VLM4Bio: A Benchmark Dataset to Evaluate Pretrained Vision-Language Models for Trait Discovery from Biological Images

M. Maruf , Arka Daw , Kazi Sajeed Mehrab , Harish Babu Manogaran , Abhilash Neog , Medha Sawhney , Mridul Khurana , James P. Balhoff , Yasin Bakış , Bahadir Altintas , Matthew J Thompson , Elizabeth G Campolongo , Josef C. Uyeda , Hilmar Lapp , Henry L. Bart Jr. , Paula M. Mabee , Yu Su , Wei-Lun Chao , Charles Stewart , Tanya Berger-Wolf , Wasila Dahdul , and Anuj Karpatne

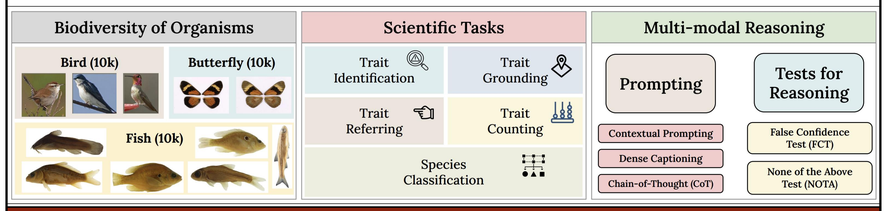
Scan for the Paper

Motivation

Images are increasingly becoming the currency for documenting biodiversity on the planet, providing novel opportunities for accelerating scientific discoveries in the field of organismal biology, especially with the advent of **large vision-language models (VLMs).** We ask if pre-trained VLMs can aid scientists in answering a range of biologically relevant questions *without any additional fine-tuning*.

Challenge: Understanding scientific images requires knowledge of domain-specific terminologies and reasoning that are not fully represented in conventional image datasets used for training VLMs.

- In this work, we evaluate the effectiveness of 12 state-of-the-art (SOTA) VLMs in the field of organismal biology using a novel dataset, **VLM4Bio**, consisting of **469K question-answer pairs** involving **30K images** from three groups of organisms: fishes, birds, and butterflies, covering five biologically relevant tasks.
- We also explore the effects of applying prompting techniques and tests for reasoning hallucination on the performance of VLMs, shedding new light on the capabilities of current SOTA VLMs in answering biologically relevant questions using images.



Organism Datasets

- We used image collections of three taxonomic groups of organisms: **Fish (containing ~10k images), Birds (containing ~10k images), and Butterflies (containing ~10k images),** obtained by taking subsets of the FishAIR dataset, the CUB dataset, and the Cambridge Butterfly dataset.
- We leveraged expert annotations of biologists to generate the ground-truth data that comprises approximately 469k question-answer pairs for the ~30k biological images across all tasks.

Statistics	Fish-10K	Bird-10K	Butterfly-10K	Fish-500	Bird-500
# Images	10,347	11,092	10,013	500	492
# Species	495	188	60	60	47
# Genera	178	114	27	18	33
# Traits	10	28	-	8	5

Table 1: Key statistics of the **VLM4Bio** dataset.

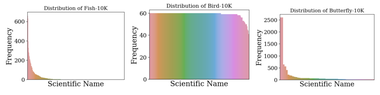
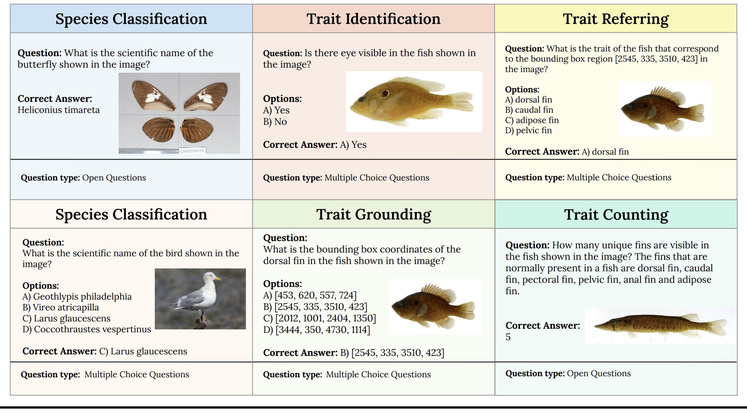


Figure 5: Dataset Distribution of Fish-10K, Bird-10K, and Butterfly-10K.

