# BIOINTERACT: A LARGE-SCALE MULTIMODAL DATASET FOR EVALUATING FINE-GRAINED SEMANTIC UNDERSTANDING OF BIOTIC INTERACTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

While recent advances in vision-language models (VLMs) have spurred the development of domain-specific datasets and benchmarks, these evaluations often fail to assess fine-grained semantic understanding, allowing models to achieve high scores without robust visual grounding. We address this evaluation gap through the lens of biotic interactions: directional, asymmetric relationships between organisms (e.g., wasp parasitizes caterpillar vs. caterpillar parasitizes wasp). This relational complexity yields naturally adversarial instances that expose superficial reasoning in current VLMs. To this end, we introduce BIOINTERACT, the largest multimodal dataset for evaluating VLM robustness on real-world biodiversity challenges. Curated from iNaturalist and validated against established interaction literature, the dataset contains 15.4K unique interactions spanning 6.5K taxa across 256K images. Each interaction is structured as a source-relation-target triplet, enabling controlled semantic perturbations. We further introduce BIOINTERACT100, an image retrieval benchmark revealing that state-of-the-art VLMs suffer from severe consistency gaps and are highly brittle to relation-direction reversals. BIOINTERACT provides a faithful evaluation of multimodal AI, while encouraging the development of robust systems for ecological research.

## 1 INTRODUCTION

Recent advances in artificial intelligence (AI) technologies have led to the widespread adoption of foundation models (FMs) across diverse domains, enabling scalable and increasingly automated analysis of complex data (Liu et al., 2026a). In particular, multimodal FMs trained on heterogeneous data (e.g., text, images) exhibit emergent capabilities that extend beyond their original training objectives, opening new opportunities for data-driven discovery. This progress has led to a growing class of domain-specific datasets and benchmarks aimed at evaluating such capabilities in realistic settings, including health, medicine, and chemistry (Chen et al., 2024; Zhou et al., 2024; Liu et al., 2026b; Yue et al., 2025). However, most domain-specific benchmarks evaluate task performance, typically using a single canonical query formulation, and do not test fine-grained semantic understanding. As a result, models may achieve high benchmark performance without robustly grounding subtle semantic distinctions in visual content, providing an incomplete assessment of their multimodal capabilities. Addressing this evaluation gap is essential for ensuring reliable deployments of such systems in specialized fields (Gu et al., 2025).

In ecology, advances in FMs and computer vision have accelerated biodiversity research, with specialized benchmarks spanning tasks from basic visual interpretation (e.g., taxonomic classification, attribute prediction, and trait inference) to higher-level ecological understanding (e.g., disease detection) (Xian et al., 2018; Khan et al., 2023; Singh et al., 2020; Manolache et al., 2026). However, understanding biodiversity and coping with the biodiversity crisis in human-modified ecosystems, requires knowledge about biotic interactions between multiple organisms. Biotic interactions are highly opportunistic in nature and, therefore, difficult to study at scale; direct observations are labor-intensive, and often restricted to specific habitats or timeframes, limiting their generalizability. Recent studies highlight the potential of community science data as a scalable resource for analyzing interactions. Platforms such as iNaturalist and Observation.org now host hundreds of millions of geo-referenced observations paired with images. Although these datasets are primarily used to document species



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: **Example images depicting biotic interactions from BIOINTERACT.** Biotic interactions are annotated triplets—source taxon, interaction type, target taxon (e.g., *Megachile centuncularis* eats *Centaurea nigra*), standardized to nine types defined by the Open Biological and Biomedical Ontology Foundry (OBO) Relation Ontology (RO) (see Table 2). Images originate from iNaturalist *research-graded* observations, where the source taxon occurrence is indexed (i.e., taxonomic identity, location, time), and the interaction type and target taxon are additional annotations.

