TOWARDS BIOLOGICAL INTELLIGENCE: EVALUATING AND ALIGNING MLLMS FOR TAXONOMIC PERCEPTION AND KNOWLEDGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Biological understanding demands the integration of visual perception with structured knowledge reasoning. Although Multimodal Large Language Models (MLLMs) have made substantial progress, their capabilities in perceiving and reasoning about biological species remain largely unexplored. To address this gap, we introduce **BioBench**, a comprehensive benchmark comprising 5,919 images and 8,821 questions across 3,611 species, spanning five taxonomic ranks and five task types: morphological attribute recognition, taxonomic hierarchy classification, fine-grained species identification, biological knowledge reasoning, and humanistic knowledge reasoning. Evaluations of 23 MLLMs reveal significant difficulties in fine-grained discrimination and biologically grounded reasoning. To mitigate these limitations, we develop **BioInstruct**, a large-scale instruction-tuning dataset of 160K multimodal samples covering over 38K species, designed with few-shot species presentation to align visual traits with taxonomic hierarchies and promote generalization. Using standard Supervised Fine-Tuning on BioInstruct, a 3B-parameter model achieves substantial performance improvements, even surpassing larger frontier models. Our work establishes a foundation for biologically grounded AI, highlighting that bridging fine-grained perception with structured knowledge is critical for specialized intelligence.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Li et al., 2024a; Bai et al., 2023; Team et al., 2025) have achieved remarkable progress in integrating visual perception and linguistic reasoning across diverse domains. Advances in benchmarks are key to this progress, driving focused research in areas such as mathematics (Sun et al., 2025; Lu et al., 2023; Wang et al., 2024a), medicine (Ye et al., 2024b; Hao et al., 2025; Hu et al., 2024), scientific discovery (Yue et al., 2024; 2025), and remote sensing (Zhan et al., 2023; Li et al., 2024b). These domain-oriented datasets enable systematic evaluation of MLLMs under task-specific multimodal scenarios and accelerate development toward more general intelligence.

Despite recent advances, the biological domain, characterized by its rigorous taxonomic hierarchy and intricate morphological diversity, remains largely unexplored. Unlike conventional object recognition, understanding biological species requires models to perceive fine-grained morphological details, reason over hierarchical taxonomic relationships, and integrate domain-specific knowledge. Current state-of-the-art MLLMs, most of which are general-purpose models, still exhibit limited capability in biological perception and knowledge reasoning. Moreover, the absence of a comprehensive benchmark makes it difficult to systematically evaluate current models and identify their specific limitations, which is a necessary step before improvements can be made on these biologically grounded tasks.

To address this gap, we present **BioBench**, a taxonomy-centered multimodal benchmark designed to evaluate MLLMs on biological perception and reasoning. BioBench comprises 5,919 images and 8,821 questions across five tasks, including *morphological attribute recognition*, *taxonomic hierarchy classification*, *fine-grained species identification*, and *both biological and humanistic knowledge reasoning*. To ensure fine-grained evaluation, the benchmark encompasses a broad taxonomic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

hierarchy spanning multiple ranks from class to species.These tasks reflect a progression from visual perception to conceptual reasoning, offering a comprehensive evaluation of how MLLMs integrate fine-grained visual cues with structured taxonomic knowledge.

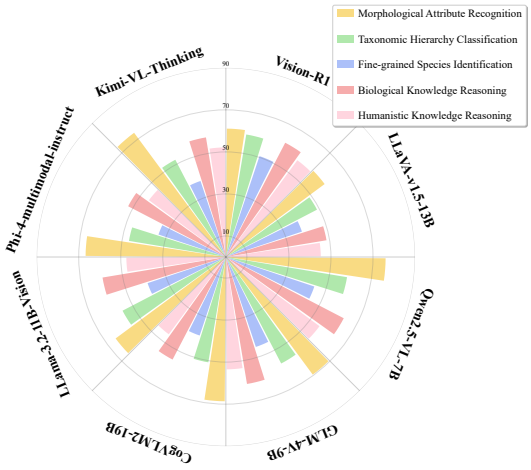


Figure 1: Performance comparison of current MLLMs on **BioBench**.

We conduct extensive experiments on BioBench to evaluate the biological perception and reasoning abilities of 23 state-of-the-art MLLMs, including general-purpose and reasoning-focused models such as Qwen2.5-VL (Bai et al., 2025a), GLM-4V (GLM et al., 2024), Llama-3.2-Vision (Grattafiori et al., 2024), Kimi-VL-Think (Team et al., 2025), and Vision-R1 (Huang et al., 2025). Figure 1 shows the performance of eight MLLMs, including six generic and two reasoning models. Our analysis reveals four key findings: (1) BioBench presents substantial challenges, with the strongest models such as Qwen-2.5VL-7B achieving only a 59.48% average score. (2) Performance declines sharply when tasks require fine-grained visual discrimination across closely related species. (3) Reasoning-oriented MLLMs show limited ability to apply biological knowledge for complex question answering. (4) Existing MLLMs underperform in open-ended biologic classification, revealing a significant generalization gap in real-world biological perception.

Building upon these insights, we construct **BioInstruct**, a large-scale multimodal instruction tuning dataset with 160K image–text pairs covering over 38K species. BioInstruct includes descriptive, dialog, and reasoning-oriented samples, providing supervision that connects fine-grained visual and anatomical characteristics to taxonomic knowledge. Fine-tuning Qwen2.5-VL-3B on BioInstruct yields **BioQwenVL**, improving biological perception and taxonomic reasoning, with an absolute average accuracy gain of approximately 4.2% on BioBench.

In summary, this work makes three contributions. First, we present BioBench, the first taxonomy-oriented multimodal benchmark for assessing MLLMs on fine-grained biological perception and reasoning. Second, we introduce BioInstruct, a large-scale instruction dataset that aligns multimodal representations with structured biological knowledge. Third, we develop BioQwenVL through instruction tuning, which improves performance on BioBench and provides insights into the gap between perception and reasoning in current models. Together, these contributions establish a foundation for biologically grounded AI, bridging fine-grained perception with structured taxonomic knowledge to advance the understanding of biodiversity and ecological complexity.

2 Related Work

Multimodal Large Language Models. MLLMs have evolved from basic visual-language alignment to large-scale unified architectures. Early models like LLaVA (Liu et al., 2023) and Flamingo (Alayrac et al., 2022) connected vision encoders with frozen language models via cross-attention, enabling efficient multimodal instruction tuning. Subsequent works, including LLaVA-Next (Li et al., 2024a), InternVL (Chen et al., 2024b), and Qwen2-VL (Wang et al., 2024b), introduced monolithic pretraining and fine-grained spatial modeling, achieving strong performance on general visual perception benchmarks. More recently, reasoning-oriented models such as Kimi-VL-Thinking (Team et al., 2025), R1-VL (Zhang et al., 2025), and DeepPerception (Ma et al., 2025) incorporate step-by-step reasoning and reinforcement learning to simulate cognitive processing for fine-grained multimodal reasoning. However, existing MLLMs still struggle in specialized domains like biology, where object-category misalignment and insufficient integration of visual traits with taxonomic knowledge limit their applicability. This underscores the need for a benchmark that systematically evaluates fine-grained species perception and biologically grounded reasoning.

Benchmarks for MLLMs. Multimodal benchmarks fall broadly into general-purpose and domain-specific categories. General-purpose benchmarks such as MME (Fu et al., 2025), SEED-Bench

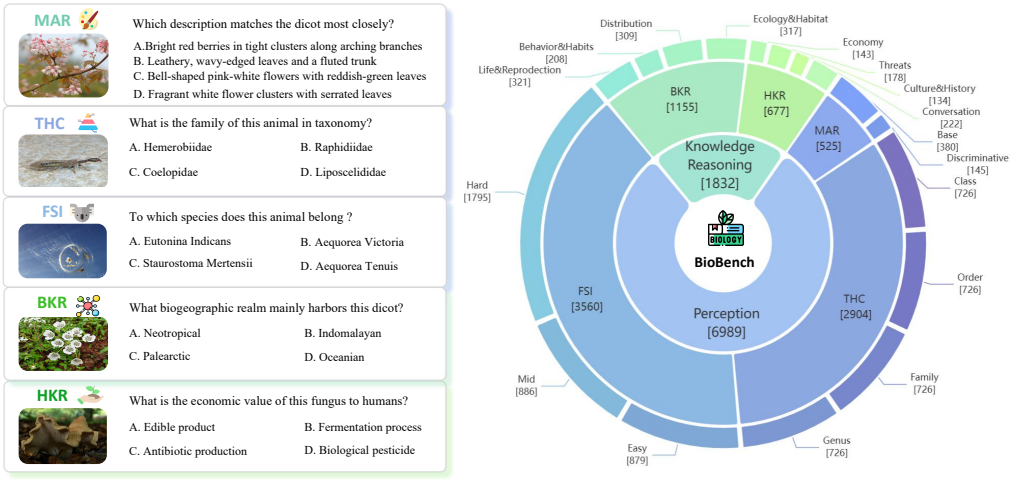


Figure 2: BioBench overview. Left: representative BioBench examples with species images and corresponding questions. Right: evaluation dimensions and data distribution, with perception tasks(MAR, THC, FSI) ordered by difficulty and reasoning tasks(BKR, HKR) by knowledge type.