Fish Traits		Bird Traits			
		Color		Pattern	Measurements
. Eye . Head . Mouth . Barbel . Dorsal fin	. Pectoral fin . Pelvic fin . Anal fin . Two dorsal fins . Adipose fin	. Bill-color . Crown-color . Eye-color . Forehead-color . Nape-color . Primary-color . Throat-color . Back-color	. Belly-color . Breast-color . Leg-color . Under-tail-color . Underparts-color . Upper-tail-color . Upperparts-color . Wing-color	. Head-pattern . Back-pattern . Breast-pattern . Wing-pattern . Tail-pattern . Belly-pattern	. Bill-length . Bill-shape . Shape . Size . Tail-shape . Wing-shape

Figure 6: Trait list for Trait Identification task.

Scientific Tasks



Prompts

Task	Prompt Format
Species Classification	*<image>* What is the scientific name of the *<organism>* shown in the image? *<options>* Write the answer after writing the answer is: .
Trait Identification	*<image>* Is there *<trait>* visible in the *<organism>* shown in the image? *<options>* Write the answer after writing the answer is: .
Trait Grounding	*<image>* What is the bounding box coordinates of the *<trait>* in the fish shown in the image? *<options>* Write the answer after writing the answer is: .
Trait Referring	*<image>* What is the trait of the *<organism>* that corresponds to the bounding box region *<coordinates>* in the image? *<options>* Write the answer after writing the answer is: .
Trait Counting	*<image>* How many unique *<trait>* are visible in the *<organism>* shown in the image? *<options>* Write the answer after writing the answer is: .
Contextual Prompting	*<image>* Each biological species has a unique scientific name composed of two parts: the first for the genus and the second for the species within that genus. What is the scientific name of the *<organism>* shown in the image? *<options>* Write the answer after writing the answer is: .
Dense Caption Prompting	*<image> <dense caption>.* Use the above dense caption and the image to answer the following question. What is the scientific name of the *<organism>* shown in the image? *<options>* Write the answer after writing the answer is: .
Chain-of-Thought Prompting	*<image>* What is the scientific name of the *<organism>* shown in the image? *<options>* Please consider the following reasoning to formulate your answer. *<reasoning>*. Write the answer after writing the answer is: .
False Confidence Test (FCT)	*<image>* What is the scientific name of the *<organism>* shown in the image? *<options>* Chosen Answer: *<suggested answer>*. Please provide: 1) Whether the chosen answer is correct (True/False). 2) The correct answer.
None of the Above Test (NOTA)	*<image>* What is the scientific name of the *<organism>* shown in the image? *<options: A) _ B) _ C) _ D) None of the above.>* Write the answer after writing the answer is: .

Results

Dataset	Question type	gpt-4v	llava v1.5-7b	llava v1.5-13b	cogvlm chat	BLIP flan-xl	BLIP flan-xxl	minigpt4 vicuna-7B	minigpt4 vicuna-13B	instruct flant5xl	instruct flant5xxl	instruct vicuna7B	instruct vicuna13B	Random Choice
Species Classification														
Fish-10K	Open	1.01	2.32	0.40	0.11	0.01	1.59	0.50	0.38	0.00	1.46	0.00	0.00	0.20
	MC	35.91	40.20	32.27	31.72	29.76	33.36	29.02	27.45	30.86	31.70	27.27	26.57	25.00
Bird-10K	Open	17.40	1.45	2.06	0.86	0.00	0.57	2.80	2.56	0.00	0.50	0.07	0.00	0.53
	MC	82.58	50.32	55.36	44.73	33.68	34.75	23.95	27.62	36.36	35.83	44.00	46.55	25.00
Butterfly-10K	Open	0.04	0.05	0.00	0.01	0.00	0.00	0.07	0.01	0.00	0.00	9.94	0.00	1.54
	MC	28.91	50.24	44.58	36.45	25.14	28.88	33.06	28.90	25.28	36.67	41.70	34.48	25.00
Trait Identification														
Fish-10K	MC	82.18	56.84	45.15	46.92	68.36	39.33	55.08	51.87	64.34	39.26	81.95	20.69	50.0
Bird-10K	MC	62.22	34.68	46.14	63.93	50.11	41.38	39.11	40.44	47.89	45.52	77.91	89.98	31.12
Trait Grounding														
Fish-500	MC	29.41	24.87	17.98	23.42	23.32	25.14	22.18	25.58	7.20	27.09	33.51	26.90	25.00
Bird-500	MC	8.1	26.92	35.36	23.2	11.83	10.52	15.39	24.22	3.48	0.81	30.24	13.91	25.00
Trait Referring														
Fish-500	MC	28.15	27.07	29.14	28.19	24.93	25.68	39.24	31.21	31.75	25.78	28.04	32.73	25.00
Bird-500	MC	42.28	30.5	29.64	18.45	35.16	40.59	26.04	35.88	27.52	41.69	23.03	22.69	25.00
Trait Counting														
Fish-500	Open	16.4	47.4	52.0	14.8	37.6	63.4	13.6	31.53	50.2	61.4	61.4	0.0	25.00
	MC	44.80	13.20	54.80	21.00	64.8	78.2	22.00	25.00	74.0	69.4	15.80	11.80	25.00
Overall		34.24	29.0	31.78	25.27	28.91	30.24	23.0	25.19	28.49	29.79	33.92	23.31	22.03