occurrences, the images are captured *in situ* and frequently contain substantially richer ecological information as a by-product (Pernat et al., 2024a), including morphological traits, behavior, habitat, and interactions. For instance, pollinator images may reveal host flower preferences (Pernat et al., 2024b), while images of predatory species can provide evidence of trophic interactions (Hu et al., 2025). Despite the rapid growth of such repositories, extracting richer ecological descriptors at scale remains a significant challenge. Fewer than 25% of over 300 million observations on iNaturalist include detailed annotations (Vendrow et al., 2026). Although indispensable, manual labeling is both time-consuming and dependent on expert knowledge, rendering it impractical at this scale and limiting the ability of researchers to fully exploit these data. As ecosystem degradation accelerates globally (Abdala-Roberts et al., 2025), there is an urgent need for methods capable of leveraging these vast resources to automatically extract latent ecological information from existing observations.

Vision-language models (VLMs) offer a promising approach for zero-shot multimodal understanding of ecologically relevant images through tasks such as visual question answering (VQA) and image retrieval (Vendrow et al., 2024; Maruf et al., 2024; Wang et al., 2026; Gauba et al., 2026). By design, these tasks rely on unconstrained, free-form natural language inputs to enable flexible, user-friendly image queries. However, biotic interactions are particularly challenging: their meaning is inherently relational and directional, encoding asymmetric effects between organisms where small semantic changes can completely alter the meaning, yielding naturally adversarial interaction instances (e.g., *wasp parasitizes caterpillar* and *caterpillar parasitized by wasp* are semantically equivalent, whereas *wasp parasitizes caterpillar* and *caterpillar parasitizes wasp* are syntactically equivalent, yet represent distinct interactions). Current benchmarks remain limited in scale and scope (see Table 1), typically fine-tuning and evaluating answer accuracy using a single canonical query formulation per image in VQA such as CoralVQA (Wang et al., 2026) and AgMMU (Gauba et al., 2026), or in text-to-image retrieval such as INQUIRE (Vendrow et al., 2024). While these benchmarks assess whether a model produces the correct answer on fixed image-query pairs, they fail to capture the complexities of open-world settings, leaving a substantial gap in understanding how effectively VLMs operate in realistic scenarios. A system capable of fine-grained semantic understanding in an open-world setting would unlock meaningful progress in AI-driven biodiversity research.

In this work, we challenge the biodiversity AI research by curating and releasing BIOINTERACT, a large-scale multimodal dataset containing 15.4K unique biotic interactions spanning 6.5K taxa across 256K images, enabling systematic generation of controlled adversarial interaction formulations to

Table 1: **Existing biotic interactions benchmarks.** These benchmarks provide limited insight into a model's ability to reason about biotic interactions under input perturbation—conditions that commonly arise in open-world settings. BIOINTERACT (ours) scales both in size and annotation richness (see Table 3).

| Tasks | Benchmark | # Images | Query | Description |
|---|---|---|---|---|
| *visual question answering (VQA)* | CoralVQA (Wang et al., 2026) | 12K | single | ecological and health-related conditions questions for coral reef analysis, in open-ended and closed-ended formats, including symbiotic relationship between corals and algae in underwater images |
| | AgMMU (Gauba et al., 2026) | 50.2K | single | agriculture knowledge multiple-choice and open-ended questions, including insect/pest identification |
| *image retrieval* | INQUIRE (Vendrow et al., 2024) | 33K | single | 200 unique queries extracted from expert interviews and academic literature, including 12 parasitism and symbiosis relationships |
| *multi-task* | BIOINTERACT (ours) | 256K | multiple | 15.4K unique biotic interactions across five kingdoms (*Animalia*, *Plantae*, *Fungi*, *Chromista*, and *incertae sedis*) denoted with both scientific and vernacular names, standardized to nine interaction types, grounded in spatiotemporal metadata |

probe fine-grained semantic understanding. Specifically, BIOINTERACT captures who interacts with whom, how, and under what context by representing interactions as structured triplets—source taxon, interaction type, and target taxon. Examples from BIOINTERACT are shown in Figure 1, while Table 1 summarizes how BIOINTERACT contrasts in scale and annotation richness. By focusing on real-world biodiversity challenges, BIOINTERACT provides a faithful evaluation of multimodal AI progress, encouraging the development of robust systems that can assist with accelerating ecological research. Our main contributions include: (1) the largest multimodal dataset of biotic interactions to date enabling systematic adversarial interaction formulations for fine-grained semantic understanding; (2) BIOINTERACT100, a novel image retrieval task on a fixed set, comprising 281 unique interactions (i.e., unique triplets—source taxon, interaction type, and target taxon) with 100 images each; (3) controlled adversarial benchmark revealing that state-of-the-art VLMs suffer from severe compositional shortcomings, failing drastically under relation-direction reversals despite high baseline accuracy.

## 2 RELATED WORK

VLMs have emerged as a powerful paradigm for learning joint representations across visual and textual modalities. Early contrastive approaches such as CLIP (Contrastive Language-Image Pre-Training) (Radford et al., 2021) learn a shared embedding space where data from two modalities can be encoded jointly. Building on this foundation, VLMs, including GPT-4o (Hurst et al., 2024), LLaVA (Li et al., 2024), and Qwen2-VL (Wang et al., 2024), extend these capabilities by connecting the outputs of visual encoders directly into language models. Recent expert-level benchmarks further challenge multimodal models in scenarios where expert-level knowledge is required (Chen et al., 2024; Zhou et al., 2024; Yue et al., 2025). For instance, GMAI-MMBench (Chen et al., 2024) and BenchX (Zhou et al., 2024) feature a collection of healthcare and medically relevant questions, while MMMU-Pro (Yue et al., 2025), comprises visual-questions related to problems from diverse fields such as business, arts, and science. However, the evaluation landscape remains fragmented: domain-specific benchmarks primarily assess whether models can successfully perform a given task, whereas fine-grained semantic understanding is predominantly evaluated on general-domain datasets depicting common objects and everyday scenes (Yuksekgonul et al., 2023; Thrush et al., 2022; Alhamoud et al., 2025; Xu et al., 2024). Consequently, continued performance gains on domain-specific benchmarks may increasingly reflect memorization rather than genuine capability (Cheng et al., 2026). Recent work therefore emphasizes the need for evaluation protocols that distinguish benchmark performance from genuine model capabilities and assess whether these capabilities generalize beyond the conditions under which they are evaluated (Gu et al., 2025).

BIOINTERACT is different from existing expert-level benchmarks for the following reasons. **(1) Focus on biotic interactions**: While prior work has addressed a range of visual tasks (Xian et al., 2018; Singh et al., 2020; Khan et al., 2023; Maruf et al., 2024; Manolache et al., 2026), these efforts largely focus on individual organisms or ecological context understanding. In contrast, BIOINTERACT targets

Table 2: **Interaction types from BIOINTERACT, defined by Open Biological and Biomedical Ontology Foundry (OBO).** Interactions follow the hierarchy of biological process types. Each process type is mapped to a corresponding interaction relation, and finer-grained relations inherit from broader ones through a subrelation structure (see Appendix A).

| Type | Definition |
|---|---|
| eats | A biotic interaction where one organism consumes a material entity through a type of mouth or other oral opening. |
| has host | X 'has host' y if and only if: x is an organism, y is an organism, and x can live on the surface of or within the body of y. |
| interacts with | An interaction relationship in which at least one of the partners is an organism and the other is either an organism or an abiotic entity with which the organism interacts. |
| parasite of | A parasite-host relationship where an organism benefits at the expense of another. |
| parasitoid of | A parasite that kills or sterilizes its host. |
| pathogen of | Inverse of has pathogen. |
| preys on | An interaction relationship involving a predation process, where the subject kills the target in order to eat it or to feed to siblings, offspring or group members. |
| visits | A broad interaction indicating that an organism makes approaches or makes contact with another organism, without specifying the purpose, mechanism, or ecological outcome of the interaction. |
| visits flowers of | A broad interaction indicating that an organism makes approaches or makes contact with the flowers of another plant, without specifying the purpose, mechanism, or ecological outcome of the interaction. |

biotic interactions, capturing relationships between organisms (e.g., feeding, predation, parasitism, or visitation), which are fundamental to understanding biodiversity. The closest related efforts are found in VQA tasks, which evaluate multimodal understanding through both open-ended and closed-ended questions (Hu et al., 2025; Gauba et al., 2026) (e.g., *Is there a symbiotic relationship between the coral depicted in the center of the image and zooxanthellae?*), and in image retrieval, where models learn joint embeddings to align images with queries describing relationships between entities, leveraging multimodal large language models as out-of-the-box rerankers (Vendrow et al., 2024) (e.g., *Does this image show {some query}? Answer with "Yes" or "No" and nothing else.*). **(2) Systematic generation of semantically controlled natural language description.** Unlike domains such as mathematics or programming, where ground truth can often be specified through formal rules, subtle linguistic changes in biotic interactions can preserve or fundamentally alter their meaning. Their structured representation as interaction triplets enables BIOINTERACT to systematically generate controlled natural-language variants that preserve or modify relational semantics. This allows fine-grained evaluation of semantic understanding while reducing the variation and biases associated with human- or prompt-generated perturbations (Felizzi et al., 2025; Koddenbrock et al., 2025; Cocchieri et al., 2026; Sadanandan & Behzadan, 2026).

## 3 BIOINTERACT

BIOINTERACT is a large-scale multimodal image dataset specifically developed for understanding biotic interactions between species, a key component of ecological and biodiversity research. Unlike conventional biodiversity datasets that primarily focus on recognizing individual species, BIOINTERACT targets interactions between organisms, requiring models to identify not only which taxa are present but also who interacts with whom and how (see Table 2). The structured source-relation-target interaction representation enables the controlled generation of natural-language variants for fine-grained semantic evaluation.

### 3.1 DATASET CURATION

We retrieve biotic interaction data from Global Biotic Interactions (GloBI, 2025), using the archive hosted on Zenodo (version 0.8). GloBI periodically collects biotic interactions from a wide range of independent datasets, such as iNaturalist and Encyclopedia of Life (EOL), indexed in tabular form. Biotic interactions are semantic relationships between entities represented as knowledge graphs, stored in a triplet format consisting of a source taxon, an interaction type, and a target taxon. To construct our dataset, we focus on records that are linked to iNaturalist observations, which include images of organisms observed *in situ*. In these records, the source taxon corresponds to the organism documented in the iNaturalist observation, while additional annotations describe its ecological context: interaction type with a target taxon. We filter the GloBI data to retain only interaction records that reference a *research-grade* observation (iNaturalist, 2024), yielding 1.27M records. To these records we aggregate corresponding iNaturalist observation data from GBIF, including taxonomic identity, spatiotemporal information and corresponding images.

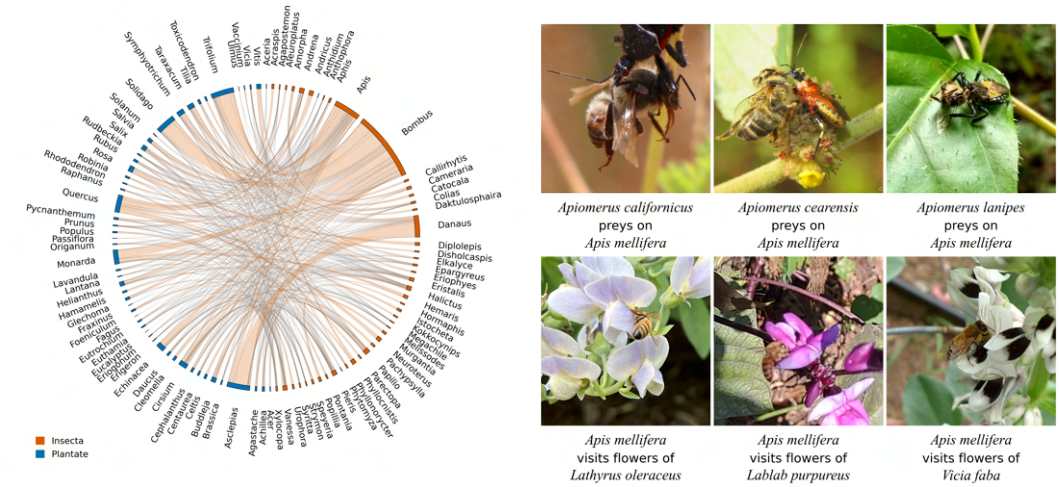


Figure 2: **(left) A chord graph showing BIOINTERACT top 100 most frequent taxa and their interactions.** The dataset is centered around *Insecta-Plantae* interactions. **(right) Example images depicting biotic interaction networks from BIOINTERACT.** Top row showcases predatory interactions of the honeybee (*Apis mellifera*) with rove beetles (genus *Apiomerus*) (Hu et al., 2025); while bottom row showcases its plant preferences, including associations with families such as *Fabaceae* (Baena-Díaz & Dáttilo, 2026).

We align all records to a standardized seven-level Linnaean taxonomy (*kingdom*, *phylum*, *class*, *order*, *family*, *genus*, *species*) using the GBIF Backbone Taxonomy (2023)—the most widely used, open-access, community-developed standard for sharing biodiversity data, providing a stable, standardized language for describing biological specimens and observations—and discard records whose taxonomic resolution falls outside this hierarchy. From the 1.27M biotic interaction records, we retain 636K records, corresponding to 215K unique biotic interactions. We further enrich the dataset with corresponding English vernacular names from GBIF Backbone Taxonomy (2023).

Consequently, we do not treat iNaturalist as a primary source of valid biotic interactions. Biotic interaction with image data are documented *a priori* via iNaturalist; however, these annotations are not routinely verified as the *research-grade* observations (iNaturalist, 2024). Instead, we retain only annotated interactions that are independently supported by at least one additional resource (a separate curated dataset, other than iNaturalist) indexed in GloBI (2025). This cross-validation procedure ensures that the interactions have been independently recorded across sources, yielding a final set of 15.4K unique biotic interactions represented by 256K image records.

### 3.2 DATASET OVERVIEW

BIOINTERACT standardizes interactions to nine types defined by Open Biological and Biomedical Ontology Foundry (OBO) (see Table 2), while interacting taxa are reported at heterogeneous taxonomic levels across five kingdoms (*Animalia*, *Plantae*, *Fungi*, *Chromista*, and *incertae sedis*), with over 99% taxa recorded at the fine-grained *species* levels (see Appendix A). Based on the most recent IUCN (2026) Red List assessment, BIOINTERACT covers 87 threatened species, and 23 of the world's worst invasive species Global Invasive Species Database (2026), enabling evaluation on ecologically critical and conservation-relevant taxa (Groom et al., 2021). The spatiotemporal metadata associated with each image is particularly valuable for studying biotic interactions, as it supports region-specific and seasonal analyses that are fundamental to ecological research, especially in the study of invasive species (Mikula et al., 2026).

Interactions are primarily recorded between insects and plants, indicating a strong skew toward plant-associated ecological relationships (see Figure 2). This focus is well justified, as recent studies report strong declines worldwide in both insects and plants, threatening plant diversity, food security and ecosystem stability (Pan et al., 2024). Interaction types are hierarchically organized, in which the

Table 3: **BIOINTERACT annotations enumerating 27 fields provided in Parquet format**. The dataset is openly available for download and browsing on HuggingFace Datasets.

| Type | Description |
|---|---|
| Source/target scientific name | Scientific taxon name represented as a string in fields `sourceTaxonName` and `targetTaxonName`. |
| Interaction type | Interactions standardized according to Open Biological and Biomedical Ontology Foundry (OBO) Relational Ontology (RO). |
| Source/target rank | The taxonomic rank at which the interacting taxa is identified (*kingdom*, *phylum*, *class*, *order*, *family*, *genus*, *species*). |
| Source/target vernacular name(s) | English common or vernacular name(s) represented as list of strings in fields `sourceVernacularName` and `targetVernacularName`. |
| Taxonomic hierarchy | Taxonomic hierarchy deterministically derived from taxa scientific name, represented as strings in separate fields for each source and target (*kingdom*, *phylum*, *class*, *order*, *family*, *genus*). |
| Spatiotemporal context | Date when the interaction was observed (separated DD, MM, YYYY), and latitude, longitude coordinates where the interaction was observed (decimals). |
| Licence | Image usage licence associated with each observation. |
| URL | Downloadable image link from Naturalist represented as a string in field `imageURL`. |

general relation *interacts with* is recursively refined into more specific relations, such as *has host*, and further into fine-grained sub-relations like *parasite of*. Notably, this hierarchy is not strictly tree-structured: relations at the same level may overlap or exhibit cross-links, reflecting the compositional nature of biological interactions. For example, a bee can both visit a plant and visit flowers of it; parasitoid wasps can be both parasitoid of and functionally parasite of a host, illustrating cross-links between branches.

**License.** The dataset contains a heterogeneous mixture of Creative Commons licenses, predominantly non-commercial licenses (CC BY-NC 4.0), all of which are provided and should be considered accordingly in downstream usage.

**Geoprivacy.** We include geolocation metadata for all records, relying on the source platforms' automated and user-specified obscuration for sensitive species (iNaturalist, 2025). This means in practice that endangered or protected taxa have deliberately imprecise coordinates, in line with geoprivacy best practices.

**Responsible use.** Models trained on this data should not be used to infer, exploit, or manipulate sensitive biotic interactions in ways that could harm species, disrupt ecosystems, or facilitate unlawful activities such as wildlife persecution, invasive species spread, or targeted ecological interference; the dataset is provided to support ecological research, conservation, and responsible environmental monitoring.

### 3.3 SEMANTIC PERTURBATIONS

Biotic interactions are inherently directional, and can be expressed through multiple equivalent formulations. Let $\mathcal{D} = \{(x_i, s_i, r_i, t_i)\}_{i=1}^{N}$ denote a dataset of $N$ samples, where $x_i$ is an image, annotated with interaction triplets $(s_i, r_i, t_i)$, where $s_i$ and $t_i$ are the source and target taxa, respectively, and $r_i$ is the interaction type. For each image $x$, BIOINTERACT enables controlled generation of meaning preserving queries, and contradictory queries that alter salient attributes such as taxa or interaction type leveraging rich image annotations (see Table 3). Controlled perturbation is preferred to ensure semantic equivalence and avoid hallucinations or unintended biases that may arise from generative approaches Wei et al. (2026).

Given an image $x$ and an interaction $(s, r, t)$, we construct queries through five perturbation types: (1) *entity removal*: the source or the target is removed; (2) *entity replacement*: the source or the target is randomly replaced with another entity; (3) *relation removal*: the relation is removed, preserving the source and the target; (4) *relation replacement*: the relation is replaced, while preserving the source or the target; (5) *relation inversion*: reverse the meaning of the relation, while preserving the source and the target. To preserve semantic validity, we ensure that replacements of entities and relations remain ecologically plausible by leveraging hierarchical structures in both taxonomy and interaction types. Specifically, we generate alternative variants of $s_i$ and $t_i$ by sampling across different taxonomic levels, and substitute $r_i$ with relations drawn from the same interaction-type hierarchy (see Appendix A). Conversely, to construct contradictory statements, we deliberately replace entities and relations with elements that fall outside these hierarchies, thereby violating ecological and semantic constraints.

