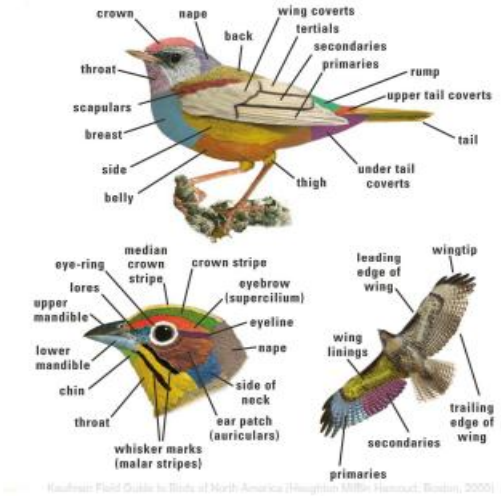
VLM4Bio: A Benchmark Dataset to Evaluate Pretrained Vision-Language Models for Trait Discovery from Biological Images

Conference - NeurIPS 2024

https://proceedings.neurips.cc/paper_files/paper/2024/file/eced4a5fbc776e81b45e2f72447f0164-Paper-Datasets_and_Benchmarks_Track.pdf

Problem Statement

- Large collections of organism images exist from museums, universities, and citizen science efforts.
- Biologists aim to extract traits from images, but manual methods are slow and labor-intensive.
- Pre-trained VLMs can handle text-image tasks, making them promising for biological research.
- It is unclear **if VLMs contain enough scientific knowledge to answer biology-related questions accurately.**
- Assessing SOTA VLMs with the VLM4Bio dataset is needed to test their effectiveness, prompting strategies, and reasoning limitations.



Traits are the externally visible characteristics of an organism that define the species. Traits can be certain regions, colors, patterns, or landmark points of the organism.

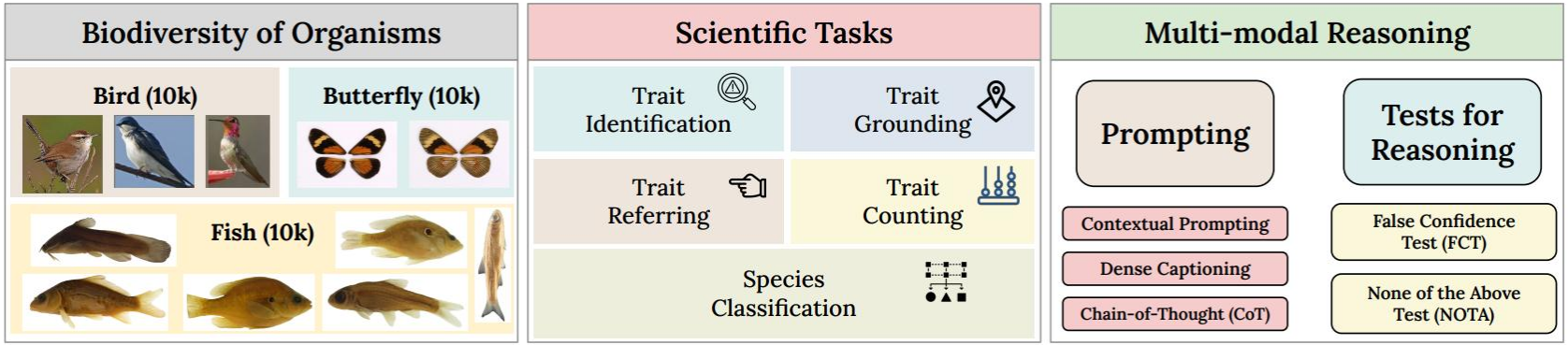
Why this paper

Novelty: It is the first benchmark dataset specifically for evaluating VLMs in organismal biology, covering a range of biologically relevant tasks.

Social Good: It enables automated biodiversity monitoring, helps scientists analyze species traits, and could help research on the impact of environmental changes on organisms.

Approach

- Creation of VLM4Bio dataset that includes **469K question-answer pairs** from 30K images of fish, birds, and butterflies.
- **Evaluating 12 VLMs** on five tasks: species classification, trait identification, trait grounding, trait referring, and trait counting.



Species Classification

Question: What is the scientific name of the butterfly shown in the image?

Correct Answer:
Heliconius timareta

Question type: Open Questions

Trait Identification

Question: Is there eye visible in the fish shown in the image?

Options:
A) Yes
B) No

Correct Answer: A) Yes

Question type: Multiple Choice Questions

Trait Referring

Question: What is the trait of the fish that correspond to the bounding box region [2545, 335, 3510, 423] in the image?