Table 2: Zero-shot accuracy comparison of VLM baselines (in % ranging from 0 to 100) for the five scientific tasks. Results are color-coded as Best , Second best , Worst , Second worst .

Dataset	Difficulty	gpt-4v	gpt-4o	llava v1.5-7b	llava v1.5-13b	cogvlm chat	BLIP flan-xl	BLIP flan-xxl	minigpt4 vicuna-7B	minigpt4 vicuna-13B	instruct flant5xl	instruct flant5xxl	instruct vicuna7B	instruct vicuna13B	CLIP	BioCLIP
Fish	Easy	44.50	37.50	47.50	46.00	24.00	34.00	27.50	29.00	19.50	32.00	28.00	33.50	33.50	36.50	55.50
	Medium	3.50	5.50	30.00	28.50	27.00	26.00	23.00	26.50	25.00	28.50	24.50	26.00	25.50	26.00	29.00
Bird	Easy	73.50	68.00	53.50	50.00	38.50	34.50	36.00	21.00	32.00	41.00	33.00	43.50	39.00	57.00	94.00
	Medium	41.00	40.50	30.50	37.00	30.00	25.50	21.00	21.00	24.00	27.00	27.00	24.50	26.50	31.00	95.00
Butterfly	Easy	18.50	17.50	19.00	20.50	24.50	30.00	25.00	34.50	26.00	24.50	22.50	19.00	24.50	21.50	65.50
	Medium	5.50	7.00	29.50	29.00	29.50	20.00	25.50	33.00	25.00	27.50	25.00	25.00	25.00	21.50	58.00
	Hard	2.00	1.50	22.00	21.00	32.00	26.50	20.00	29.50	24.00	22.50	24.00	24.00	21.00	21.50	35.00

Dataset	Prompting	gpt-4v	gpt-4o	llava v1.5-7b	llava v1.5-13b	cogvlm chat	BLIP flan-xl	BLIP flan-xxl
Fish-Prompting	No Prompting	34.40	79.00	41.60	35.40	31.00	28.60	22.60
	Contextual	30.00	77.20	40.20	35.60	25.60	27.20	26.60
	Dense Caption	18.80	78.60	26.00	27.60	32.00	28.40	29.80
	CoT	42.60	86.00	41.40	34.80	26.80	29.20	24.60
Bird-Prompting	No Prompting	78.80	97.60	44.20	49.80	45.40	35.60	35.80
	Contextual	78.60	98.60	44.00	52.00	49.40	35.60	30.40
	Dense Caption	87.40	97.00	33.40	41.00	44.00	25.60	22.80
	CoT	62.60	98.60	37.40	47.80	42.20	30.60	31.00
Butterfly-Prompting	No Prompting	13.20	56.40	27.20	26.80	25.60	24.40	21.20
	Contextual	9.20	56.20	26.00	24.60	27.20	23.60	24.60
	Dense Caption	49.60	63.20	25.20	23.80	27.00	23.20	23.20
	CoT	63.60	74.60	21.40	23.20	34.60	37.20	23.60

Table 4: Zero-shot accuracy comparison for different prompting techniques of seven VLMs (in % ranging from 0 to 100). Results are color-coded as Best and Worst .

