WILDBIOWIKI: A SPECIES-CENTERED BENCHMARK FOR WILDLIFE QUESTION ANSWERING WITH CONTROLLED INPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Wildlife question answering requires linking visual observations to species knowledge. Comparing models under different information conditions requires common questions, reference answers, and supporting evidence. We introduce WildBioWiki, a species-centered benchmark that organizes these links for reuse. It connects 264,563 images of 392 vertebrate species to 3,136 unique species-level question–answer pairs and retained evidence, alongside image-specific silver captions. Sharing fixed questions across species images induces 2,116,504 image-conditioned QA instances. The aligned records support matched comparisons that vary access to images, species names, and knowledge while preserving the biological target. We establish broad baselines with twelve zero-shot models. Targeted evaluations use balanced claims, matched facts, and human review to examine what these comparisons establish. Human checks support the observed information effects and expose local errors hidden by high automatic agreement; matched-fact comparisons leave an additional combination penalty unresolved. WildBioWiki provides reusable targets and matched input protocols for studying wildlife question answering and interpreting performance.

1 INTRODUCTION

A wildlife assistant may recognize an animal yet describe the wrong diet, or attach correct ecology to the wrong species. Comparing these outcomes requires common biological targets: images, questions, reference answers, and evidence must remain aligned as input information changes. WildBioWiki organizes these links so that researchers can reuse the same targets across controlled comparisons.

Natural-world benchmarks evaluate recognition, distribution shift, and retrieval (Beery et al., 2018; Van Horn et al., 2021; Vendrow et al., 2024). Knowledge-intensive visual question answering (VQA) already links images to external facts (Marino et al., 2019; Chen et al., 2023; Mensink et al., 2023), while biological benchmarks assess species and traits (Maruf et al., 2024). WildBioWiki contributes a fixed ecological question schema coupled to matched input protocols, allowing researchers to compare information conditions without rebuilding question targets and their evidence alignment.

Each species has eight questions shared across its images; captions describe individual photographs. This distinction preserves the species-level target when image correspondence, supplied identity, or knowledge access changes. Retained evidence makes reference answers inspectable, while species-aware aggregation accounts for shared questions and repeated appearances.

Our contributions are threefold:

(1) A species-centered image–knowledge resource. WildBioWiki links 264,563 images from 392 vertebrate species to 3,136 species-level question–answer pairs and supporting evidence. Fixed questions provide reusable targets across images, with unique knowledge annotations distinguished from induced QA instances.

(2) Matched protocols for studying information use. We define conditions that vary images, species names, and retained knowledge while keeping questions and reference targets fixed. The linked records support comparisons without reconstructing the image–question–evidence alignment.

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.

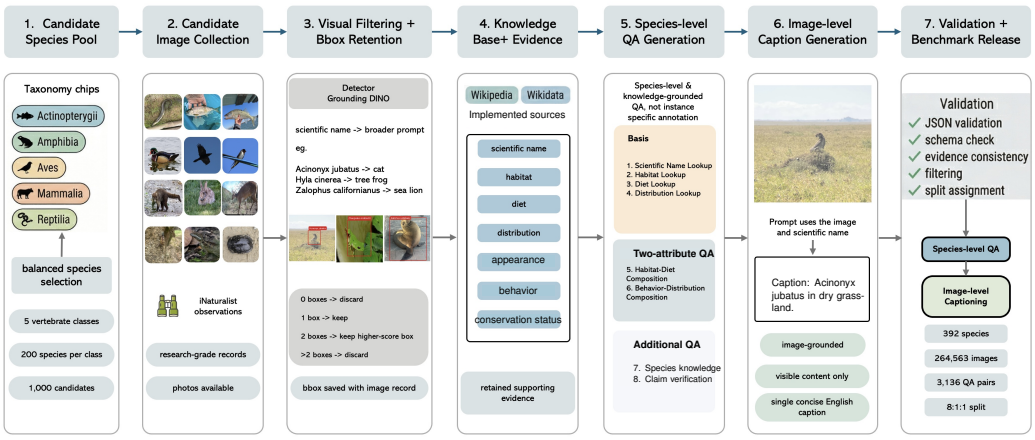


Figure 1: Aligning images with a shared biological target. The construction pipeline links species metadata, Wikipedia text, and Wikidata-linked records. Retained knowledge supports species-level QA; images and scientific names support caption references. Question slots are defined in Table 2 and audited in Section 6.

(3) Baseline results and targeted validation. We evaluate twelve zero-shot models and study balanced claim verification, matched-fact answering, and human–automatic agreement. These analyses demonstrate benchmark uses and clarify which performance interpretations the evidence supports.

2 Related Work

Knowledge-intensive VQA uses external facts (Marino et al., 2019; Schwenk et al., 2022; Wang et al., 2017). InfoSeek evaluates visual information seeking with and without a knowledge base, including entity-recognition pipelines (Chen et al., 2023). Encyclopedic-VQA pairs questions with multiple category images, retains Wikipedia evidence, and evaluates oracle retrieval (Mensink et al., 2023). These precedents establish image–knowledge alignment and evidence access. WildBioWiki builds on them with eight fixed ecological slots across its species inventory and five matched input conditions on shared targets. Researchers can reuse the question–answer–evidence mapping and condition definitions when testing another model, rather than reconstructing these choices.

Natural-world resources support recognition and retrieval (Van Horn et al., 2021; Vendrow et al., 2024). VLM4Bio evaluates species classification and visible-trait identification, grounding, referring, and counting (Maruf et al., 2024); our species-level ecology targets address facts that need not be visible in a photograph. Alongside retrieval-augmented VQA (Qiu et al., 2024; Choi et al., 2025), our supplied-record condition is a reference for answering with evidence, not retrieval quality. Rubric scores (Chan et al., 2023), balanced claims, matched facts, and human checks connect the resource to inspectable outcomes. Appendix G details the comparison; Appendix **??** maps the files and reuse steps.

3 WildBioWiki: Resource Design and Construction

3.1 Species-level evidence as the shared target

The reusable unit is a species record: its questions, answers, and supporting evidence are linked to multiple images. Captions describe individual photographs; shared questions fix the biological target across images. The same retained evidence supports reference construction and knowledge-supplied comparisons. For species s, let $\mathcal{I}_s$ be its retained images and $\mathcal{K}_s$ its structured knowledge record. Each image $x \in \mathcal{I}_s$ has a silver caption C_x, while the species has a shared question set

$$\mathcal{Q}_s = \{(q_{sj}, a_{sj}, e_{sj})\}_{j=1}^{8}, \qquad e_{sj} \subseteq \mathcal{K}_s. \tag{1}$$

Table 1: WildBioWiki inventory. Induced QA instances reuse species-level questions across images; they are not independent knowledge annotations.

Resource unit	Count	Descriptive mean	Value
Initial / qualified / aligned species	1,000 / 396 / 392	Images per species	674.91
Images / silver captions	264,563 / 264,563	Question words	12.58
Unique species–question pairs	3,136	Answer words	18.45
Induced image–question instances	2,116,504	Caption words	15.60

Here a_{sj} is the reference answer and e_{sj} retains supporting text and field provenance. Closed-book QA takes (x, q_{sj}) as input; the species identity and knowledge are withheld unless a control condition supplies them. Captioning takes the image and asks for a species-aware description. Reference construction can use the known scientific name, so it must be distinguished from the information available to the evaluated model.