(Li et al., 2023), MMBench (Liu et al., 2024c), MM-Vet (Yu et al., 2023), and MME-RealWorld (Zhang et al., 2024b) assess open-domain visual perception, multimodal understanding, and basic reasoning across diverse scenarios. Domain-specific benchmarks target professional applications requiring expert knowledge and fine-grained reasoning—MathVista (Lu et al., 2023) and MathVerse (Zhang et al., 2024a) evaluate diagram-based mathematical reasoning; OmniMedVQA (Hu et al., 2024) and MedBench (Liu et al., 2024b) assess clinical image understanding and medical decision-making; VRSBench (Li et al., 2024b) and RSGPT (Zhan et al., 2023) measure remote sensing scene interpretation. These specialized benchmarks highlight the value of evaluating models in high-stakes, expert-driven scenarios not covered by general-purpose tests. Despite these advances, large-scale, fine-grained biological species perception remains largely unexplored. Existing efforts, such as DeepPerception (Ma et al., 2025) and Geigle et al. (Geigle et al., 2024), address limited species subsets or single classification tasks, leaving a gap in comprehensive biologically grounded evaluation. Our work fills this gap by providing a benchmark that combines broad species coverage with diverse visual and reasoning tasks.

3 BioBench

3.1 Overview

Existing biology-related datasets often suffer from limited biological diversity and narrow evaluation scopes, hindering a comprehensive evaluation of current MLLMs. We propose BioBench, a novel multimodal benchmark that spans an extensive taxonomic spectrum and extends evaluation, thereby offering a more holistic evaluation of MLLMs in realistic natural scenarios.

3.1.1 Data Composition

Detailed statistics of BioBench are presented in Table 1. The benchmark encompasses a broad taxonomic hierarchy across 108 classes, 613 orders, 1,590 families, 2,729 genera and 3,611 species, ensuring extensive biological and visual diversity. To facilitate MLLMs evaluation, all tasks are formulated as multiple-choice questions (MCQs) with four answer options. Approximately 79.2% of the questions focus on perception-oriented recognition, while 20.8% involve knowledge reasoning. Collectively, BioBench comprises 5,919 images and 8,821 carefully curated questions, providing a comprehensive and biologically diverse resource for assessing MLLMs' multimodal understanding and reasoning capabilities within biological domains.

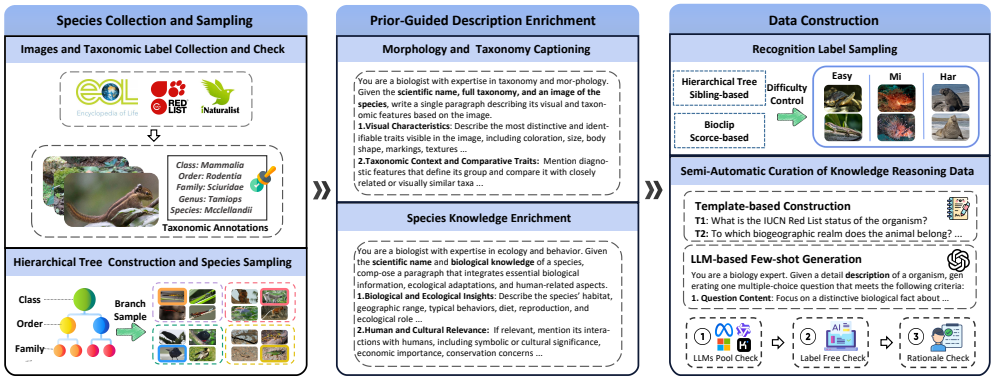


Figure 3: Overview of the BioBench data curation pipeline in three stages. Stage 1 : collection, validation, and sampling of species data. Stage 2 : LLM-based enrichment of structured biological knowledge for each species. Stage 3 : final dataset construction using outputs from prior stages, with rigorous quality control.

3.1.2 Evaluation Dimensions

BioBench evaluates MLLMs along two primary dimensions: biological perception and knowledge reasoning. These dimensions are further decomposed into five perspectives: morphological attribute recognition, taxonomic hierarchy classification, fine-grained species identification, biological knowledge reasoning, and humanistic knowledge reasoning. Figure 2 presents representative examples. Further details are as follows:

Statistic	Number
Total questions	8821
- Recognition questions	6,989 (79.2%)
- Knowledge reasoning questions	1,832 (20.8%)
Hierarchy levels	5
- Covered Class	108
- Covered Order	613
- Covered Family	1,590
- Covered Genera	2,729
- Covered Species	3,611
Average question length	10.66
Average answer length	2.81
Average choice number	4.00

Table 1: Key statistics of our proposed **BioBench**.

Morphological Attribute Recognition (MAR): This task evaluates the model's ability to select the textual description that best matches an organism's visual attributes. The core challenge lies in distinguishing candidates that differ by subtle yet biologically meaningful traits, such as coloration, shape, or distinctive markings. It includes two settings: base, with coarse attribute differences, and discriminative, emphasizing fine-grained morphological distinctions.

Taxonomic Hierarchy Classification (THC): This task assesses the model's ability to comprehend and apply the biological taxonomic hierarchy across multiple classification levels. It requires determining an organism's taxonomic ranks—from class to genus—by linking observable morphological traits to hierarchical categories. The challenge lies in mapping visual traits to a multi-level hierarchy, where each taxonomic rank imposes distinct constraints that guide classification at finer granularities.

Fine-grained Species Identification (FSI): This task requires identifying exact species from visually similar candidates, with distractors sampled at varying difficulty levels based on similarity. Ground-truth species span diverse taxonomic branches to ensure broad coverage. The task challenges models to detect subtle morphological differences among closely related species while generalizing across lineages, assessing both fine-grained discrimination and cross-branch generalization.

Biological Knowledge Reasoning (BKR): This task requires models to first identify an organism's taxonomic category and then perform multi-dimensional reasoning over core biological concepts, including life cycles, reproductive strategies, behavior, geographic distribution, habitat, and ecological niches. It demands integration of species-specific morphological traits with ecological knowledge.

The challenge lies in accurately integrating visual identification with coherent biological reasoning about ecological relationships.

Humanistic Knowledge Reasoning (HKR): This task requires models to identify species and reason about their interactions with human contexts, including cultural significance, historical roles, economic value, human-induced threats, and societal impacts. It demands integration of species-specific information with broader socio-cultural and environmental knowledge. The challenge lies in linking biological entities to human-centered contexts and reasoning about their societal and ethical implications

3.2 Data Collection and Annotation

3.2.1 Species Collection and Sampling

To ensure broad taxonomic coverage, we collect images and their taxonomic annotations from multiple sources, including the Encyclopedia of Life (Parr et al., 2014), the IUCN Red List (Rodrigues et al., 2006), and the iNaturalist2021 (Van Horn et al., 2021) dataset. Given the inherent noise from inconsistent taxonomic annotations, we remove entries with incomplete or ambiguous taxonomy and perform cross-source consistency checks combined with LLM-assisted validation. Through this pre-filtering process, we obtain a high-quality species pool covering 81,308 species.

To balance evaluation efficiency with taxonomic coverage, we construct a taxonomic tree with uniform branch sampling, further incorporating species rarity, geographic distribution, and IUCN status for diversity. The resulting 3,611 representative species undergo manual image screening to retain clear single-species samples and exclude visually confounding cases such as parasitic or entangled specimens.

3.2.2 Prior-Guided Description Enrichment

To enrich the benchmark with comprehensive biological information, we adopt a prior-guided approach to generate structured textual descriptions for each species. We first employ Qwen2.5-VL-72B (Bai et al., 2025a) to produce distinctive visual descriptions conditioned on hierarchical taxonomic labels and verified images, capturing observable traits such as coloration, body structure, and notable markings while emphasizing comparative features that distinguish the species from closely related or visually similar taxa. Concurrently, we gather biological and humanistic context from multiple authoritative sources and use Qwen2.5-72B (Qwen et al., 2025) to synthesize and reorganize the knowledge. All generated descriptions undergo a post-editing pipeline involving automated filtering and LLM-assisted revision to ensure factual accuracy and linguistic consistency, with benchmark-included species additionally manually verified against Wikipedia and IUCN Red List entries.

3.2.3 Data Construction

Recognition Label Sampling. To calibrate recognition difficulty and ensure label diversity, we adopt a taxonomy-grounded sampling strategy. For each target label, distractors are sampled from sibling taxa under the same parent node based on taxonomic proximity, with closely related taxa for hard cases, broader ranks for intermediate ones, and distant groups for easy ones. Candidate distractors are further ranked using BioCLIP (Stevens et al., 2024) similarity scores for fine-grained difficulty control. Label frequencies are tracked to restrict each label to at most three appearances, ensuring balanced coverage.

Semi-Automatic Curation of Knowledge Reasoning Data. To enrich question diversity and biological coverage, we design an LLM-assisted semi-automatic generation pipeline with two components. (1) *Template-based Construction* uses structured resources (IUCN, 2025) to design domain-specific templates, enabling scalable generation of high-quality MCQs with verified answers. (2) *LLM-based Few-shot Generation* employs GPT-4o (Achiam et al., 2023) to produce species-specific MCQs guided by taxonomic labels and species descriptions. Carefully designed prompts and few-shot exemplars ensure broad coverage across hierarchical and species levels, with outputs post-processed for structural consistency.