Options:
A) dorsal fin
B) caudal fin
C) adipose fin
D) pelvic fin

Correct Answer: A) dorsal fin

Question type: Multiple Choice Questions

Species Classification

Question:
What is the scientific name of the bird shown in the image?

Options:
A) Geothlypis philadelphia
B) Vireo atricapilla
C) Larus glaucescens
D) Coccothraustes vespertinus

Correct Answer: C) Larus glaucescens

Question type: Multiple Choice Questions

Trait Grounding

Question:
What is the bounding box coordinates of the dorsal fin in the fish shown in the image?

Options:
A) [453, 620, 557, 724]
B) [2545, 335, 3510, 423]
C) [2012, 1001, 2404, 1350]
D) [3444, 350, 4730, 1114]

Correct Answer: B) [2545, 335, 3510, 423]

Question type: Multiple Choice Questions

Trait Counting

Question: How many unique fins are visible in the fish shown in the image? The fins that are normally present in a fish are dorsal fin, caudal fin, pectoral fin, pelvic fin, anal fin and adipose fin.

Correct Answer:
5

Question type: Open Questions

Results

Dataset	Difficulty	Models														
		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

Table 3: Zero-Shot accuracy comparison for *easy, medium, and hard* datasets. Results are color-coded as Best, Second best, Worst, Second worst.

Results

Dataset	Prompting	Models						
		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 .

Three prompting techniques

1. Contextual Prompting:

- Provided a single-line description of the tasks with the question.
- For example, for species classification task:

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.

2. Dense Caption Prompting:

- Prompt the VLM to generate a dense caption for the specimen image.
- Then add the dense caption before the question and prompt, “Use the above dense caption and the image to answer the following question.” to generate responses.

3. Chain-of-Thought (CoT) Prompting:

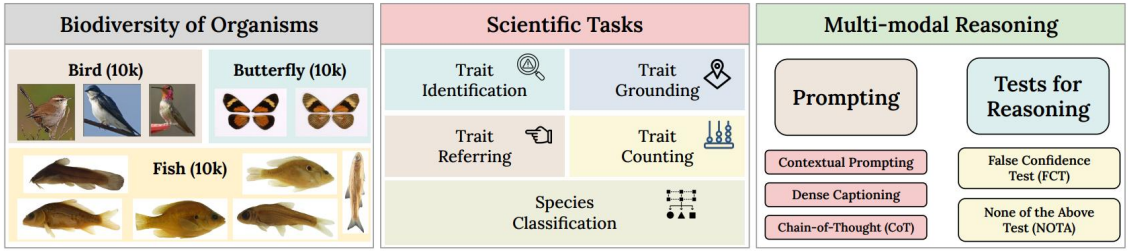
- Prompt “Let’s think step by step” to the VLM to generate the reasoning for a given VQA and multiple choices.
- Then add the reasoning after the VQA and prompt, “Please consider the following reasoning to formulate your answer.” to generate the VLM response.