We begin with 1,000 common species ranked by observation volume, with 200 candidates per vertebrate class. Wikipedia text and Wikidata-linked metadata are normalized into core fields covering appearance, habitat, diet, behavior, distribution, and conservation status, together with up to three non-overlapping extra knowledge blocks. Unsupported fields remain empty. Extraction retains supporting text, removes duplication, and excludes local anecdotes and overlapping extras. Requiring at most two missing core fields and at least one evidence-backed extra block leaves 396 knowledge-qualified species. The resulting records have four to six populated core fields and one to three extras, favoring documented, commonly observed species.

3.2 Image selection and cross-taxon coverage

To reduce ambiguous framing, we use Grounding DINO (Liu et al., 2024) with broader visual prompts. Images with no detected boxes or more than two boxes are discarded; among two boxes, the higher-confidence box is retained. This heuristic reduces scene ambiguity without independently verifying fine-grained identity. Appendix A gives the prompt mapping and detector settings.

The rule removes 16.2% of candidate images and changes any vertebrate class's image share by at most 2.57 percentage points (Appendix A). Requiring at least 300 retained images per species then removes four knowledge-qualified candidates, yielding 392 aligned species. Table 1 distinguishes sampling units; Figure 3 reports annotation composition. The image split contains 211,492/26,283/26,788 training/validation/test images. A separate, taxonomically stratified species-disjoint protocol assigns 274/39/79 species to training/validation/test, keeping every image and question for a species together. The broad evaluation is zero-shot; the species-disjoint protocol supports future supervised comparisons.

3.3 Grounded questions and caption references

Figure 2 illustrates the two tasks. QA references use species knowledge rather than instance-specific visual facts. Questions omit species names and retain supporting text and field tags. Construction review and repairs cover all 396 candidate records; structural checks cover 3,168 candidate questions (Appendix A). Table 2 defines the fixed slots. Their requested content provides broad coverage of identity and species attributes; Section 6 tests what their operational complexity supports.

Caption references are generated from the image and scientific name as one concise sentence restricted to visible content. Full-set automatic checks cover name retention and camera wording (Appendix A). A taxonomically stratified human reference audit accepted 491/500 QA references (98.2%) and 471/500 captions (94.2%); caption review used the source image. These are reference-quality estimates, separate from assessment of model outputs. The captions remain silver annotations, and automatic validation does not replace exhaustive visual review.

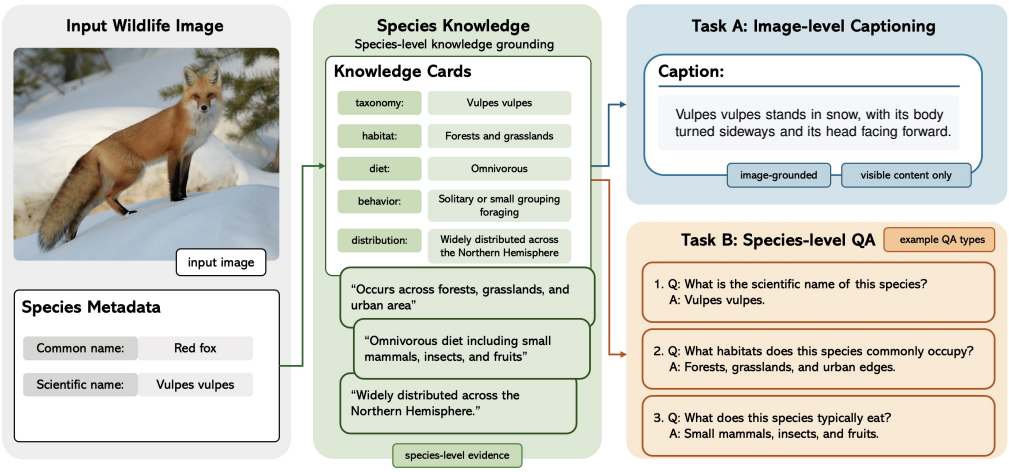


Figure 2: Two tasks with different evidence needs. The caption describes the visible fox and snow, whereas the shared QA references describe species-level habitat and diet. The schematic illustrates how image-specific descriptions and biological knowledge coexist in one record.

Table 2: The eight fixed question slots. Slot labels specify requested content; they do not certify an internal reasoning operation. The original Q7/Q8 grouping is retained for traceability and audited separately.

Slot	Requested content	Interpretation
Q1	Scientific name	Species identity; excluded from name-given contrasts
Q2–Q4	Habitat; diet; distribution	Individual species attributes
Q5–Q6	Habitat + diet; behavior + distribution	Multiple requested attributes
Q7	Core/extra species knowledge	Additional knowledge; operational complexity audited
Q8	Claim decision and explanation	Verification; original answer polarity audited

4 Evaluation Protocols for Reusing Aligned Targets

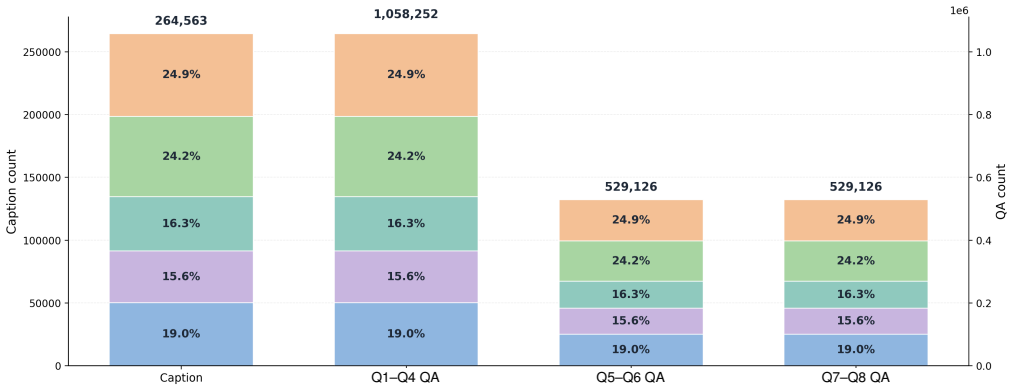
4.1 Species-aware scoring for broad comparisons

We evaluate twelve zero-shot systems listed in Table 4, with full model identifiers in Appendix B. None is trained on WildBioWiki. Retained local runners use greedy decoding and a 512-token output cap, except LLaVA-OneVision, which uses 192. Broad evaluation establishes a reference point; matched studies test information effects on fixed targets.

The broad semantic metric, LLM-J, uses a frozen text-only GPT-5.4-mini judge (snapshot 2026-03-17; rubric configuration 2026-03-30). It compares captions with silver references and QA answers with evidence-grounded references. Its anchored scores are 5 (wrong), 35 (weak with material errors), 65 (partial but core-correct), 80 (good), and 95 (correct and complete). Acc@65 records whether an individual response reaches the first core-correct band. These are rubric scores, not percentages of facts recovered. For model m, score r_{msjx}, and evaluated images $\mathcal{V}_s$,

$$\bar{r}_{msj} = \frac{1}{|\mathcal{V}_s|} \sum_{x \in \mathcal{V}_s} r_{msjx}, \quad R_m = \frac{1}{8|\mathcal{S}|} \sum_{s \in \mathcal{S}} \sum_{j=1}^{8} \bar{r}_{msj}. \tag{2}$$

Acc@65 applies the same aggregation to $\mathbf{1}[r_{msjx} \geq 65]$, rather than thresholding the mean score. Caption LLM-J is image-averaged. QA intervals resample species clusters 2,000 times (seed 2026071804), retaining all questions and images within each sampled species. Thus the brackets



	Caption	Q1–Q4 QA	Q5–Q6 QA	Q7–Q8 QA
Actinopterygii	50,171 (19.0%)	200,684 (19.0%)	100,342 (19.0%)	100,342 (19.0%)
Amphibia	41,221 (15.6%)	164,884 (15.6%)	82,442 (15.6%)	82,442 (15.6%)
Aves	43,134 (16.3%)	172,536 (16.3%)	86,268 (16.3%)	86,268 (16.3%)
Mammalia	64,060 (24.2%)	256,240 (24.2%)	128,120 (24.2%)	128,120 (24.2%)
Reptilia	65,977 (24.9%)	263,908 (24.9%)	131,954 (24.9%)	131,954 (24.9%)

Figure 3: Annotation inventory across five vertebrate classes. QA bars count image-conditioned instances of the indicated question slots. Each species shares four Q1–Q4, two Q5–Q6, and two Q7–Q8 questions across its images; the grouping describes the schema, not a hierarchy of reasoning abilities.

in Table 4 quantify uncertainty in an aggregate estimate, not the minimum and maximum species scores.

4.2 Matched inputs and targeted task controls

The base control set fixes 100 species, one image per species, and Q2–Q8, giving 700 matched items per condition for Qwen3-VL, SAIL-VL, and Gemma-3. Question-only omits the image; closed-book supplies image and question; name-given additionally supplies the correct scientific name; KB-oracle further supplies retained knowledge; shuffled-image replaces the image with one from another species. Q1 is excluded because giving the name would directly answer it. A wrong-name diagnostic covers 20 species and 140 items. These comparisons estimate input effects under fixed prompts and items, without uniquely identifying internal processes.

Two extensions test balanced verification and matched-fact correctness (Table 3). Balanced Q8 uses one true and one false claim for each of 100 species. Both members refer to the same evidence, and the false member changes a scoped fact to an incompatible alternative. Absence from the record alone is not treated as falsity. Each of three models answers both claims under closed and KB-oracle conditions, producing 1,200 responses. The second extension asks about habitat, diet, and their combination for the same 100 species. It holds image, name, and knowledge constant and yields 900 responses. All separate and combined responses are newly generated in the same CUDA environment for each model, avoiding a mixture of historical and new runs. Together, the extensions provide 2,100 generated and scored responses.

4.3 Outcome-specific scoring and human review

Balanced Q8 separates *decision accuracy*, *supported-explanation rate*, and *joint accuracy*, which requires both. Decisions use explicit leading Yes/True, No/False, or Uncertain labels; all 1,200 outputs are parseable, and uncertainty earns no correctness credit. A bare decision earns no explanation credit. Explanation validity concerns the tested fact, not every incidental assertion. We use direct parsing because the judge mis-extracted 120/1,200 decisions; raw fields and this post-detection correction are documented in Appendix D.

Table 3: Evaluation scopes. Targeted studies test specific explanations of the broad results. Human labels cover a fixed subset of the new experiments.

Study	Species	Models	Unit and scope
Broad evaluation	392	12	Validation images; eight questions/species
Base input controls	100	3	700 Q2–Q8 items per full condition
Balanced Q8	100	3	200 claims × two conditions/model
Matched facts	100	3	Two single questions + one combined/model
Human main / overlap	20 / 5	3	420 / 105 responses; overlap counted once

Table 4: Broad-validation results under the frozen rubric. QA uses species–question macro-averaging and 95% species-cluster intervals; caption scores are image-level reference-agreement means. Acc@65 is in percent. New binary and fact-level studies use distinct metrics and samples.

Model	Caption LLM-J	QA LLM-J [95% CI]	QA Acc@65 [95% CI]
InternVL3.5	45.7	32.4 [31.3, 33.5]	36.2 [34.7, 37.8]
Ovis2.5	34.3	41.7 [40.5, 43.0]	50.0 [48.3, 51.8]
Qwen3-VL	**55.9**	**45.5 [44.4, 46.5]**	**54.6 [53.1, 56.1]**
SAIL-VL	53.9	42.2 [41.1, 43.3]	50.0 [48.4, 51.7]
Gemma-3	51.2	34.8 [33.7, 35.9]	40.4 [38.9, 42.0]
Phi-4-MM	43.9	31.7 [30.8, 32.5]	35.0 [33.8, 36.3]
LLaVA-OV1.5	51.2	37.9 [36.8, 39.0]	43.9 [42.3, 45.6]
MiniCPM-V4.5	53.0	41.8 [40.7, 42.8]	49.6 [48.0, 51.1]
Idefics3	44.5	28.2 [27.4, 29.0]	28.1 [27.0, 29.3]
Molmo-7B-D	39.2	27.0 [25.9, 28.1]	27.9 [26.4, 29.4]
GPT-5.4-nano	40.1	32.0 [31.0, 33.2]	36.6 [35.2, 38.2]
Grok-4.1-fast	26.2	37.5 [36.0, 39.1]	43.8 [41.8, 45.9]

The matched-fact study labels each required proposition as correct, omitted, contradicted, or uncertain. The same supported facts are scored in separate and combined responses. Each model has 201 biological fact atoms per comparison condition; 13 record-insufficiency diagnostics are scored separately. A stricter completeness outcome requires all necessary facts and no incompatible extra claims. It compares *both* single answers jointly against the combined answer on the same 87 eligible species, rather than comparing a two-fact answer against one single answer.

Human review covers four species per class (20 species; 420 responses), with a second-reviewer overlap of five species and 105 responses. Reviewers did not see automatic scores; six disagreements were adjudicated without double-counting overlap. Confirmed-correct rates retain fixed denominators, while binary agreement excludes indeterminate labels. Human intervals use 2,000 paired, class-stratified species bootstrap samples (seed 20260908); full automatic intervals use unstratified paired species resampling. Appendix E details sample counts, uncertainty, and pairing.

5 Baseline Results and Controlled Input Comparisons

5.1 Broad performance across twelve models

Table 4 establishes broad baselines: Qwen3-VL leads QA LLM-J (45.5), Acc@65 (54.6%), and caption reference agreement (55.9). Small point-estimate differences do not establish reliable rankings. The text-only caption judge measures reference agreement rather than visual factuality. The broad-evaluation human output audit reports binary agreement of 81.7% on 1,000 QA responses and 86.4% on 1,000 captions. These results support aggregate agreement with the thresholded rubric, not validation of every five-band score difference or model ranking.

The base question groups vary across taxa. Averaged over twelve models, Q5–Q6 scores range from 28.9 for amphibians to 38.0 for birds, below the other base groups in each class (Appendix C). These differences motivate two questions: does supplying identity or evidence improve the same answers,

Table 5: Base matched input controls: species-macro LLM-J on 100 species and Q2–Q8 (700 items per condition). KB adds retained evidence to name-given input. Values are rubric points, not percentage accuracies.

Model	Question-only	Closed	Name-given	KB-oracle	Shuffled
Qwen3-VL	21.7	43.7	53.8	84.9	26.4
SAIL-VL	15.0	41.2	47.0	82.7	26.5
Gemma-3	10.7	37.7	43.1	79.8	23.1

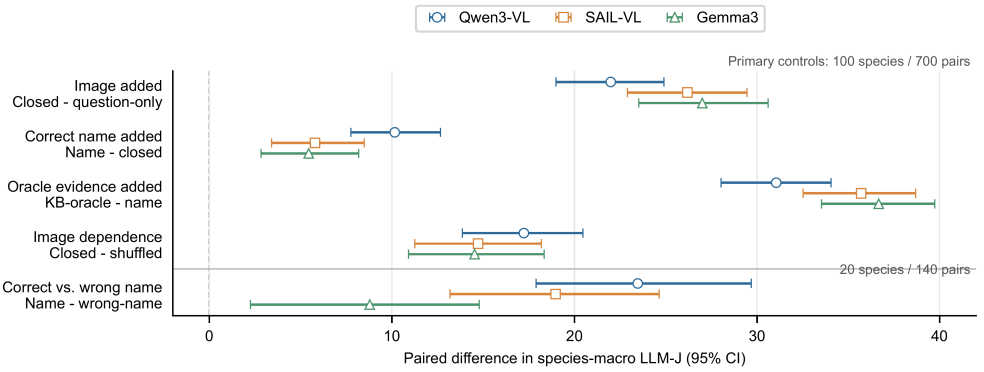


Figure 4: Paired input contrasts. Points are species-macro LLM-J differences with 95% species-cluster bootstrap intervals (2,000 replicates). Primary controls use 100 species and 700 Q2–Q8 items; wrong-name uses 20 species and 140 items. Oracle evidence is supplied directly, not retrieved.

and does requesting multiple attributes reduce correctness when the required facts are matched? The following experiments address these questions directly, accounting for input access and task content.

5.2 Matched comparisons across information conditions

The aligned records let us compare input conditions on the same 700 questions and reference targets. All three models benefit from information linking a question to the correct species and its knowledge. Adding the correct image to question-only input improves LLM-J by 22.0–27.0 points. Supplying the correct scientific name adds 5.4–10.2 points, and adding retained knowledge beyond the name adds a further 31.1–36.7 points. Shuffling the image lowers closed-book scores by 14.5–17.2 points (Table 5; differences use unrounded estimates). Paired intervals exclude zero for each comparison (Figure 4). The wrong-name diagnostic also lowers scores relative to correct names by 23.5, 19.0, and 8.8 points for Qwen3-VL, SAIL-VL, and Gemma-3, respectively. Within this input sequence, retained knowledge gives the largest additional gain, even after the correct name is supplied. These contrasts measure input benefits, not a decomposition of internal abilities.

The identity and evidence benefits persist when Q8 is excluded. Reaggregating the same base-control responses under a common judge on Q2–Q7 leaves 600 items per condition. Name-given improves Acc@65 over closed-book by 16.00 [11.67, 20.33], 7.83 [2.83, 12.50], and 8.17 [3.67, 12.67] percentage points for Qwen3-VL, SAIL-VL, and Gemma-3. KB-oracle adds a further 36.17 [31.67, 41.00], 43.67 [39.33, 48.00], and 47.00 [41.83, 52.17] points. Both effects therefore persist beyond the polarity-sensitive slot under the same base-control scoring protocol.

5.3 Validating input comparisons with balanced claims

Balanced claims demonstrate a second use of the retained evidence: constructing paired verification targets with controlled answer polarity. The base Q8 references begin with No/False for 368/392 species (93.9%), motivating an extension with one true and one false claim per species. Balancing

Table 6: Balanced Q8 on all 100 species: 200 claims per model and condition. Decision (D), explanation (E), and joint (J) rates are percentages. ΔJ is KB-oracle minus closed-book with a paired 95% species-bootstrap interval, in percentage points.

	Closed-book			KB-oracle			
Model	D	E	J	D	E	J	ΔJ [95% CI]
Qwen3-VL	67.5	52.0	51.5	99.5	99.5	99.5	48.0 [40.5, 55.5]
SAIL-VL	56.5	39.5	39.0	95.0	95.5	94.5	55.5 [48.5, 62.5]
Gemma-3	59.0	43.0	41.5	86.0	88.5	85.5	44.0 [36.0, 52.0]

Table 7: Automatic (A) and human (H) effects on the same 20 species, in percentage points. Q8 compares KB-oracle with closed-book (40 claims per condition/model); facts compare combined with separate answers (44 biological atoms per condition/model). Brackets are paired 95% human species-bootstrap intervals. Q8 gains persist under human review; small fact losses remain unresolved.

	Q8 joint correctness		Biological-fact accuracy	
Model	A	H [95% CI]	A	H [95% CI]
Qwen3-VL	37.5	40.0 [25.0, 55.0]	2.3	−2.3 [−6.5, 0.0]
SAIL-VL	65.0	65.0 [52.5, 75.0]	0.0	−2.3 [−7.0, 0.0]
Gemma-3	45.0	42.5 [30.0, 55.0]	2.3	−4.5 [−10.9, 0.0]

removes this label prior, while scoring explanations requires evidence beyond a bare decision. Appendix Figure 5 illustrates the distinction between a correct decision and a supported explanation. An always-No response without explanation obtains 50% decision accuracy but 0% explanation, joint, and both-members-correct accuracy on the paired set.

Joint accuracy rises from 39.0–51.5% to 85.5–99.5% when identity and knowledge are supplied (Table 6). On the human-audited 20-species subset, the corresponding rates change from 60.0 to 100.0% for Qwen3-VL, 35.0 to 100.0% for SAIL-VL, and 45.0 to 87.5% for Gemma-3. Table 7 compares automatic and human gains on these same species: both support the intervention, and every human interval excludes zero. Adversarial assignment of unresolved human labels also preserves positive gains (Appendix E). The benefit thus extends to supported verification under balanced answer polarity and human scoring. Because this experiment supplies name and knowledge together, its effect measures their joint benefit.

6 What the Scores Establish About Task Difficulty

6.1 Q7 primarily requests directly stated facts

An author-confirmed audit of all 392 Q7 items identifies 330 single-fact lookups (84.2%), 43 multi-fact reports (11.0%), and 19 invalid or ambiguous items (4.8%). No item requires deriving a relation beyond stated facts under the audit's fact-integration criterion. We therefore interpret Q7 as additional species knowledge rather than general reasoning. Table 4 preserves the frozen eight-slot baseline. An archived sensitivity analysis removes the 19 flagged Q7 cells: total-score shifts remain below 0.08 rubric points and Acc@65 shifts below 0.12 percentage points, with unchanged point-estimate ordering. Positive name and knowledge contrasts persist. Appendix H specifies scoring versions, aggregation, and sensitivity scope; the targeted extensions use separately checked materials.

6.2 Matched facts refine the task comparison

The matched-fact experiment returns to the lower Q5–Q6 scores: does combining questions reduce correctness when both conditions require the same biological facts? Under identical images, correct names, and retained knowledge, automatic fact accuracy is 99.0–99.5% for separate questions and 99.5% for combined questions across all 100 species. The automatic comparison is near ceiling

in both conditions. Human scoring reveals small declines: on the same 44 biological facts per condition and model, accuracy falls from 97.7% to 95.5% for Qwen3-VL and SAIL-VL, and to 93.2% for Gemma-3.

Table 7 shows the human fact differences and intervals. On the 19 eligible human-audited species, complete-answer differences are -5.3, -5.3, and -10.5 points, also with intervals including zero (Appendix E). The intervals leave an additional combination penalty unresolved in this evidence-supplied setting; they do not establish equivalence. The original group gap alone therefore cannot attribute lower scores to combining facts: the requested content and information conditions must also be considered. The small human declines motivate examining which errors the automatic evaluator accepts.

6.3 High agreement can mask local fact errors

Human checks distinguish robust claim-verification gains from small fact-level effects. The explicit decision parser agrees with human decisions on 240/240 Q8 responses. Explanation agreement is 97.4% on 227 determinate responses, joint agreement is 99.1% on 232, and biological-fact agreement is 97.7% on 262/264 determinate atom judgments. Denominators retain the distinction between determinate and uncertain labels. Of eight biological atoms humans judge incorrect, the automatic judge accepts six. This small error subset shows why agreement should be checked against the inference of interest: human scoring supports the large Q8 benefits while leaving the smaller fact-level effect unresolved.

Appendix E reports pre-adjudication reviewer agreement, confusion matrices, model-specific errors, and uncertain labels. The new audit supports these binary and fact-level comparisons, not calibration of the historical five-band means.

7 Discussion and Limitations

Choosing outcomes for the intended comparison. Base-v1 preserves the eight-slot baseline for broad comparisons (Appendix H). Balanced-Q8 separates decisions from supported explanations; matched-fact accuracy measures retention of the same required facts across question formats. Human review checks these specific outcomes and reveals errors that aggregate agreement conceals. The outcomes retain separate metrics and evidence scopes.

Scope and uncertainty. The release favors common, documented vertebrates and lower-ambiguity images, limiting transfer to rare taxa, crowded scenes, and field deployment. Filtering does not verify species labels; silver captions and text-only judging can miss visual errors. The zero-shot evaluation establishes neither supervised generalization nor absence of pretraining overlap. Names and evidence can influence several internal processes. Targeted studies cover three models and 100 species; human review covers 20 species with a five-species overlap. Bootstrap intervals omit generation randomness, systematic judge error, and multiple-comparison correction. Near-ceiling scores and few human-identified errors limit small-effect conclusions.

8 Conclusion

WildBioWiki links wildlife images, fixed species-level questions, reference answers, and evidence for reusable comparisons under controlled inputs. Twelve-model baselines and targeted studies demonstrate its uses; human checks support information effects while leaving small fact-combination differences unresolved. Researchers can reuse the aligned records to compare input conditions and inspect reference evidence, with explicit limits on interpreting performance.

AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval and discovery, language editing, translation, and writing refinement. AI was also used to generate synthetic data and supporting annotations, including structured species knowledge, question and answer references, and silver captions. The authors subsequently reviewed and manually revised all AI-assisted text and verified relevant information against the underlying sources and evidence. Construction procedures, quality checks, and human reference audits for the AI-assisted synthetic data are described in Section 3 and Appendix A. The authors made the final decisions and take responsibility for the manuscript, claims, and artifacts, including all AI-assisted content.

ETHICS STATEMENT

WildBioWiki uses existing wildlife images and species knowledge to support research on model evaluation. Images retain their source licenses and attribution requirements; the code uses Apache License 2.0 and derived text CC BY 4.0. The concentration on commonly observed, well-documented vertebrates can underrepresent other taxa and ecological settings. Users should preserve this sampling context when interpreting scores or extending the resource. Species labels, generated references, and automated assessments can contain errors; benchmark outputs should not be treated as authoritative identification or conservation advice. Human annotation is reported as evaluation of references and model responses, with reviewer-identifying details omitted for anonymous review.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe construction and partitions. Appendix **??** maps the dataset files, identifier joins, code entry points, and reuse steps. The companion resource archive separates the 392-species QA/knowledge index, base construction and evaluation code, and targeted diagnostic replay. Section 4 and Appendices B–E specify evaluation and uncertainty. Appendix F describes the 2,100 targeted response/score records, human labels, and offline replay; Appendix H distinguishes scoring versions and Q7 sensitivity. The anonymous archive includes a 20-image demonstration subset and supports target inspection and diagnostic replay. The full image/caption dataset will be made publicly available upon acceptance, with source-specific licenses and attribution retained; Appendix **??** specifies the review materials and release scope.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset, 2021.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13638–13646, 2023.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14948–14968, 2023.

Changin Choi, Wonseok Lee, Jungmin Ko, and Wonjong Rhee. Multimodal iterative rag for knowledge-intensive visual question answering, 2025.

Hongyong Han, Wei Wang, Gaowei Zhang, Mingjie Li, and Yi Wang. Coralvqa: A large-scale visual question answering dataset for coral reef image understanding, 2025.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12065–12075, 2023.

Yevheniia Kryklyvets, Mohammed Irfan Kurpath, Sahal Shaji Mullappilly, Jinxing Zhou, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, and Hisham Cholakkal. Mavis: A multimodal conversational assistant for avian species. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28601–28627, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Georgiana Manolache, Gerard Schouten, and Joaquin Vanschoren. Crypticbio: A large multimodal dataset for visually confusing biodiversity, 2025.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

M Maruf, Arka Daw, Kazi S Mehrab, Harish B Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P Balhoff, Yasin Bakış, Bahadir Altintas, et al. Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. *Advances in Neural Information Processing Systems*, 37:131035–131071, 2024.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3113–3124, 2023.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19023–19034, 2022.

Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A Crook, Yifan Ethan Xu, Babak Damavandi, Xin Luna Dong, Christos Faloutsos, Lei Li, and Seungwhan Moon. Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 247–266, 2024.

Syed Nazmus Sakib, Nafiul Haque, Mohammad Zabed Hossain, and Shifat E Arman. Plantvillagevqa: A visual question answering dataset for benchmarking vision-language models in plant science, 2025.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Yunxiao Shi, Xing Zi, Zijing Shi, Haimin Zhang, Qiang Wu, and Min Xu. Enhancing retrieval and managing retrieval: A four-module synergy for improved quality and efficiency in rag systems. In *Proceedings of the 27th European Conference on Artificial Intelligence*, pp. 2258–2265. IOS Press, 2024.

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.

Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate E Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. *Advances in Neural Information Processing Systems*, 37:126500–126514, 2024.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40 (10):2413–2427, 2017.

Yuk-Kwan Wong, Haixin Liang, Zeyu Ma, Yiwei Chen, Ziqiang Zheng, Rinaldi Gotama, Pascal Sebastian, Lauren D Sparks, and Sai-Kit Yeung. Orca: Object recognition and comprehension for archiving marine species. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1597–1609, 2026.

Jinghao Xiao, Yiheng Guo, Xing Zi, Karthick Thiyagarajan, Catarina Moreira, and Mukesh Prasad. Beyond pixels: A training-free, text-to-text framework for remote sensing image retrieval. In *2025 18th International Conference on Sensing Technology*. IEEE, 2025. doi: 10.1109/ICST66402.2025.11512424.

Runjie Zhou, Youbo Shao, Haoyu Lu, Bowei Xing, Tongtong Bai, Yujie Chen, Jie Zhao, Lin Sui, Haotian Yao, Zijia Zhao, et al. Worldvqa: Measuring atomic world knowledge in multimodal large language models, 2026.

Xing Zi, Jinghao Xiao, Yunxiao Shi, Xian Tao, Jun Li, Ali Braytee, and Mukesh Prasad. Rsvlm-qa: A benchmark dataset for remote sensing vision language model-based question answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12905–12911. Association for Computing Machinery, 2025.

A Construction Details and Resource Boundaries

Figure 2 illustrates the two tasks. QA references use species knowledge rather than instance-specific visual facts. Questions omit common and scientific names, avoid name leakage and obvious image-only traits, and store supporting text with field tags. The fixed order in Table 2 supports aligned comparisons across species. Archived construction review covered all 396 candidate records: 253 were kept, 84 received minor fixes, and 59 received major fixes. Repairs addressed unsupported answers, evidence mismatch, leakage, overlap, and low-value extras. Structural validation found no remaining schema, question-order, or evidence-field errors over 3,168 candidate questions. Structural validity does not establish semantic difficulty; we audit the operations required by Q7 separately in Section 6.

Visual filtering maps scientific names to broader prompts using common names, taxonomic class, and a hand-curated head-phrase lexicon. For example, a turtle can be queried as "turtle. reptile." Grounding DINO uses box and text thresholds of 0.30 and 0.25. Images with no detected boxes or more than two boxes are discarded; one-box images are retained, and two-box images retain the higher-confidence box. No additional NMS or IoU-based merging is introduced beyond detector post-processing. The retained box is linked to species metadata; filtering does not independently verify fine-grained identity or guarantee that every image depicts one organism.

The five vertebrate classes are Actinopterygii, Amphibia, Aves, Mammalia, and Reptilia. Balancing the initial candidate pool at 200 species per class does not balance the final image counts: observation volume, knowledge completeness, and image yield differ. Table 8 reports changes in image shares caused by visual filtering. The largest absolute change is 2.57 percentage points for mammals. These aggregate proportions do not establish that filtering is unbiased within each class, for example with respect to occlusion, habitat, photographic conditions, or organism size.

Table 8: Vertebrate-class image shares before and after filtering. Differences are percentage points.

Class	Before (%)	After (%)	Change
Actinopterygii	20.71	19.00	-1.71
Amphibia	13.38	15.60	$+2.22$
Aves	16.16	16.30	$+0.14$
Mammalia	26.77	24.20	-2.57
Reptilia	22.98	24.90	$+1.92$

Four of the 396 knowledge-qualified species fall below the 300-image retention threshold: *Mirounga angustirostris*, *Zalophus californianus*, *Eumetopias jubatus*, and *Chromis viridis*. The intermediate knowledge and QA records therefore have 396 rows and 3,168 questions, whereas image-conditioned evaluation uses the 392-species intersection and its 3,136 questions. These counts refer to distinct stages of construction, not conflicting definitions of the release.

Caption generation requires natural inclusion of the scientific name, excludes ecological speculation, and discourages camera commentary. Exact-name validation covers all 264,563 captions. A camera-wording screen flagged 7,881 captions (3.0%) for targeted rewriting, which completed without recorded processing errors. These checks validate specified rules, not every aspect of visual fidelity.

The original pipeline and example figures are retained in the main text; the captions distinguish schematic content from the implemented sources and image-grounded reference requirements. Knowledge is derived from Wikipedia and Wikidata-linked metadata; the method does not claim additional ingestion from databases that are not documented in these construction records. The fixed questions are species-level: a habitat answer need not describe the exact location or background visible in the photograph. In contrast, caption references should describe visible content. This distinction prevents ecological knowledge from being mistaken for direct visual evidence.

B Models and Reproduction Settings

Table 9: Full model identifiers corresponding to the original broad-evaluation row labels. These identify the evaluated systems; the new diagnostics use only the three models in Table 10.

Main-text label	Full identifier
InternVL3.5	InternVL3.5-8B-HF
Ovis2.5	Ovis2.5-9B
Qwen3-VL	Qwen3-VL-8B-Instruct
SAIL-VL	SAIL-VL-1d6-8B
Gemma-3	Gemma-3-4B-IT
Phi-4-MM	Phi-4-multimodal-instruct
LLaVA-OV1.5	LLaVA-OneVision-1.5-8B-Instruct
MiniCPM-V4.5	MiniCPM-V-4.5
Idefics3	Idefics3-8B-Llama3
Molmo-7B-D	Molmo-7B-D-0924
GPT-5.4-nano	GPT-5.4-nano
Grok-4.1-fast	Grok-4.1-fast-non-reasoning

Table 10: Pinned configurations for the new Q8 and matched-fact experiments. All use eager attention, no sampling, and a 512-token generation cap. Full revisions and checkpoint fingerprints are retained in the evidence companion.

Model repository	Revision prefix	Precision
Qwen/Qwen3-VL-8B-Instruct	`0c351dd01ed8`	float16
BytedanceDouyinContent/SAIL-VL-1d6-8B	`6b0c85f357cf`	bfloat16
google/gemma-3-4b-it	`093f9f388b31`	bfloat16

The targeted system prompt is: "You are a wildlife visual question answering assistant. Answer the question concisely and directly." Q8 requests a leading Yes, No, or Uncertain decision and a brief explanation. Knowledge-oracle requests include the correct scientific name and the retained knowledge record. The separate habitat, separate diet, and combined questions use identical species, image, and knowledge within each matched comparison. Input and response hashes permit joins by stable identifiers rather than file order. No new model training is part of these experiments.

C Original Evaluation: Additional Descriptive Results

Table 11: Original cross-taxon means over twelve models under the frozen LLM-J rubric. Columns are fixed QID groups, not certified cognitive categories. Values are scores, not percent accuracies.

Taxon	Q2–Q4	Q5–Q6	Q7–Q8
Ray-finned fish	37.6	29.2	41.8
Amphibians	35.4	28.9	43.2
Birds	43.0	38.0	48.8
Mammals	41.2	36.7	46.3
Reptiles	34.9	30.6	41.1

Across the original 15 taxon–question-group contexts, the twelve-model rankings have Spearman correlations of 0.685–0.958 with the overall ranking. This describes context-specific ordering; it is not an independent validation of the metric. On the 79-species disjoint test partition, reaggregating frozen zero-shot predictions places Qwen3-VL highest at 47.7 LLM-J [45.5, 49.9] and 58.7%



Acanthurus tractus
Species metadata

Retained evidence: "It is diurnal and relies on grazing on algal films."

True claim: The animal is diurnal.
False claim: The animal is nocturnal.

Requested output: Yes, No, or Uncertain, with a brief explanation.

Separate outcomes: The leading label determines the decision; the reason is checked against the evidence. A bare "No" can reject the false claim but earns no explanation credit.

Figure 5: An actual balanced-Q8 item pair. Closed-book input contains the image and claim, while KB-oracle additionally supplies the species name and retained record. The metadata and evidence shown here explain construction; they are not supplied in the closed-book condition.

Acc@65 [55.3, 62.1]. The models have not been trained on a WildBioWiki training partition, so these results cannot establish supervised transfer to unseen species.

Thresholds 60 and 65 yield identical original archived rankings because the five-band rubric has no intermediate anchor between 35 and 65. Their agreement is therefore not a separate robustness test. The stricter threshold 70 retains positive rank correlation with 65 ($\rho = 0.825$) in the archived analysis. Five-band averages, thresholded correctness, the new Q8 joint outcome, and new biological-fact accuracy have different meanings and must not be pooled into a single score.

Table 12: Original wrong-name diagnostic on 20 species and 140 Q2–Q8 items. Positive differences indicate the benefit of the correct name over an incorrect one; units are LLM-J points.

Model	Correct minus wrong name	Paired 95% CI
Qwen3-VL	23.5	[17.9, 29.7]
SAIL-VL	19.0	[13.2, 24.6]
Gemma-3	8.8	[2.3, 14.8]

D New Scoring Rules and Analysis Corrections

Decision versus explanation. The Q8 decision is the explicit leading choice made by the model. A later explanation cannot silently change that choice. The raw API judge sometimes substituted the evidence-supported truth value for the model's expressed choice: for example, an answer beginning "Yes" could receive an extracted "No" when its explanation contradicted the reference. This occurred in 120 of 1,200 responses. Primary decision statistics therefore use the strict leading-label parser that was available before API scoring. All outputs are parseable, and Uncertain receives no correctness credit. Raw API decisions and the disagreement list are retained for reproducibility, while explanation judgments remain separate and unchanged.

The judge uses GPT-5.4-mini-2026-03-17, low reasoning effort, and a strict JSON output schema. It sees the answer and reference evidence without the responding model's identity. The explanation categories are valid, invalid, absent, and uncertain. A bare Yes/No/Uncertain is absent; restating a rejection without the relevant supported reason also earns no explanation credit. During execution, a schema constraint was applied to 30 bare responses to enforce this existing rule after a validation failure. This was an execution correction, not a change to the underlying question pairs, biological facts, or generated responses. Original outputs and analysis versions are preserved.

Fact accuracy and complete answers. Each required fact has an explicit proposition, supporting record field, acceptable paraphrases, optional details, and contradiction criteria. Unlisted content is not automatically false. A known conflict overrides an otherwise matching phrase, and an answer that omits a required proposition is distinct from one that contradicts it. Only supported biological facts contribute to fact accuracy. Questions whose records lack a required fact contribute separate insufficiency diagnostics rather than being treated as biological errors by construction.

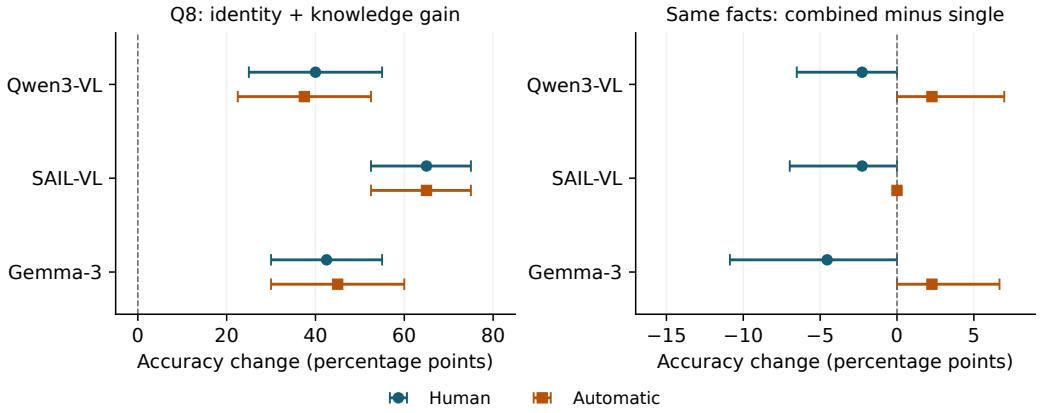


Figure 6: Human and automatic contrasts on the *same* 20-species sample. Left: change in balanced-Q8 joint correctness when identity and knowledge are supplied. Right: change in biological-fact accuracy when identical facts are requested together rather than separately. Bars are paired, class-stratified 95% species-bootstrap intervals. Human evidence supports the input gain; it does not establish a stable additional combination penalty.

For condition $c \in \{\text{separate}, \text{combined}\}$ and supported fact set $\mathcal{A}_s$, the primary micro-averaged outcome is

$$F_{mc} = \frac{\sum_s \sum_{a \in \mathcal{A}_s} \mathbf{1}[\ell_{msac} = \text{correct}]}{\sum_s |\mathcal{A}_s|}. \tag{3}$$

Species bootstrap recomputes this ratio from sampled counts while preserving all atoms and conditions of each sampled species. Separate accuracy evaluates each fact only in its corresponding single answer; the other single answer cannot rescue an omission. Complete-answer accuracy requires all requested facts to be correct and no incompatible extra assertion. The separate condition requires both single answers to satisfy that rule. This avoids an artificial disadvantage from comparing a joint requirement with only one constituent question.

E Human Evidence and Remaining Uncertainty

Human intervals use 2,000 paired, class-stratified species bootstrap samples (seed 20260908), preserving four species per class and all within-species pairings. Full 100-species automatic intervals use paired species resampling without this human-sample stratification. Figure 6 expands the paired effects summarized in main-text Table 7.

The final human sample contains 20 species, selected as four per class, and 420 responses. The second-reviewer overlap covers five species and 105 responses, with one species per class. Reviewers were not shown item-level automatic scores. The recorded provenance does not separately establish masking to one another's original labels; we therefore report the observed pre-adjudication agreement without asserting stronger blinding. Six label disagreements were resolved by joint adjudication. One adjudicated explanation legitimately remains uncertain, and uncertainty elsewhere in the main sample is preserved rather than forced into a binary label.

Table 13: Final human versus automatic binary outcomes. Correctness is the positive class: FP means automatic acceptance of a human-incorrect item. Uncertain human labels are excluded only from these matrices, not silently dropped from fixed-denominator confirmed-correct rates.

Outcome	Determinate / total	TP	TN	FP	FN	Agree (%)
Explicit Q8 decision	240 / 240	195	45	0	0	100.0
Q8 explanation	227 / 240	171	50	1	5	97.4
Q8 joint correctness	232 / 240	169	61	0	2	99.1
Biological fact	262 / 264	254	2	6	0	97.7

The explicit decision row validates extraction of the expressed choice, not the LLM judge's faulty decision field. For the latter diagnostic, agreement is 91.7% on 240 human-audited responses, with 20 false positives. Joint correctness can be determinate even when the explanation is uncertain if the expressed decision is already wrong, which explains why its determinate denominator exceeds that of explanation alone. The six record-insufficiency judgments are separate from the 264 biological facts and have complete human–automatic agreement in this small sample.

Table 14: Model-specific agreement with final human labels on determinate items. These group sizes support descriptive checks, not universal estimates of model-specific judge bias.

Model	Explanation agree (n)	FP / FN	Fact agree (n)	FP / FN
Qwen3-VL	98.7% (77)	0 / 1	97.7% (88)	2 / 0
SAIL-VL	100.0% (74)	0 / 0	98.8% (86)	1 / 0
Gemma-3	93.4% (76)	1 / 4	96.6% (88)	3 / 0

Table 15: Human complete-answer comparisons on the same 19 eligible species. Separate requires both single answers to be complete. Differences are combined minus separate, in percentage points.

Model	Separate (%)	Combined (%)	Difference [95% CI]
Qwen3-VL	94.7	89.5	-5.3 [-15.8, 0.0]
SAIL-VL	89.5	84.2	-5.3 [-15.8, 0.0]
Gemma-3	94.7	84.2	-10.5 [-25.0, 0.0]

There remain 13 uncertain Q8 explanations and two uncertain biological-fact judgments. For fixed-denominator joint accuracy, assigning all unresolved outcomes adversarially gives human KB-minus-closed gain ranges of [35.0, 40.0], [57.5, 65.0], and [37.5, 45.0] percentage points for Qwen3-VL, SAIL-VL, and Gemma-3. These are sensitivity bounds on point estimates, not confidence intervals; the input benefit remains positive throughout. For matched facts, uncertainty and sample size leave the interpretation of small differences unresolved. Intervals containing zero are not evidence of equivalence.

F Evidence Companion and Offline Replay

The portable companion retains 100 true/false pairs, 100 matched fact-rubric records, pinned model configurations, all 2,100 new input–response–score records, 240 joined human Q8 rows, and 270 joined human atom rows. The atom rows comprise 264 biological-fact and six record-insufficiency judgments. Reviewer identities and free-text internal notes are omitted from the portable labels. Stable species, pair, atom, and job identifiers connect the records; joins should not depend on row order.

Running `python replay_companion.py` with Python 3 and its standard library recomputes 15 paired contrasts, 45 human/automatic/difference-of-differences intervals, and 24 overall or model-specific confusion matrices. The replay preserves seed 20260908, species ordering, within-class sampling, and the archived quantile-index rule, and asserts exact equality with the final adjudicated statistics. It requires no API key, GPU, model weights, or image download. It reproduces the new human comparisons, not the original broad-validation scores. The included worked example records the exact input, response, judge request, and score; it is an automatic example, not an additional human-audited response.

The evidence-supplied experiments are bounded diagnostic settings. They leave practical retrieval, richer ecological relations, difficult multi-organism scenes, and supervised species-disjoint learning for future work. These are possible uses of the resource rather than unperformed experiments claimed by this paper.

G Extended Related Work

Knowledge-intensive visual evaluation. OK-VQA and A-OKVQA require knowledge beyond direct image description (Marino et al., 2019; Schwenk et al., 2022); FVQA connects questions to supporting facts (Wang et al., 2017). InfoSeek and Encyclopedic-VQA emphasize information seeking and detailed properties of fine-grained categories (Chen et al., 2023; Mensink et al., 2023). OVEN and WorldVQA address visual entity recognition and atomic world knowledge (Hu et al., 2023; Zhou et al., 2026). WildBioWiki shares their concern with entity-linked knowledge, but fixes an ecological question schema across vertebrate species and repeats each species-level question over multiple images. Its input controls ask how access to identity and evidence changes answers under this shared structure.

Natural-world and biological resources. Terra Incognita, iWildCam, iNat2021, and Animal Kingdom support recognition, distribution-shift, and behavior studies (Beery et al., 2018; 2021; Van Horn et al., 2021; Ng et al., 2022). INQUIRE evaluates natural-world image retrieval (Vendrow et al., 2024). VLM4Bio, ORCA, CrypticBio, and MAviS extend evaluation toward biological traits, marine species comprehension, visually confusing taxa, and avian language interaction (Maruf et al., 2024; Wong et al., 2026; Manolache et al., 2025; Kryklyvets et al., 2025). CoralVQA and PlantVillageVQA provide domain-specific QA settings (Han et al., 2025; Sakib et al., 2025). Our distinguishing unit is a species record with fixed questions, retained ecological evidence, and multiple images, coupled to controlled input changes. This complements task-specific biological benchmarks; dataset size alone is not our novelty claim.

Generation, retrieval, and evaluation. Novel-object captioning extends recognition toward description (Agrawal et al., 2019), while SnapNTell and MI-RAG study retrieval-augmented knowledge-intensive VQA (Qiu et al., 2024; Choi et al., 2025). Retrieval management also examines query reformulation, filtering, memory, and triggering (Shi et al., 2024). Related remote-sensing work studies VQA and retrieval from textual descriptions (Zi et al., 2025; Xiao et al., 2025). WildBioWiki's oracle condition supplies retained evidence directly and therefore measures neither retrieval quality nor a deployable RAG system. Rubric-based semantic evaluation is informed by caption evaluator work such as CLAIR (Chan et al., 2023). We complement reference-based scores with human comparisons of specific input effects and explicitly account for the limits of text-only judging.

Interpreting task difficulty. Q7's operational audit and the matched-fact study constrain a reasoning-based interpretation of low group scores. Evidence-supplied fact recovery is not arbitrary multi-step reasoning, and human support for balanced Q8 does not certify other rankings or tasks. Future comparisons can use these explicit outcome definitions to distinguish missing knowledge, incomplete reporting, and disagreement with reference evidence.

H Scoring Versions and Q7 Sensitivity

The published comparisons distinguish a frozen baseline from targeted diagnostic modules. Table 16 fixes the intended use of each. The default for reproducing Table 4 is Base-v1, with all eight original question slots and the frozen rubric. It remains an archival baseline, including the disclosed Q7 limitations. For new verification or fact-coverage claims, the recommended diagnostic outcomes are Balanced-Q8 joint correctness and Matched-facts accuracy, reported separately. They are not averaged with Base-v1 into a composite score. A Q7-filtered result must carry its own label and retained-cell count; it must not silently replace an unfiltered baseline.

Filtered aggregation. Let $w_{sj} = 0$ only for a Q7 item classified as invalid or ambiguous, and $w_{sj} = 1$ for other eligible cells. The filtered statistic is

$$R_m^{\text{filter}} = \frac{\sum_s \sum_j w_{sj} \overline{r}_{msj}}{\sum_s \sum_j w_{sj}}. \tag{4}$$

The same rule applies to cell-level Acc@65. This preserves equal weight per retained species–question cell; a species with seven eligible questions contributes seven cells rather than being renormalized to eight. Species-cluster bootstrap resamples both cell sums and their retained counts. Only

Table 16: Protocol names used in this manuscript and their scoring scope.

Protocol	Unit / metric	Intended use
Base-v1	Eight slots; LLM-J, Acc@65	Reproduce frozen broad results and original matched controls
Q7-filtered	Eligible cells; same rubric	Sensitivity to invalid/ambiguous Q7; 3,117 broad cells or 697 base-control cells
Balanced-Q8	Paired claims; D, E, J	Test claim verification and joint identity/evidence assistance
Matched-facts	Required facts; accuracy	Compare separate and combined answers on the same facts

Q7 is filtered; no species or its other questions is removed. The full audit flags 19 of 3,136 cells, and the 100-species base-control subset contains three of those flagged cells.

Sensitivity evidence. The following comparison uses the archived prediction subset and its fixed scoring configuration. All before/after differences are calculated within that same archive; they do not replace full-validation Table 4. The archived filter was recorded under `exclude_AI_ambiguous_Q7`; the author subsequently confirmed the audit's 330/43/19 operation counts. Table 17 reports the resulting changes. Both LLM-J and Acc@65 point-estimate orderings are unchanged across twelve models. This is a sensitivity result, not proof that close ranks are statistically distinguishable or that every retained item is semantically valid.

Table 17: Within-archive sensitivity to excluding 19 Q7 cells. LLM-J values are rubric points; ΔAcc@65 is in percentage points. Deltas use unrounded estimates.

Model	Base LLM-J	Filtered LLM-J	ΔLLM-J	ΔAcc@65
InternVL3.5	32.406	32.393	-0.012	-0.022
Ovis2.5	41.723	41.752	+0.029	+0.022
Qwen3-VL	45.481	45.538	+0.056	+0.071
SAIL-VL	42.168	42.171	+0.002	-0.009
Gemma-3	34.802	34.872	+0.071	+0.109
Phi-4-MM	31.691	31.701	+0.009	+0.000
LLaVA-OV1.5	37.907	37.888	-0.020	-0.032
MiniCPM-V4.5	41.751	41.794	+0.043	+0.058
Idefics3	28.169	28.228	+0.059	+0.079
Molmo-7B-D	26.984	26.974	-0.010	-0.018
GPT-5.4-nano	32.048	32.040	-0.008	-0.011
Grok-4.1-fast	37.535	37.525	-0.010	-0.015

On the matched base-control archive, the same filter retains 697 of 700 questions. Name-given minus closed-book LLM-J remains 10.22 [7.77, 12.70], 5.72 [2.95, 8.41], and 5.36 [2.68, 8.09] for Qwen3-VL, SAIL-VL, and Gemma-3. KB-oracle minus name-given remains 31.25 [28.43, 34.21], 35.83 [32.76, 38.76], and 36.80 [33.74, 40.24]. These paired species-bootstrap intervals all exclude zero. The companion retains the archived summaries and a deterministic comparison script; filtering and summary comparison introduce no new model generations.