Dataset	Question type	gpt-4v	llava v1.5-7b	cogvlm chat	CLIP	BioCLIP
Species Classification						
Fish-10K	Open	1.01	2.32	0.11	0.57	1.24
	MC	35.91	40.20	31.72	42.45	50.65
Bird-10K	Open	17.40	1.45	0.86	7.74	67.12
	MC	82.58	50.32	44.73	45.78	93.93
Butterfly-10K	Open	0.04	0.05	0.01	5.33	15.95
	MC	28.91	50.24	36.45	45.60	62.32

Table 7: Zero-shot accuracy comparison of VLM baselines (in % ranging from 0 to 100) with BioCLIP for the species classification task. Results are color-coded as Best , and Worst .

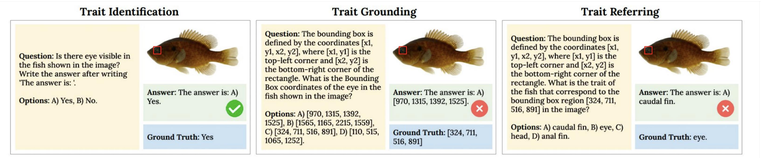


Figure 3: Examples of correct and incorrect predictions of GPT-4V for trait identification, trait grounding, and trait-referring tasks related to the "eye". For visualization assistance, a red-colored bounding box is added around the "eye" in the image.

Dataset	Metrics	Models: gpt-4v	gpt-4o	llava v1.5-7b	llava v1.5-13b	cogvlm chat	BLIP flan-xl	BLIP flan-xxl
False Confidence Test (FCT)								
Fish-Prompting	Accuracy	34.20	73.60	25.00	28.60	24.60	0.00	7.00
	Agreement Score	4.40	16.60	99.80	19.20	74.40	0.00	28.4
Bird-Prompting	Accuracy	73.40	99.00	25.40	35.80	19.80	0.00	20.20
	Agreement Score	11.40	21.00	93.20	17.80	47.80	0.00	79.80
Butterfly-Prompting	Accuracy	5.20	53.40	27.20	26.60	6.20	0.00	5.00
	Agreement Score	2.60	12.40	95.40	5.60	13.80	0.00	19.00
None of the Above (NOTA) Test								
Fish-Prompting	Accuracy	81.40	44.80	3.40	3.80	0.00	4.00	0.00
Bird-Prompting	Accuracy	75.00	91.40	1.00	1.20	0.00	31.40	0.00
Butterfly-Prompting	Accuracy	50.40	4.60	1.00	4.60	0.00	51.00	0.00

Table 5: Performance of seven VLMs on the NOTA and FCT reasoning tests. Results are color-coded as Best and Worst .

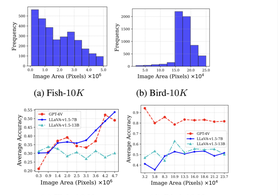


Figure 9: Distribution of image resolutions for Fish-10K and Bird-10K are shown in Figures (a) and (b), respectively. The average score over image resolution for the GPT-4V, LLaVA-v1.5-7B, and LLaVA-v1.5-13B models on Fish-10K and Bird-10K are presented in Figures (c) and (d). We conduct the experiment in the context of the Species Classification task with Multiple-Choice (MC) questions.

Key Findings

- All VLMs show poor accuracy on open questions but perform better on MC questions.
- The Bird dataset shows better accuracy than the Fish or Butterfly dataset.
- VLMs struggle to localize traits in images.
- Counting biological traits is difficult for VLMs.
- The pretrained VLMs generally perform best on the easy set and worst on the hard set for each organism.
- By comparing BioCLIP with CLIP, we can see that finetuning foundation models with biological data provide large gains in classification performance.
- From our prompting experiments, providing extra context and caption is more useful for GPT-4V and GPT-4o than the smaller models.
- GPT-4V often responded by apologetic expressions, admissions of an inability to visualize the organism precisely, and disclaimers regarding prediction without sufficient expert data and guidance.
- Image resolution influences the VLM performance for the Fish-10K dataset since higher resolution helps recognize the details of the biological traits and correct species.

Acknowledgments

This research is supported by National Science Foundation (NSF) awards for the HDR Imageomics Institute (OAC-2118240). We are thankful for the support of computational resources provided by the Advanced Research Computing (ARC) Center at Virginia Tech. This poster has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains, and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript or allow others to do so for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan).