 4

[45] Q. Chen, Q. Liu, and E. Lin, "A knowledge-guide hierarchical learning method for long-tailed image classification," *Neurocomputing*, vol. 459, pp. 408–418, 2021. 4

[46] S. Beery, D. Morris, and P. Perona, "The iwildcam 2019 challenge dataset," *arXiv preprint arXiv:1907.07617*, 2019. 4

[47] Z. Miao, K. M. Gaynor, J. Wang, Z. Liu, O. Muellerklein, M. S. Norouzzadeh, A. McInturff, R. C. Bowie, R. Nathan, S. X. Yu *et al.*, "Insights and approaches using deep learning to classify wildlife," *Scientific reports*, vol. 9, no. 1, p. 8137, 2019. 5, 6, 8, 10

[48] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," *arXiv preprint arXiv:2206.04769*, 2022. 6

[49] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," *arXiv preprint arXiv:1704.01444*, 2017. 6

[50] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1792–1801. 6, 10

[51] B. Li, Y. Shen, J. Yang, Y. Wang, J. Ren, T. Che, J. Zhang, and Z. Liu, "Sparse mixture-of-experts are domain generalizable learners," *arXiv preprint arXiv:2206.04046*, 2022. 6

[52] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. 6, 10

[53] Z. Fabian, Z. Miao, C. Li, Y. Zhang, Z. Liu, A. Hernández, A. Montes-Rojas, R. Escucha, L. Siabatto, A. Link *et al.*, "Multimodal foundation models for zero-shot animal species recognition in camera trap images," *arXiv preprint arXiv:2311.01064*, 2023. 8, 9, 13, 16, 17

[54] C. To-Anun, I. Hidayat, J. Meeboon *et al.*, "Genus cercospora in thailand: taxonomy and phylogeny (with a dichotomous key to species)," *Plant Pathology & Quarantine*, vol. 1, no. 1, pp. 11–87, 2011. 9

[55] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open- vocabulary semantic segmentation with mask-adapted clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7061–7070. 8, 9

[56] X. Wu, F. Zhu, R. Zhao, and H. Li, "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*, 2023, pp. 7031–7040. 8

[57] B. Jiao, L. Liu, L. Gao, R. Wu, G. Lin, P. Wang, and Y. Zhang, "Toward re-identifying any animal," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 8

[58] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021. 8

[59] J. Sherman, M. Ancrenaz, and E. Meijaard, "Shifting apes: Conservation and welfare outcomes of bornean orangutan rescue and release in kalimantan, indonesia," *Journal for Nature Conservation*, vol. 55, p. 125807, 2020. 8

[60] Z. Miao, B. Elizalde, S. Deshmukh, J. Kitzes, H. Wang, R. Dodhia, and J. M. L. Ferres, "Zero-shot transfer for wildlife bioacoustics detection," 2023. 10

[61] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 10, 13, 16

[62] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, "Solving olympiad geometry without human demonstrations," *Nature*, vol. 625, no. 7995, pp. 476–482, 2024. 10, 16

[63] W. Bao, L. Chen, H. Huang, and Y. Kong, "Prompting language-informed distribution for compositional zero-shot learning," *arXiv preprint arXiv:2305.14428*, 2023. 10

[64] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. 10

[65] OpenAI, "Introducing chatgpt." 2023. 11, 12

[66] G. team, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024. 11, 16, 17

[67] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021. 11

[68] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," *arXiv preprint arXiv:2311.16502*, 2023. 11

[69] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 11

[70] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," *arXiv preprint arXiv:2007.10310*, 2020. 11

[71] Y. Zhang, K. Zhou, and Z. Liu, "What makes good examples for visual in-context learning?" *Advances in Neural Information Processing Systems*, vol. 36, 2024. 13

[72] D.-W. Zhou, H.-L. Sun, J. Ning, H.-J. Ye, and D.-C. Zhan, "Continual learning with pre-trained models: A survey," *arXiv preprint arXiv:2401.16386*, 2024. 13

[73] J. M. Kim, A. Koepke, C. Schmid, and Z. Akata, "Exposing and mitigating spurious correlations for cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2584–2594. 13

[74] S. Menon and C. Vondrick, "Visual classification via description from large language models," *ICLR*, 2023. 13

[75] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505. 13

[76] K. Roth, J. M. Kim, A. S. Koepke, O. Vinyals, C. Schmid, and Z. Akata, "Waffling around for performance: Visual classification with random words and broad concepts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 15 746–15 757. 13

[77] A. Yan, Y. Wang, Y. Zhong, C. Dong, Z. He, Y. Lu, W. Y. Wang, J. Shang, and J. McAuley, "Learning concise and descriptive attributes for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3090–3100. 13

[78] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar, "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*, 2023, pp. 19 187–19 197. 13

[79] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozi`ere, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 13

[80] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu *et al.*, "Lima: Less is more for alignment," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 13

[81] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. 16

[82] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021. 16, 17

[83] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford alpaca, 2023. 16

[84] H. Liu, K. Son, J. Yang, C. Liu, J. Gao, Y. J. Lee, and C. Li, "Learning customized visual models with retrieval-augmented knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 148–15 158. 16

[85] Y. Zhang, K. Zhou, and Z. Liu, "Neural prompt search," 2022. 16

[86] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799. 16

[87] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727. 16

[88] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A survey on model compression for large language models," *arXiv preprint arXiv:2308.07633*, 2023. 16

[89] D. Kuznedelev, S. Tabesh, K. Noorbakhsh, E. Frantar, S. Beery, E. Kurtic, and D. Alistarh, "Vision models can be efficiently specialized via few-shot task-aware compression," *arXiv preprint arXiv:2303.14409*, 2023. 16

[90] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021. 16

[91] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8769–8778. 17

[92] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf *et al.*, "Bioclip: A vision foundation model for the tree of life," *arXiv preprint arXiv:2311.18803*, 2023. 17

[93] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023. 17

[94] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," Jul. 2021, if you use this software, please cite it as below. [Online]. Available: https://doi.org/10.5281/zenodo.5143773 17