Table 2: Overview of the BioInstruct dataset, including covered species and subset sizes.

Species	Subset	Description	Size
38,892	Profile	Single-turn instruction–response pairs for species knowledge.	44K
	Conversation	Multi-turn dialogues from species descriptions.	70K
	Reasoning	Single QA pairs for species recognition and reasoning.	47K

To ensure reliability, we apply a three-stage verification process. First, an LLM pool comprising LLaMA3.1 (Grattafiori et al., 2024), Qwen3 (Yang et al., 2025a), Phi4 (Abdin et al., 2024), and Moonlight (Liu et al., 2025) independently answers each question given the ground-truth label; only questions answered correctly and consistently by all models are retained. Second, species labels are masked to filter out questions solvable without visual input, and the remaining questions are evaluated by multiple MLLMs to remove those that are overly difficult or ambiguous. Third, manual verification cross-compares model-generated explanations to ensure interpretive consistency. This pipeline yields a diverse and reliable question set for biological knowledge perception and reasoning.

4 BioQwenVL

To enhance MLLMs' perceptual and reasoning capabilities in biological domains, we construct **BioInstruct**, a large-scale multimodal instruction-tuning dataset spanning diverse taxa and biological contexts.

BioInstruct. As summarized in Table 2, BioInstruct comprises over 160K multimodal instruction samples across 38K species and diverse taxonomic levels, with each species presented in a few-shot format to promote cross-taxa generalization. The dataset includes three subsets: (1) *Profile*, integrating species-level textual descriptions covering morphology, taxonomy, ecology, and ethology with visual traits to ground fine-grained biological knowledge; (2) *Conversation*, featuring multi-turn dialogues that simulate expert interactions to reinforce hierarchical taxonomic alignment and biological knowledge application; and (3) *Reasoning*, providing knowledge-grounded data to guide models in reasoning-based problem solving through integration of visual cues and biological knowledge. Profile descriptions are derived from our enriched taxonomic corpus, while Conversation and Reasoning samples are generated via the LLM-assisted semi-automatic curation pipeline, enabling BioInstruct to capture both static descriptive knowledge and dynamic reasoning behaviors essential for biodiversity understanding.

Supervised Fine-Tuning. We perform full-parameter fine-tuning of Qwen2.5VL-3B (Bai et al., 2025b) on BioInstruct, updating all model parameters including the vision encoder, projection layers, and language model via cross-entropy loss to align outputs with instruction data. The resulting model, **BioQwenVL**, integrates fine-grained visual perception, hierarchical taxonomic understanding, and biological reasoning, delivering reliable and biologically grounded performance on biodiversity tasks.

5 Experiments

In this section, we first describe the experimental setup. We then present the main results, including the evaluation of representative MLLMs on BioBench and the improvements introduced by our data construction and training pipeline, comparing outcomes across different stages. Finally, we provide comprehensive analyses and ablation studies, highlighting key findings across multiple dimensions.

5.1 Experimental Setups

Evaluation Settings. We evaluate 23 MLLMs on BioBench across all tasks. Table 3 reports the main results. We select 17 open-source generic MLLMs, including LLaVA-v1.5 (Liu et al., 2023), LLaVA-Next (Liu et al., 2024a), InternVL2.5 (Chen et al., 2024a), InternVL3 (Zhu et al., 2025), Qwen2-VL (Wang et al., 2024b), Qwen2.5-VL (Bai et al., 2025a), Phi-4-multimodal-instruct (Abdin et al., 2024), GLM-4V (GLM et al., 2024), Kimi-VL-Instruction (Team et al., 2025), CogVLM2

Table 3: Main results on the benchmark. MAR = Morphological Attribute Recognition, THC = Taxonomic Hierarchy Classification, FSI = Fine-grained Species Identification, BKR = Biological Knowledge Reasoning, HKR = Humanistic Knowledge Reasoning. The terms *disc.*, *distr.*, *econ.* and *cons.* are short for discriminative, distribution, economic and conservation. The best performance is highlighted in red, and the second-best is shown in blue.

Model	Size	MAR		THC				FSI			BKR				HKR				Avg.
		base	*disc.*	*class*	*order*	*family*	*genus*	*easy*	*mid*	*hard*	*life*	*behavior*	*distr.*	*ecology*	*econ.*	*threats*	*culture*	*cons.*	
Open-Source Generic MLLMs																			
LLaVA-v1.5	7B	55.79	37.67	65.98	46.28	33.20	36.50	47.15	33.86	27.44	46.42	42.31	42.39	47.00	46.85	32.58	51.49	30.63	42.72
LLaVA-v1.5	13B	70.26	46.58	76.03	48.62	34.44	36.36	46.47	37.13	30.83	48.91	46.63	43.69	54.89	52.45	41.57	55.97	30.63	48.14
LLaVA-Next	7B	62.11	47.95	70.52	44.90	32.37	37.33	44.53	33.86	26.17	42.06	46.63	41.42	46.69	46.15	38.76	49.25	30.18	44.13
LLaVA-Next	13B	66.58	45.89	68.18	46.42	31.82	32.09	45.67	36.68	27.33	50.47	50.00	41.10	51.10	48.25	38.20	52.99	28.83	45.57
InternVL2.5	8B	50.00	48.63	48.35	35.95	24.38	31.54	29.38	27.65	21.78	41.74	34.13	37.54	45.11	37.76	30.34	55.97	26.58	37.14
InternVL3	14B	51.32	54.79	54.82	40.91	27.96	33.88	32.46	33.75	25.33	42.37	34.13	36.57	43.85	41.26	33.71	70.90	29.73	40.39
InternVL3	38B	57.63	57.53	48.90	37.33	26.31	35.12	33.03	32.39	26.61	42.99	30.77	36.57	45.74	42.66	33.71	66.42	28.83	40.89
Qwen2-VL	7B	78.42	48.63	78.79	61.57	40.36	38.71	61.89	44.02	31.80	68.22	57.21	54.05	67.51	55.94	47.19	72.39	39.64	56.21
Qwen2.5-VL	7B	81.58	61.38	84.99	64.05	43.11	41.60	62.68	**45.49**	34.85	68.54	59.62	52.10	70.98	**62.24**	51.69	**76.12**	41.44	59.48
GLM-4V	9B	81.58	57.53	87.47	61.16	41.46	38.29	59.91	41.87	33.50	63.86	61.06	53.40	66.56	61.54	46.63	66.42	39.19	57.33
Kimi-VL-Instruction	16B	80.53	**69.18**	79.48	59.09	**43.66**	42.01	56.61	41.53	32.74	59.19	62.02	51.78	66.25	51.75	46.63	66.42	36.94	56.47
Phi-4-multimodal-instruct	5.6B	73.42	48.63	71.35	49.31	31.13	33.61	47.21	33.86	27.40	62.50	53.14	**76.47**	61.95	57.65	45.45	68.66	38.30	51.91
CogVLM2	19B	76.58	47.95	77.41	53.04	36.91	35.81	51.82	42.78	31.61	54.52	61.06	49.84	56.47	50.35	44.38	61.19	31.98	51.72
Emu3-Chat	8B	61.84	50.00	50.41	27.13	24.66	24.93	34.58	27.43	22.96	37.69	38.94	34.63	36.28	37.76	31.46	41.04	29.28	37.59
mPLUG-Owl3	7B	73.68	52.74	72.18	51.93	33.61	34.57	49.15	35.21	27.29	60.12	55.77	45.95	61.20	46.85	48.88	64.18	37.84	50.86
Llama-3.2-Vision	11B	70.00	52.74	77.55	60.74	40.22	42.15	52.56	39.28	32.30	60.75	61.54	56.31	59.94	46.85	44.38	64.93	39.19	53.23
Open-Source Reasoning MLLMs																			
Kimi-VL-Thinking	16B	77.89	60.96	79.75	54.68	37.33	36.64	54.04	38.71	29.96	66.67	60.10	43.04	60.88	60.84	51.69	66.42	37.84	54.59
R1-VL	7B	71.84	48.63	78.79	55.79	38.15	31.40	56.54	41.42	28.74	62.31	55.77	48.54	63.41	58.04	42.70	61.19	36.94	52.13
Vision-R1	7B	67.63	43.15	85.67	64.60	41.32	42.56	60.07	44.13	31.63	66.04	62.02	49.51	65.62	58.04	47.19	63.43	36.94	54.25
R1-Onevision	7B	73.95	58.22	82.78	61.57	36.50	37.33	57.45	42.89	32.91	62.93	57.21	47.25	57.73	55.24	42.70	52.99	38.74	53.59
DeepPerception-FGVR	7B	71.05	56.16	71.07	59.64	39.39	38.29	55.52	41.65	30.41	54.83	47.60	44.66	54.26	44.76	43.82	64.93	35.14	50.88
Qwen2.5-VL(baseline)	3B	80.53	56.55	78.51	61.98	42.15	39.67	58.48	41.87	33.24	62.31	**66.83**	55.02	65.93	50.35	51.69	71.64	40.99	57.02
Bioqwen-VL(ours)	3B	**82.37**	57.24	**90.50**	**67.91**	42.29	**43.53**	**63.71**	44.47	**36.80**	**71.96**	66.35	60.84	**75.39**	60.84	**54.49**	71.64	**42.34**	**61.20**

(Wang et al., 2024c), Emu3-Chat (Wang et al., 2024d), mPLUG-Owl3 (Ye et al., 2024a), and Llama-3.2-Vision (Grattafiori et al., 2024). To ensure standardized outputs, we append "Answer with the option's letter from the given choices directly" to all prompts. We also evaluate five open-source multimodal reasoning models: Kimi-VL-Thinking (Team et al., 2025), R1-VL (Zhang et al., 2025), R1-Onevision (Yang et al., 2025b), Vision-R1 (Huang et al., 2025), and DeepPerception-FGVR (Ma et al., 2025).