Results

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

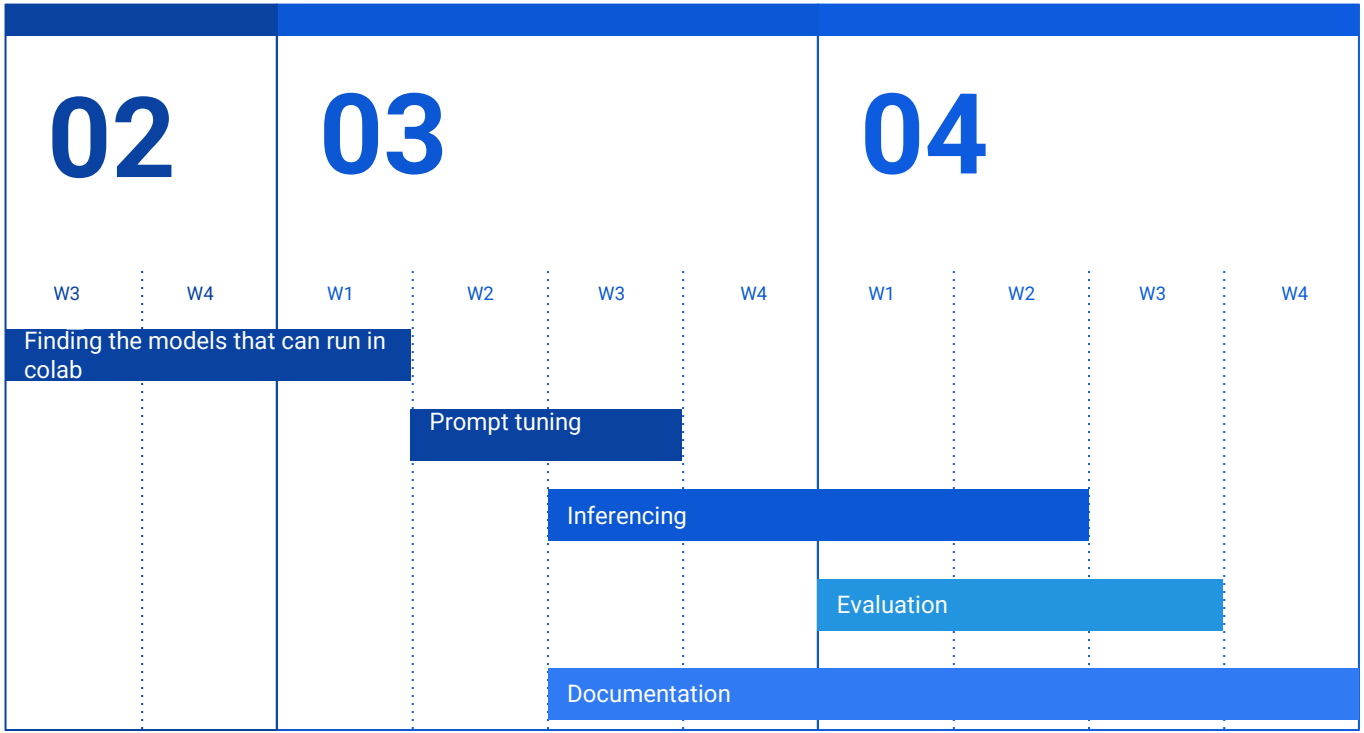
Table 5: Performance of seven VLMs on the NOTA and FCT reasoning tests. Results are color-coded as Best and Worst.

Feasibility Analysis



- The dataset and it's meta data are available - https://huggingface.co/datasets/imageomics/VLM4Bio
- Instead of using large-scale models like GPT-4V, We will test with a smaller VLM, this will give the idea of the **performance trade-offs**.
  - Medium sized model
  - Quantized large model
- Experimenting with different prompting techniques to improve the model reasoning.
- The code for evaluation process is available - https://github.com/imageomics/VLM4Bio

Gantt Chart


02
03
04
W3
W4
W1
W2
W3
W4
W1
W2
W3
W4
Finding the models that can run in colab
Prompt tuning
Inferencing
Evaluation
Documentation

thankyou!

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

Conference - NeurIPS 2024

Presented by: Aamod, Aryaman, Diya, Gaurav, Isha

Problem Statement

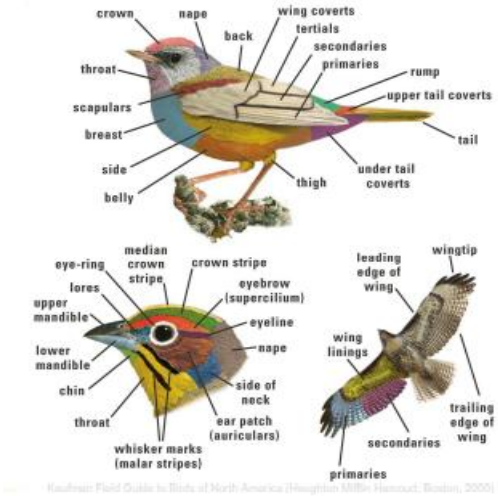
– Large collections of organism images exist from museums, universities, and citizen science efforts.

– Biologists aim to extract traits from images, but manual methods are slow and labor-intensive.

– Pre-trained VLMs can handle text-image tasks, making them promising for biological research.

– It is unclear if VLMs contain enough scientific knowledge to answer biology-related questions accurately.

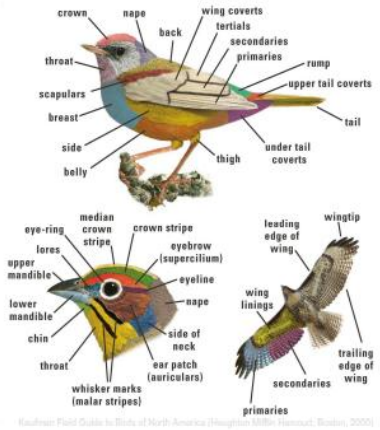
– Assessing SOTA VLMs with the VLM4Bio dataset is needed to test their effectiveness, prompting strategies, and reasoning limitations.



Traits are the externally visible characteristics of an organism that define the species. Traits can be certain regions, colors, patterns, or landmark points of the organism.

Problem Statement

- Large repositories of organism images are available.
  - Museum and university library collection.
  - Citizen science data.
- Biologists are interested in discovering biological traits directly from the organism's images.
- Large Vision-Language Models (VLMs) can solve a diverse range of tasks involving text and images.
- **Do pre-trained VLMs contain the necessary scientific knowledge to help biologists in answering a variety of questions related to the discovery of biological traits from images?**



Traits are the externally visible characteristics of an organism that define the species. Traits can be certain regions, colors, patterns, or landmark points of the organism.

Novelty

1. **Comprehensive Evaluation of Pretrained VLMs**
   - Tests **12 state-of-the-art models** (GPT-4V, LLaVA, MiniGPT-4) on **zero-shot** biodiversity tasks.
   - Assesses both **predictive accuracy** and **reasoning ability** in scientific applications.
2. **Unique Dataset & Tasks**
   - **VLM4Bio** includes **469,000 Q&A pairs** across **30,000 images** from three taxonomic groups (fish, birds, butterflies).
   - Covers **five key biological tasks**: species classification, trait identification, trait grounding, trait referring, and trait counting.
3. **New Reasoning Tests for AI Models**
   - Introduces **False Confidence Test (FCT)** and **None of the Above Test (NOTA)** to detect **hallucinations** in AI-generated scientific answers.
   - Studies the effect of **advanced prompting techniques** (Contextual, Dense Captioning, Chain-of-Thought).

Social Good

1. **Accelerating Biodiversity Research**
   - Reduces the **manual effort** needed for species identification and trait analysis.
   - Enables **faster scientific discovery** in organismal biology and ecology.
2. **Enhancing Conservation Efforts**
   - Helps monitor **species populations and ecological changes** due to climate change.
   - Supports **automated biodiversity tracking** with AI-powered image analysis.
3. **Democratizing Scientific Knowledge**
   - Open-source dataset and benchmarks help **scientists, researchers, and policymakers** use AI for environmental sustainability.
   - Supports the **global fight against biodiversity loss** with AI-driven solutions.

Approach

VLM4Bio Dataset:

- **30K** images from Fish, Birds, and Butterflies.
- **469K QA** pairs across **5 tasks**
- Annotation Strategy:
  - **Automated:** Trait matrices and metadata for large datasets **(Fish-10K, Bird-10K)**
  - **Manual:** Bounding box annotations for smaller subsets (Fish-500, Bird-500)

Prompting Techniques

- Contextual Prompting: Adds task-specific context
- Dense Captioning: Generates detailed image descriptions as input
- Chain-of-Thought (CoT): Encourages step-by-step reasoning

Species Classification

Question: What is the scientific name of the butterfly shown in the image?

Correct Answer:
Heliconius timareta

Question type: Open Questions

Trait Identification

Question: Is there eye visible in the fish shown in the image?

Options:
A) Yes
B) No

Correct Answer: A) Yes

Question type: Multiple Choice Questions

Trait Referring

Question: What is the trait of the fish that correspond to the bounding box region [2545, 335, 3510, 423] in the image?

Options:
A) dorsal fin
B) caudal fin
C) adipose fin
D) pelvic fin

Correct Answer: A) dorsal fin

Question type: Multiple Choice Questions

Species Classification

Question:
What is the scientific name of the bird shown in the image?

Options:
A) Geothlypis philadelphia
B) Vireo atricapilla
C) Larus glaucescens
D) Coccothraustes vespertinus

Correct Answer: C) Larus glaucescens

Question type: Multiple Choice Questions

Trait Grounding

Question:
What is the bounding box coordinates of the dorsal fin in the fish shown in the image?

Options:
A) [453, 620, 557, 724]
B) [2545, 335, 3510, 423]
C) [2012, 1001, 2404, 1350]
D) [3444, 350, 4730, 1114]

Correct Answer: B) [2545, 335, 3510, 423]

Question type: Multiple Choice Questions

Trait Counting

Question: How many unique fins are visible in the fish shown in the image? The fins that are normally present in a fish are dorsal fin, caudal fin, pectoral fin, pelvic fin, anal fin and adipose fin.

Correct Answer:
5

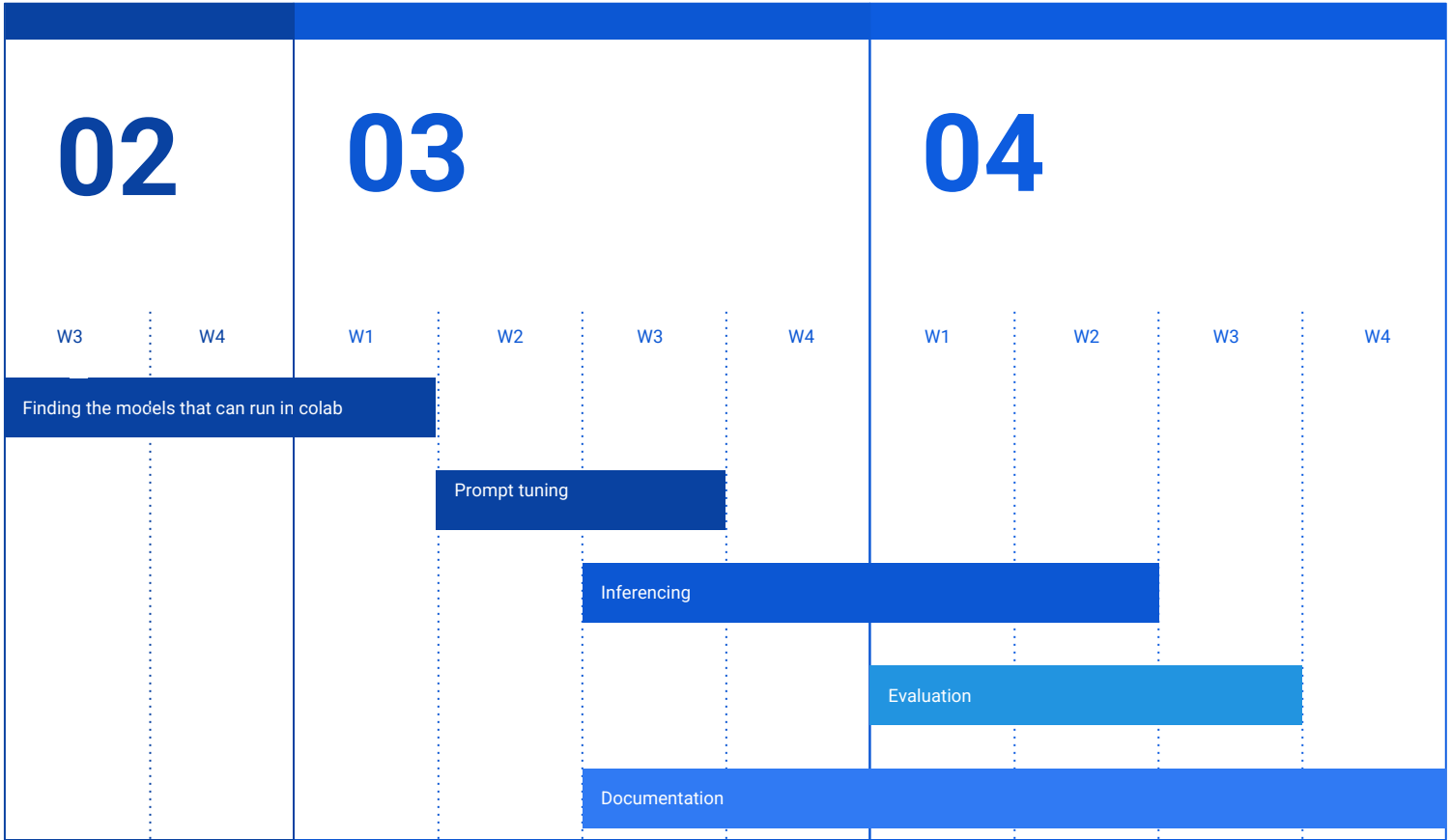
Question type: Open Questions

Approach to Replicate

- Code and data already available.
- https://huggingface.co/datasets/imageomics/VLM4Bio
- https://github.com/Imageomics/VLM4Bio/

We will evaluate on smaller and/or medium sized quantized models due to limited computational resources

Gantt chart


02
03
04
W3
W4
W1
W2
W3
W4
W1
W2
W3
W4
Finding the models that can run in colab
Prompt tuning
Inferencing
Evaluation
Documentation

Conclusions

- Checking on smaller model will be beneficial for everyone to use

Updated —--