For these reasoning-centric models, we extract the content enclosed within `<answer></answer>` tags and then identify the selected option. In case where the model fail to follow the instruction format, we leverage Qwen2.5-3B with a customized prompt to determine the correctness. Furthermore, to more realistically evaluate model performance in open-ended scenarios, we sample a subset of the benchmark to assess several models' recognition capabilities. For this open-ended evaluation, we use a customized prompt with Qwen2.5-3B to compare model predictions against ground-truth labels, enabling a binary assessment of recognition accuracy beyond closed-set conditions. All models are evaluated under the same inference settings, with temperature set to 0.2, number of beams fixed at 1, and a maximum sequence length of 2048 tokens.

Training Details. We use the ms-swift framework to perform full model fine-tuning on 6 NVIDIA A6000 GPUs. We use the AdamW optimizer with a learning rate of 5×10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1e - 8$. The per-device batch size is 4 with gradient accumulation steps of 4, resulting in a total batch size of 96. The maximum sequence length is set to 4096 tokens. Training is conducted for 1 epoch using BFloat16 precision, and DeepSpeed ZeRO Stage 3 is employed for memory optimization.

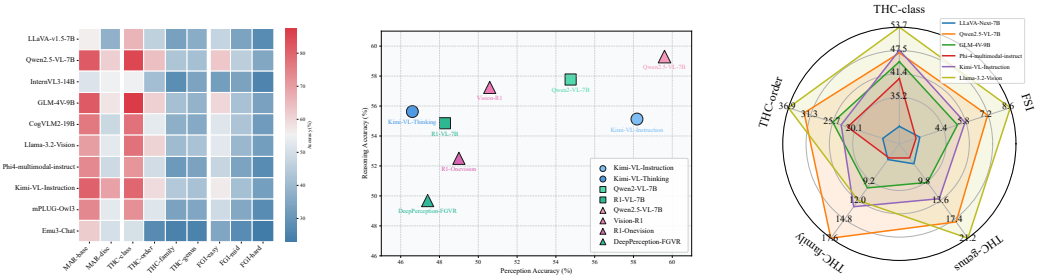
5.2 Main Results

Table 3 presents the main results, from which we derive the following key findings:

Finding 1: BioBench poses a significant challenge to current state-of-the-art MLLMs. Even the most advanced open-source model, Qwen2.5-VL-7B, yields an average accuracy of merely 59.48%. This performance bottleneck reveals fundamental limitations in current multimodal models, including fine-grained visual perception, hierarchical taxonomic understanding, and the integration of species-specific knowledge. These findings underscore the inherent constraints of MLLMs in biological visual understanding and reasoning, emphasizing the urgent need for developing expert-level models tailored for biodiversity understanding.

Table 4: Ablation study of BioInstruct under different supervised fine-tuning (SFT) data. The full integration of all three subsets yields superior performance, confirming the effectiveness of the BioInstruct design.

Model	SFT			Perception				Reasoning			Avg.
	Profile	Conversation	Reasoning	MAR	THC	FSI	Overall	BKR	HKR	Overall	
Qwen2.5-VL-3B	✗	✗	✗	73.90	55.27	41.61	57.03	62.16	51.85	57.01	57.02
BioqwenVL	✓	✗	✗	76.25	54.88	41.47	57.53	64.57	53.18	58.87	58.20
BioqwenVL	✗	✓	✗	73.90	59.02	43.94	58.96	65.08	54.04	59.56	59.19
BioqwenVL	✗	✗	✓	75.24	55.92	41.08	57.41	63.54	52.84	58.19	57.80
BioqwenVL	✓	✓	✓	75.24	58.71	44.61	59.52	65.54	53.62	59.58	59.54



(a) Comparison across difficulty levels.

(b) Comparison on perception *vs.* reasoning.

(c) Comparison on open-ended recognition.

Figure 4: Performance comparison across multiple MLLMs from three perspectives.

Finding 2: MLLMs exhibit marked performance decay in fine-grained biological perception. As shown in Figure 4 (a), accuracy declines as tasks demand finer visual discrimination. MAR drops from *base* to *discriminative*; THC accuracy falls sharply from class to genus; and FSI plunges from 62.68% (*easy*) to 34.85% (*hard*). The decline is most pronounced in FSI, where subtle, morphology-dependent differences challenge models. Similar trends across GLM-4V and Kimi-VL-Instruction confirm that current MLLMs fail to capture fine-grained morphological cues essential for biodiversity understanding.

Finding 3: Enhanced reasoning fails to yield consistent gains in biological perception. As illustrated in Figure 4 (b), reasoning-oriented models show only marginal improvements on reasoning tasks while generally underperforming on perception tasks compared to their instruction-tuned baselines. Kimi-VL-Thinking achieves limited gains on a few reasoning tasks but declines across most perception tasks relative to Kimi-VL-Instruction. Similarly, R1-based variants—R1-VL *vs.* Qwen2-VL-7B and Vision-R1 *vs.* Qwen2.5-VL-7B—consistently underperform on perception tasks. These results suggest that current reasoning mechanisms struggle to integrate effectively with fine-grained visual perception in biological domains.

Finding 4: MLLMs show limited open-ended generalization in biology. As shown in Figure 4 (c), we evaluate model performance on two open-ended perception benchmarks: THC, spanning taxonomic levels from class to genus, and FSI, focusing on species-level recognition. The results demonstrate a substantial performance degradation compared to their closed-ended counterparts. This performance gap indicates that while models can select correctly from predefined candidates, they struggle to accurately map fine-grained visual features to hierarchical taxonomic labels. This misalignment between visual perception and structured taxonomy represents a key bottleneck for reliable biological understanding in open-ended settings.

5.3 Ablation Study of BioInstruct

To assess the contribution of each BioInstruct subset, we fine-tune the base model on *Profile*, *Conversation*, and *Reasoning* individually, each with 30K samples for fair comparison. As shown in Table 4, all three subsets consistently improve over the base model, indicating complementary contributions. Among them, *Conversation* yields the most significant overall gain, as its multi-turn

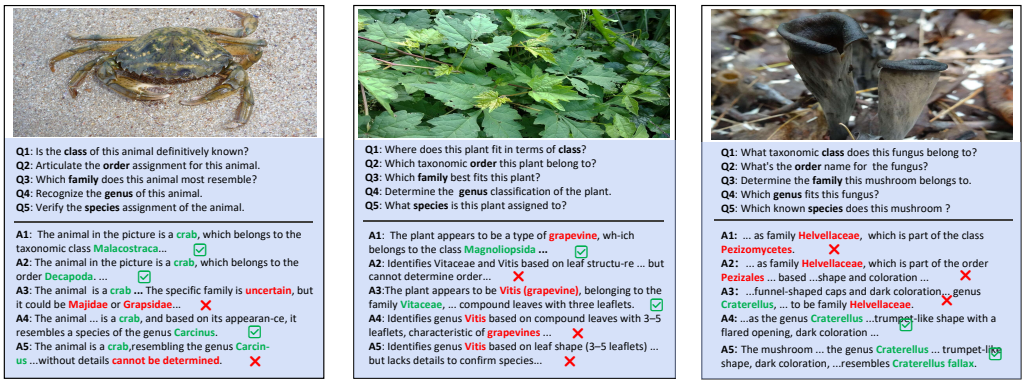


Figure 5: Case study of hierarchical misalignment in MLLM visual taxonomic understanding: Qwen2.5-VL-7B correctly predicts low-level categories but produces inconsistent higher-level predictions, highlighting failures incross-level consistency from top to down and mapping fine-grained visual evidence to accurate taxonomic labels. Correct predictions are highlighted in green, and incorrect ones in red.

dialogues emphasize aligning species with taxonomic labels, enabling the model to better capture hierarchical relationships and contextualize traits within broader taxonomic structures. Combining all three subsets produces further synergistic improvements across both perception and reasoning tasks, validating the integrated design.

Despite these gains, performance remains limited by domain complexity and coarse perception, with room for improvement via richer instruction data and multimodal alignment. More critically, achieving stronger hierarchical alignment is essential to move beyond superficial visual similarities, enable coherent cross-taxonomic reasoning, and support biologically meaningful predictions.

5.4 Case Study

To qualitatively examine how MLLMs align visual perception with taxonomic hierarchies, we query Qwen2.5-VL-7B with a single image across multiple taxonomic levels, from class down to species, in an open-ended setting. We observe frequent cases where the model correctly identifies lower-level categories such as species or genus but fails at higher levels such as order or class, suggesting that current MLLMs lack hierarchically consistent visual representations. Instead of performing systematic top-down reasoning, the model tends to rely on salient low-level visual attributes, including color, shape, and texture, to directly predict a fine-grained taxon, and subsequently uses language priors or reverse inference to assign broader categories. This behavior points to a fundamental limitation in the training paradigm: models are optimized to map visual inputs directly to the most specific taxonomic labels, inadvertently encouraging shortcut learning that bypasses the intermediate hierarchical structure. As a result, they acquire features that are locally predictive for species-level recognition but fail to develop a principled understanding of how taxa are embedded within broader biological contexts, leaving cross-level reasoning inherently unreliable.

6 Conclusion

This work presents BioBench, the first taxonomy-oriented multimodal benchmark for evaluating MLLMs on fine-grained biological perception and reasoning, spanning hierarchical taxonomic levels and diverse task types. Our evaluation of 23 models reveals key challenges, such as difficulty in discriminating closely related species, and suboptimal optimization of current reasoning models. To further advance biological understanding in MLLMs, we introduce BioInstruct and develop BioQwen-VL via supervised fine-tuning, improving the alignment between visual features and taxonomic knowledge. Overall, this study lays the groundwork for biologically grounded multimodal intelligence by integrating fine-grained recognition with hierarchical reasoning.

7 AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript and for retrieval and discovery of relevant literature.

REPRODUCIBILITY STATEMENT

Our work primarily contributes two datasets: BioBench and BioInstruct. Appendix A provides comprehensive supporting materials. Specifically, Appendix A.1 illustrates example cases of species data collection and preprocessing. Appendix A.2 details the BioBench curation process, including the prompts used and relevant statistics. Appendix A.3 presents visualizations of the BioInstruct dataset. Appendix A.4 reports supplementary evaluation results, while Appendix A.5 provides additional case studies. Finally, Appendix A.6 discusses the limitations of our work and outlines directions for future research. The datasets and evaluation code will be released with the final version.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. *arXiv preprint arXiv:2406.14496*, 2024.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lizhuo Lin, Jinrong Yang, Qi Yong H Ai, Lun M Wong, Hao Tang, and Kuo Feng Hung. Towards better dental ai: A multimodal benchmark and instruction dataset for panoramic x-ray analysis. *arXiv preprint arXiv:2509.09254*, 2025.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

IUCN. Iucn red list of threatened species, version 2025-2. https://www.iucnredlist.org, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. In *Proceedings of Advances in Neural Information Processing Systems*, volume 37, pp. 3229–3242, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Mianxin Liu, Weiguo Hu, Jinru Ding, Jie Xu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, et al. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Big Data Mining and Analytics*, 7(4):1116–1128, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *Proceedings of European Conference on Computer Vision*, pp. 216–233, 2024c.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*, 2025.

Cynthia S Parr, Mr Nathan Wilson, Mr Patrick Leary, Katja S Schulz, Ms Kristen Lans, Ms Lisa Walley, Jennifer A Hammock, Mr Anthony Goddard, Mr Jeremy Rice, Mr Marie Studer, et al. The encyclopedia of life v2: providing global access to knowledge about life on earth. *Biodiversity data journal*, 2:e1079, 2014. doi: 10.3897/BDJ.2.e1079.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Ana SL Rodrigues, John D Pilgrim, John F Lamoreux, Michael Hoffmann, and Thomas M Brooks. The value of the iucn red list for conservation. *Trends in ecology & evolution*, 21(2):71–76, 2006.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19412–19424, 2024.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. Mathglance: Multimodal large language models do not know where to look in mathematical diagrams. *arXiv preprint arXiv:2503.20745*, 2025.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12884–12893, 2021.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *Proceedings of Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Proceedings of Advances in Neural Information Processing Systems*, 37:121475–121499, 2024c.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024d.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025b.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024a.

Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. In *Proceedings of Advances in Neural Information Processing Systems*, volume 37, pp. 94327–94427, 2024b.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 15134–15186, 2025.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Proceedings of European Conference on Computer Vision*, pp. 169–186, 2024a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix

A.1 Data Preparation



Class: Aves **Order**: Passeriformes
Family: Panuridae **Genus**:Panurus
Common name: Bearded Reedling
Scientific name: Panurus Biarmicus

IUCN Assessment: {"population_trend": "Increasing", "red_list_category":"Least Concern","movement_pattern": "Not a Migrant", "actions": ["Systematic monitoring scheme","Included in international legislation"], documentation: { "population": "In Europe, the total population size is stimated at 343,000-783,000 ...", "habitats": "This speci-es is found in extensive reedbeds and is associated with dense non-woody vegetation...", "threats": "Populations have al-ways been subject to regular annual population fluctuations...", "measures": "This species will use nestboxes, particularly when nest sites are in short supply ...", "use_trade": "This species has been reported as used in trade ...", "rationale": This species has an extremely large range and the population size is extremely large ..." }, "biogeographical_realms":" ["Palearctic"] ... }

Wikipedia Knowledge: The bearded reedling (Panurus biarmicus) is a small, long-tailed passerine bird found in reed beds near water in the temperate zone of Eurasia. It is frequently known as the bearded tit or bearded parrotbill, as it historically was believed to be closely related to tits or parrotbills. Today it is known to lack close relatives and it is the only species in the family Panuridae. Bearded reedlings are strongly sexually dimorphic and form life-long pairs. They are highly productive and can breed several times in a season. They mainly feed on small invertebrates in summer and plant seeds in winter. **Taxonomy and systematics**: He placed it with the tits in the genus Parus and coined the ... **Distribution and habitat**: The bearded reedling is native to temperate Europe and Asia ... **Appearance and voice:** This is a small bird ...

Figure 6: Example of collected data sources. The left panel shows sampled images with their corresponding taxonomic annotations, while the right panel presents species knowledge collected from IUCN and Wikipedia.

We begin by obtaining data from BioCLIP (Stevens et al., 2024), which is mainly sourced from EOL (Parr et al., 2014) and iNaturalist2021 (Van Horn et al., 2021). This repository contains over 450K unique classes and 10.4M biological images, along with their corresponding taxonomic labels. However, this raw data exhibits substantial noise, including low-quality images and incorrect taxonomic annotations.

To improve label reliability, we additionally collect 169,542 species and their taxonomic information from the IUCN Red List (Rodrigues et al., 2006). We then take the intersection of the two sources and perform cross-source taxonomic consistency checks. To further ensure annotation accuracy, we prompt Qwen2.5-72B (Qwen et al., 2025) with the full taxonomic labels of each species to verify the correctness of their hierarchical taxonomy. After this multi-stage verification, we retain 81,308 species with validated taxonomic annotations, comprising animals, plants, and fungi.

For image preparation, we sample biological images from BioCLIP and use Qwen2.5-VL-7B (Bai et al., 2025b) to verify species presence with the prompt "Is the species species present in the image, and can it be clearly observed without significant occlusion?". We also collect additional candidate images for each species from Wikipedia and Google Images based on species names, using them as supplementary sources when BioCLIP images are insufficient or required validation. All images used in **BioBench** are then manually reviewed to ensure that each depicts a single, clearly visible individual of the target species. This multi-stage filtering process ensures the overall visual quality and reliability of the final dataset.

In addition to images and taxonomic labels, we further gather species-related textual knowledge from the IUCN API (v4) (IUCN, 2025) and Wikipedia whenever available, providing richer ecological, morphological, and conservation information that supports downstream task construction and enhances the biological grounding of our benchmark. Figure 6 illustrates representative examples of collected species.

A.2 BioBench Curation Details

A.2.1 Ground-Truth Species Sampling

Using the collected taxonomic annotations, we construct a five-level hierarchical taxonomy from class to species. To maximize taxonomic coverage under a limited sampling budget, we perform family-level stratified sampling by uniformly selecting one species from each family branch. This strategy reduces bias toward species-rich families and ensures broad coverage across diverse biological lineages.

To further improve ecological and conservation diversity, we supplement the initial species pool using species-level knowledge, including geographic distribution and IUCN conservation status.

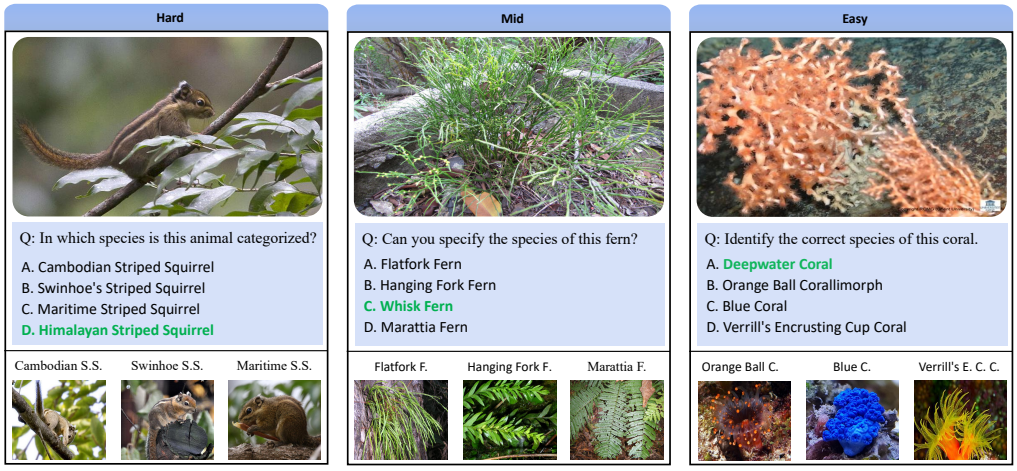


Figure 7: Representative examples from the Fine-grained Species Identification (FSI) task, sampled across varying difficulty levels from easy to hard.

Specifically, additional species are uniformly selected across different countries/regions and conservation categories to increase representation of diverse habitats and threat levels. The number of supplemented species is determined based on coverage requirements rather than natural species frequency.

Following this two-stage sampling procedure, BioBench contains 3,611 ground-truth species spanning diverse taxonomic, ecological, and conservation contexts.

A.2.2 Species Description Generation

To support knowledge reasoning tasks and the construction of species-centered dialogue and reasoning data, we consolidate and enrich species descriptions using the collected annotations. Since individual sources typically capture only partial aspects of species information, we aim to construct a unified textual knowledge representation that integrates visual characteristics with biological knowledge, covering morphology, taxonomy, ecology, distribution, natural history, and conservation-related information.

As illustrated in Figure 8, we generate species descriptions through two complementary prompts. The first prompt, executed with Qwen2.5-VL-72B Bai et al. (2025b), focuses on image-grounded visual and taxonomic descriptions. Given species images and associated annotations, the model identifies observable morphological characteristics and taxonomic cues, providing a visual perspective of the species. The second prompt, powered by Qwen2.5-72B Qwen et al. (2025), incorporates supplementary species knowledge to generate descriptions from a biological and human-related perspective, including geographic distribution, habitat, ecological roles, life-history traits, and conservation information.

The outputs from the two prompts are subsequently integrated into a single coherent species description, combining image-grounded observations with broader biological knowledge. This process enables each species to be represented from both perceptual and semantic perspectives, producing comprehensive descriptions that serve as the textual knowledge foundation for subsequent knowledge reasoning tasks and the construction of BioBench-related instruction data.

A.2.3 Recognition Label Sampling

For the Taxonomic Hierarchy Classification and Fine-grained Species Identification tasks, we follow prior work Geigle et al. (2024) in sampling confusing labels, but introduce more fine-grained, difficulty-aware control.

You are an expert biologist and field naturalist with deep knowledge of species taxonomy and morphology. You will be given the **scientific name** of a species, its complete **taxonomic hierarchy** (class, order, family, genus, species), and an **image** of the organism. Your task is to produce a single, cohesive, and richly detailed paragraph that describes the species' visual appearance and taxonomic context based solely on what can be observed in the image.

Begin naturally with an **introductory phrase** such as "The image shows a ..." or "Depicted here is a ..." followed by the species name.

Your paragraph must seamlessly integrate two complementary dimensions:

1. **Visual Diagnostic Features** — Describe the most distinctive and identifiable traits visible in the image, including coloration, body proportions, textures, patterns, structural features, and other morphological cues that support species recognition. Emphasize the traits most characteristic or striking in this particular image.

2. **Taxonomic Context and Comparative Traits** — Situate the species within its broader taxonomic ranks (class, order, family, genus). Briefly mention key morphological hallmarks of its taxonomic group and provide meaningful comparisons to closely related or visually similar taxa, highlighting features that distinguish the species within its lineage.

Keep the description strictly visual and morphological. **Do not include** behavior, habitat, ecology, distribution, or conservation information unless it is directly inferable from the image. The paragraph should read like a polished excerpt from a professional zoological or natural history field guide—precise, fluent, authoritative, and strongly **grounded in the visual evidence**.

Example:
Image: <a photo of a species_name>
Species name: {input_species_name}
Taxonomy: {input_taxonomy}
Output: {output_example}

Now describe the organism in the image below:

You are a highly knowledgeable biologist and field naturalist with expertise in species taxonomy, morphology, ecology, and biogeography. You will be provided with the **scientific name** of a species along with additional **supplementary information** about this species. Your task is to compose a single, cohesive, and richly informative paragraph describing the species' **biological, ecological, and human-related significance**.

Your description must focus exclusively on the following two dimensions:

1. **Biological and Ecological Insights** — Summarize the species' natural history, including its habitat preferences, geographic distribution, feeding strategies, reproductive biology, life cycle, behavioral ecology, and functional role within its ecosystem. Explain how these traits represent adaptations to its environment or contribute to broader ecological dynamics.

2. **Human and Cultural Relevance** — If applicable, discuss the species' interactions with humans, such as cultural symbolism, traditional uses, domestication or management history, economic importance, conservation concerns, major anthropogenic threats, and ongoing protection or restoration measures.

The paragraph should read like an authoritative entry from a modern field guide or zoological monograph—scientifically accurate, contextually rich, and written in a fluid narrative style. **Integrate** the provided supplementary information naturally into the description, synthesizing it into a coherent account of the species' biology and relevance. Avoid listing facts mechanically.

Example:
Species name: {input_species_name}
Supplementary information: {supplementary_info}
Output: {output_example}

Now generate the final description using the following inputs:

Figure 8: Overview of our two complementary description-generation prompts. **Top:** The visual–taxonomic prompt guides MLLMs to produce image-grounded descriptions focusing on morphology and taxonomy. **Bottom:** The biological–humanistic prompt constructs species-level natural history narratives using supplemental information beyond the visual input.

Animals Templates

- What reproductive strategy does the animal employ?
- What is the average lifespan of this species?
- What is the primary activity pattern of this animal?
- What is the typical social structure of this animal?
- What is the dominant diet type of this animal?
- What ecological role does this animal mainly play?
- What is the main ecological function of the animal?
- How does this animal respond to threats?
- ...

Plants Templates

- What type of life cycle or growth pattern does this plant have?
- What is the pollination method of this plant?
- When does this plant typically produce flowers?
- How does this plant negatively affect human health or safety?
- ...

Fungi Templates

- How does this fungus produce its spores?
- Which environment does the fungus species typically grow?
- What reproductive method does this fungus mainly use?
- ...

General Templates

- What is the IUCN Red List status of the organism?
- What is the primary biogeographical realm where the species is found?
- What country serves as the main range for this species?
- Which ecosystem supports the life of this organism?
- Which actions have been taken to improve the survival chances of the species?
- Which of the following factors are considered risks for the species?
- ...

Figure 9: Representative template designs used in the semi-automatic construction of knowledge-reasoning questions. These templates are derived from curated species knowledge and serve as the basis for generating diverse biological queries.

You are a biology expert specializing in the creation of high-quality multiple-choice questions related to species-specific biological knowledge. I will provide you with the name of a **species**. Based on this, generate a **multiple-choice question** that satisfies all constraints below, and output the result strictly in **JSON** format.

1. **Question Content**: The question must focus on **ONE** distinctive biological fact about the given species. Avoid generic traits shared across higher taxa. Avoid trivial, ambiguous, or overly broad questions.
2. **Options**: Provide exactly four options labeled "A" through "D". Exactly one option must be correct. Distractors must be plausible, biologically reasonable, and distinct from the correct answer. Avoid obviously wrong, irrelevant, or nonsensical options.
3. **Correct Answer**: Clearly indicate which option is correct.
4. **Explanation**: Provide a concise, scientifically accurate explanation. Justify why the correct option is true. Briefly clarify why incorrect options are invalid.
5. **Category**: Classify the question under one of the following 8 categories.
   - Life Cycle: Reproduction mode, developmental stages, metamorphosis, lifespan.
   - Behavior: Feeding, communication, mating, migration, or social behavior.
   - Geographic Distribution: Native range or natural habitats.
   - Ecological Role: Trophic function or ecological interactions (predator/herbivore/decomposer/etc.).
   - Socioeconomic Importance: Human-related economic uses (trade, medicine, food, tourism).
   - Threats: Threats faced, caused, or posed by the species.
   - Cultural History: Traditional, symbolic, or historical roles of the species.
   - Conservation: Conservation status, protections, or restoration programs.

Output JSON format:

```
{
 "Species": "<species name>",
 "Question": "<question>",
 "Options": { "A": "<Option A>", "B": "<Option B>", "C": "<Option C>", "D": "<Option D>"},
 "Correct_Answer": "<A/B/C/D>",
 "Explanation": "<concise explanation>",
 "Category": "<chosen category>"
}
```

Few-shot Examples: {Examples}

Now generate the multiple-choice question based on following description: {description}

Figure 10: Illustration of the LLM-based prompt used to generate multiple-choice questions (MCQs). This prompt leverages few-shot examples derived from curated species descriptions to produce biologically grounded, diverse, and well-structured MCQs.

For hierarchical label sampling, we first collect all sibling labels under the same parent taxonomic category. We then compute their similarity to the ground-truth label using BioCLIP Stevens et al. (2024), enabling us to select confusing labels that are maximally similar within the same hierarchy level.

For species-level label sampling, we design three difficulty settings based on taxonomic distance:

- **Hard setting**: We preferentially sample species from the closest taxonomic neighborhood (e.g., same genus), expanding to higher taxonomic levels only when necessary.



(a) The word cloud for Profile. (b) The word cloud for Conversation. (c) The word cloud for Reasoning.

Figure 11: Word-cloud visualizations for the three subsets of BioInstruct.

Table 5: Evaluation results of 25 existing LVLMs on BioBench. MAR = Morphological Attribute Recognition, THC = Taxonomic Hierarchy Classification, FSI = Fine-grained Species Identification, BKR = Biological Knowledge Reasoning, HKR = Humanistic Knowledge Reasoning. The terms *disc.*, *distr.*, *econ.* and *cons.* are short for discriminative, distribution, economic and conservation. The best performance is highlighted in red, and the second-best is shown in blue.

Model	Size	MAR		THC				FSI			BKR				HKR				Avg.
		base	*disc.*	*class*	*order*	*family*	*genus*	*easy*	*mid*	*hard*	*life*	*behavior*	*distr.*	*ecology*	*econ.*	*threats*	*culture*	*cons.*	
Open-Source Generic MLLMs																			
LLaVA-v1.5	7B	55.79	37.67	65.98	46.28	33.20	36.50	47.15	33.86	27.44	46.42	42.31	42.39	47.00	46.85	32.58	51.49	30.63	42.72
LLaVA-v1.5	13B	70.26	46.58	76.03	48.62	34.44	36.36	46.47	37.13	30.83	48.91	46.63	43.69	54.89	52.45	41.57	55.97	30.63	48.14
LLaVA-Next	7B	62.11	47.95	70.52	44.90	32.37	37.33	44.53	33.86	26.17	42.06	46.63	41.42	46.69	46.15	38.76	49.25	30.18	44.13
LLaVA-Next	13B	66.58	45.89	68.18	46.42	31.82	32.09	45.67	36.68	27.33	50.47	50.00	41.10	51.10	48.25	38.20	52.99	28.83	45.57
InternVL2.5	8B	50.00	48.63	48.35	35.95	24.38	31.54	29.38	27.65	21.78	41.74	34.13	37.54	45.11	37.76	30.34	55.97	26.58	37.14
InternVL3	14B	51.32	54.79	54.82	40.91	27.96	33.88	32.46	33.75	25.33	42.37	34.13	36.57	43.85	41.26	33.71	70.90	29.73	40.39
InternVL3	38B	57.63	57.53	48.90	37.33	26.31	35.12	33.03	32.39	26.61	42.99	30.77	36.57	45.74	42.66	33.71	66.42	28.83	40.89
Qwen2-VL	7B	78.42	48.63	78.79	61.57	40.36	38.71	61.89	44.02	31.80	68.22	57.21	54.05	67.51	55.94	47.19	72.39	39.64	56.21
Qwen2.5-VL	3B	80.53	56.55	78.51	61.98	42.15	39.67	58.48	41.87	33.24	62.31	66.83	55.02	65.93	50.35	51.69	71.64	40.99	57.02
Qwen2.5-VL	7B	81.58	61.38	84.99	64.05	43.11	41.60	62.68	45.49	34.85	68.54	59.62	52.10	70.98	62.24	51.69	76.12	41.44	59.48
Qwen2.5-VL	32B	84.47	63.70	88.83	69.42	50.55	47.80	68.30	48.14	38.35	77.88	67.31	63.43	76.03	65.73	60.11	82.09	44.59	64.62
Qwen2.5-VL	72B	85.53	69.18	92.56	72.31	51.79	50.55	69.25	50.00	40.33	79.75	74.04	65.37	76.34	66.43	58.99	76.87	50.00	66.55
GLM-4V	9B	81.58	57.53	87.47	61.16	41.46	38.29	59.91	41.87	33.50	63.86	61.06	53.40	66.56	61.54	46.63	66.42	39.19	57.33
Kimi-VL-Instruction	16B	80.53	69.18	79.48	59.09	43.66	42.01	56.61	41.53	32.74	59.19	62.02	51.78	66.25	51.75	46.63	66.42	36.94	56.47
Phi-4-multimodal-instruct	5.6B	73.42	48.63	71.35	49.31	31.13	33.61	47.21	33.86	27.40	62.50	53.14	76.47	61.95	57.65	45.45	68.66	38.30	51.91
CogVLM2	19B	76.58	47.95	77.41	53.04	36.91	35.81	51.82	42.78	31.61	54.52	61.06	49.84	56.47	50.35	44.38	61.19	31.98	51.72
Emu3-Chat	8B	61.84	50.00	50.41	27.13	24.66	24.93	34.58	27.43	22.96	37.69	38.94	34.63	36.28	37.76	31.46	41.04	29.28	37.59
mPLUG-Owl3	7B	73.68	52.74	72.18	51.93	33.61	34.57	49.15	35.21	27.29	60.12	55.77	45.95	61.20	46.85	48.88	64.18	37.84	50.86
Llama-3.2-Vision	11B	70.00	52.74	77.55	60.74	40.22	42.15	52.56	39.28	32.30	60.75	61.54	56.31	59.94	46.85	44.38	64.93	39.19	53.23
Ovis2	34B	83.95	64.38	89.26	64.33	49.31	45.18	64.62	45.82	38.91	78.50	68.75	57.93	76.97	62.94	59.55	81.34	47.30	63.78
Open-Source Reasoning MLLMs																			
Kimi-VL-Thinking	16B	77.89	60.96	79.75	54.68	37.33	36.64	54.04	38.71	29.96	66.67	60.10	43.04	60.88	60.84	51.69	66.42	37.84	54.59
R1-VL	7B	71.84	48.63	78.79	55.79	38.15	31.40	56.54	41.42	28.74	62.31	55.77	48.54	63.41	58.04	42.70	61.19	36.94	52.13
Vision-R1	7B	67.63	43.15	85.67	64.60	41.32	42.56	60.07	44.13	31.63	66.04	62.02	49.51	65.62	58.04	47.19	63.43	36.94	54.25
R1-Onevision	7B	73.95	58.22	82.78	61.57	36.50	37.33	57.45	42.89	32.91	62.93	57.21	47.25	57.73	55.24	42.70	52.99	38.74	53.59
DeepPerception-FGVR	7B	71.05	56.16	71.07	59.64	39.39	38.29	55.52	41.65	30.41	54.83	47.60	44.66	54.26	44.76	43.82	64.93	35.14	50.88

- **Mid setting**: We sample one species each from the same genus, family, and order as the ground-truth species.

- **Easy setting**: We sample species from progressively broader categories: family, order, and class.

Within each difficulty group, we further refine label selection using BioCLIP Stevens et al. (2024) similarity scores to ensure fine-grained and consistent difficulty control. Examples of different difficulty settings are shown in Figure 7.

A.2.4 Semi-Automatic Curation of Knowledge Reasoning Data

To construct biological and humanistic knowledge reasoning data, we develop a semi-automatic pipeline that combines template-based design and LLM-based generation. We first collect species-level knowledge from Wikipedia and the IUCN API, and use an LLM to integrate these heterogeneous sources into comprehensive species descriptions. Based on these descriptions, we manually design question templates covering diverse reasoning patterns by filling the corresponding knowledge fields. Meanwhile, LLM-based questions are generated through few-shot prompting with examples derived from the generated descriptions, enabling more diverse question formats while preserving biological grounding. The generated questions are parsed into structured multiple-choice formats, and the detailed quality control process is described in the paper. Representative templates and the MCQ-generation prompt are provided in Figure 9 and Figure 10.

A.3 BioInstruct Visualization

To further illustrate the content diversity of BioInstruct, we visualize the most frequently occurring words in the Description, Conversation, and Reasoning subsets using word-cloud maps, as shown in Figure 11. The word distributions reveal distinct characteristics across different subsets. The Description subset mainly contains species-level biological concepts, including taxonomic categories, morphological attributes, habitats, and ecological information. The Conversation subset reflects diverse instruction-following patterns for interacting with biological knowledge, while the Reasoning subset highlights concepts related to biological inference, adaptation, and conservation reasoning. These observations demonstrate that BioInstruct provides broad coverage of both factual biological knowledge and higher-level reasoning patterns. Representative instruction samples are further presented in Figure 12.

A.4 Supplementary Evaluation Results

We conduct zero-shot evaluations across 25 MLLMs on BioBench, with full results provided in Table 5. Additional large-scale models, such as Qwen2.5-VL-32B Bai et al. (2025b), Qwen2.5-VL-72B Bai et al. (2025b), and Ovis-32B Lu et al. (2024), are also included. Future work will extend the evaluation to stronger closed-source MLLMs. Results indicate that the newly added large-scale models exhibit consistent and substantial improvements across nearly all evaluation dimensions. For instance, scaling Qwen2.5-VL from 3B to 72B parameters leads to steady and monotonic performance gains across most benchmarks, reflecting the strong impact of increased model capacity. Despite these overall improvements, the relative performance gaps between adjacent taxonomic levels remain largely preserved, suggesting that scaling primarily enhances general representational strength rather than fundamentally altering the models' understanding of biological structure.

A.5 More Case Study

In this section, we present additional case studies to further analyze model behaviors and failure patterns under different task settings. As shown in Figure 13, models can often recognize visually similar species when explicit candidate references are provided, benefiting from relative visual comparisons. However, without candidate references, they struggle to associate subtle morphological characteristics with species-level concepts and often make predictions at broader taxonomic levels or confuse closely related species. Figure 14 illustrates that biological knowledge reasoning is highly dependent on accurate species grounding. When species recognition is reliable, models can effectively leverage prior knowledge for reasoning about biological attributes, ecology, or conservation; otherwise, recognition errors propagate to subsequent reasoning and lead to incorrect conclusions. Figure 15 further highlights challenging cases involving species with subtle visual differences, where current MLLMs tend to rely on coarse visual patterns and fail to integrate fine-grained perception with biological knowledge. These examples reveal persistent limitations of existing models in achieving biologically grounded multimodal understanding.

A.6 Limitations and Future Work

A.6.1 BioInstruct Data

Although we collected additional species information, knowledge for certain species remains extremely scarce and difficult to obtain. For these species, the generated descriptions rely heavily on the GT label prior to guide high-capacity models, which may lead to potential inaccuracies. Future work should explore more refined description-construction pipelines and post-processing procedures for quality verification and error correction. Moreover, due to resource constraints, we currently include only 160K instances covering approximately 50,000 species. Subsequent work may further expand both the diversity of species categories and the overall data scale.

A.6.2 Model Architecture

The models used in BioBench rely on existing MLLM architectures, which are not necessarily optimized for capturing fine-grained biological attributes or performing hierarchical taxonomic reasoning. Future work could explore architecture-level improvements, such as integrating a biology-

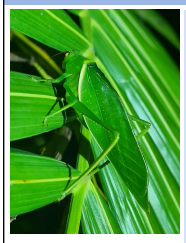
Profile



The Boreal Owl

User: How can the species in the photo be described in detail?
Assistant: The Boreal Owl, also known as the Tengmalm's Owl, is a small to medium-sized owl chara-cterized by its distinctive appearance and adaptations for life in boreal forests. **Morphologically**, it exhibits a compact body with a relatively large head and small, rounded wings, which contribute to its agile flight capabilities in dense forest habitats. Its plumage is predominantly brownish-gray with intricate white spotting ... This species belongs to the order **Strigiformes** and the family **Strigidae**, placing it among the "typical owls." The Boreal Owl is primarily distributed across the boreal and subarctic regions of **North America and Eurasia** ... The Boreal Owl is less vocal han many other owl species and is more active during dusk and dawn, though it can also be heard calling at night.**Ecologically**, it plays a crucial role as a mesopredator, helping to regulate populations of small rodents and contributing to the balance of forest ecosystems. Despite its wide dist-ribution, the species faces challenges from **habitat fragmentation and climate change**... **Interactions with humans** are minimal, as the owl is typically elusive and prefers remote forested areas, it is **occasionally** observed in managed forestry landscapes ...Overall, the Boreal Owl is a fascinating example of avian adaptation to **cold, forested** ...

Conversation



Serrated Bush Katydid

User: What would this insect be called?
Assistant: Serrated Bush Katydid (Paracaedicia serrata)
User: What are the key physical features that distinguish this insect from other similar species?
Assistant: **Bright green coloration**, serrated wing edges, and leaf-like body shape ...
User: To which family does this insect belong, and what are some related groups?
Assistant: Belongs to the **Tettigoniidae** family; related to bush crickets and long-horned ...
User: How does this species typically communicate or attract mates?
Assistant: Uses **stridulation** (rubbing wings together) to produce sounds for mating calls.
User: What are the primary threats to this species, and is it protected?
Assistant: Habitat loss from **deforestation**; ... but populations are **declining**.
User: Describe its typical behavior during the day versus night.
Assistant: **Nocturnal**; rests motionless on leaves during the day, becomes active at night ...

Reasoning



Turquoise Bluet

User: What is the taxonomic rank of the blue insect shown in the image?
Assistant: The insect in the image has a slender, elongated body with a **vibrant turquoise-blue coloration** on its thorax and abdomen. Its transparent wings are held outstretched, a characteristic feature of **damselflies**. Additionally, its segmented legs and large **compound eyes** are typical of the order **Odonata**. Based on these visual traits, the species can be identified as part of the family Coenagrionidae within the order Odonata. This confirms it is a damselfly, specifically the Turquoise Bluet (Enallagma divagans).

Figure 12: Representative BioInstruct examples, illustrating the typical structure and content of the Profile, Conversation, and Reasoning subsets.

oriented visual encoder (e.g., BioCLIP Stevens et al. (2024)), incorporating taxonomy-aware modules, enhancing cross-modal fusion mechanisms, or designing hierarchical reasoning heads tailored to biodiversity-related tasks.

A.6.3 Reasoning Model

Current reasoning models have been shown to be suboptimal for tasks in BioBench, possibly because they have not yet learned to integrate fine-grained visual cues with hierarchical taxonomic reasoning. These models typically excel at general reasoning but struggle when subtle morphological differences or taxonomy-dependent relations are required. Future work should focus on exploring how to jointly model fine-grained visual representations and structured taxonomic knowledge, such as through taxonomy-aware training objectives, multi-hop hierarchical supervision, or reasoning modules explicitly designed for biological classification. Such developments may enable reasoning models to more effectively handle visually similar species and long-tail categories in biodiversity tasks.



Figure 13: Additional case studies illustrating model behavior under close-ended versus open-ended conditions. In closed-ended settings, the models can leverage similarity comparisons among candidate labels to arrive at the correct answer. However, in open-ended settings, they can only roughly identify the broader taxonomic group and struggle to distinguish between fine-grained species.

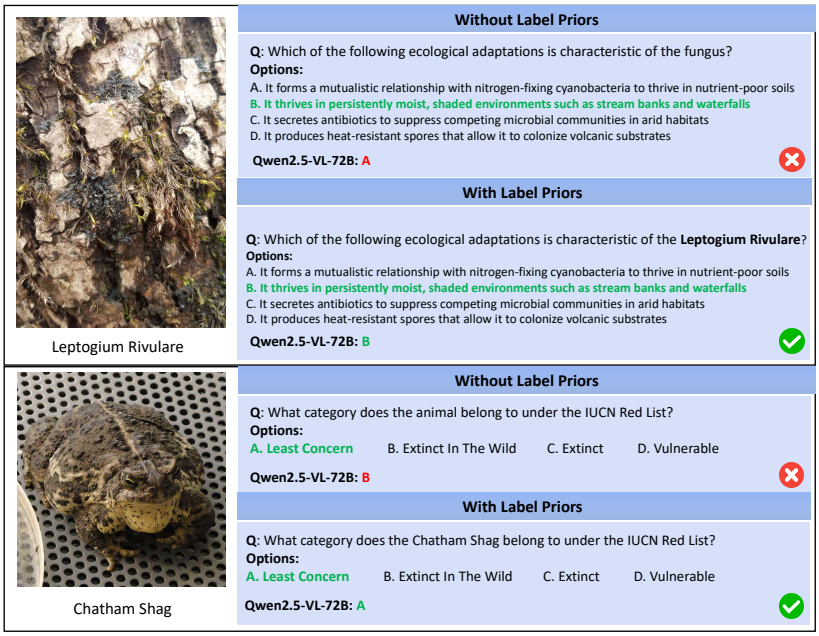


Figure 14: A case study highlighting the role of precise species identification in knowledge-based reasoning tasks. Correctly associating the query with the underlying species is crucial for accurate answers. The figure illustrates both the strengths of current models in leveraging prior knowledge and their limitations when species recognition is ambiguous or challenging.

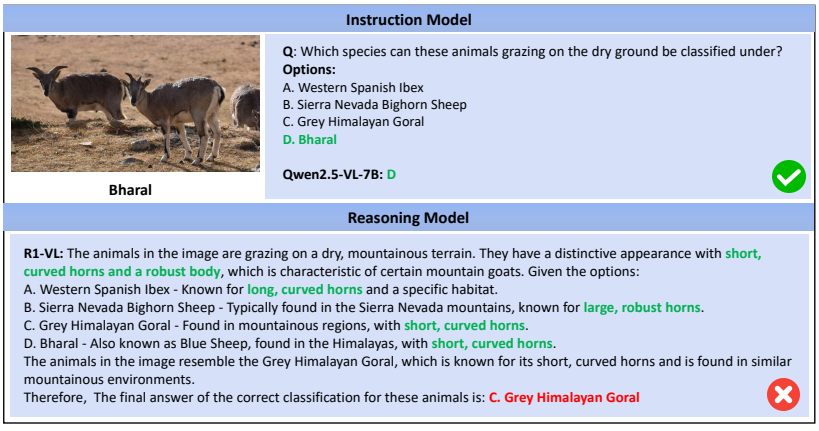


Figure 15: A case study showing limitations of current reasoning MLLMs in fine-grained species identification. Although models compare visual features among candidate options, they relies on coarse visual cues and still fails to distinguish highly similar species.