Table 4: **Performance of different CLIP-style models over semantically equivalent queries.** We report the mean scores over five semantically equivalent queries. **Bold** and underlined entries indicate the **best** and the second best results, respectively.

| Model | Method | Params (M) | mAP@50 | mMRR | mRecall@10 |
|---|---|---|---|---|---|
| CLIP | ViT-L-14 | 427 | $0.14 \pm 0.02$ | $0.34 \pm 0.06$ | $0.56 \pm 0.05$ |
| MetaCLIP | ViT-L-14-quickgelu | 427 | $0.25 \pm 0.06$ | $0.49 \pm 0.10$ | $0.72 \pm 0.11$ |
| SigLIP | SO400M-14-SigLIP | 877 | $0.35 \pm 0.10$ | $0.60 \pm 0.11$ | $0.78 \pm 0.08$ |
| SigLIP2 | ViT-L-16-SigLIP2-256 | 881 | $0.37 \pm 0.10$ | $0.63 \pm 0.10$ | $0.83 \pm 0.11$ |
| BioTrove-BioCLIP | ViT-B-16 | 149 | $0.30 \pm 0.08$ | $0.57 \pm 0.11$ | $0.78 \pm 0.09$ |
| BioCAP | ViT-B-16 | 149 | $0.50 \pm 0.06$ | $0.78 \pm 0.04$ | $0.93 \pm 0.03$ |
| BioCLIP | ViT-B-16 | 149 | $0.30 \pm 0.06$ | $0.63 \pm 0.07$ | $0.82 \pm 0.07$ |
| BioCLIP2 | ViT-L-14 | 427 | $\mathbf{0.61 \pm 0.08}$ | $\mathbf{0.83 \pm 0.05}$ | $\mathbf{0.95 \pm 0.02}$ |

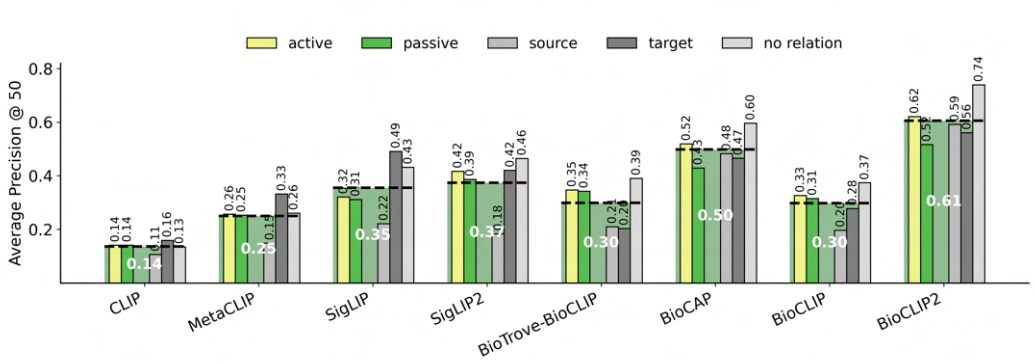


Figure 3: **Performance of different CLIP-style models per semantically equivalent query.** Queries are based on interaction triples $(s, r, t)$: **active** queries specify the full triple $(s, r, t)$, **passive** queries invert the interaction $(t, r', s)$, **source** queries condition on $(s, r)$, **target** queries on $(t, r')$, and **no relation** queries specify only entity pairs without the relation $(s, t)$.

Each perturbation can be augmented with multiple levels of semantic granularity by incorporating both scientific and vernacular names, as well as hierarchical variants of taxonomy and interactions types.

## 4 Evaluation

Rather than focusing on accuracy under a canonical query, our evaluation considers how predictions change under semantic perturbations, while the visual input remains constant. We introduce a ranking task similar to INQUIRE (Vendrow et al., 2024), which fixes images for each query and uses models like GPT-4o Hurst et al. (2024) to improve over initial text-to-image CLIP-style retrieval. We, thus, measure fine-grained semantic understanding in embedding similarity and large language models with vision capabilities binary (yes/no) questions (see Appendix C).

**BioInteract100.** We construct a subset dataset via stratified sampling, selecting 100 image samples per interaction triplet group. Samples are prioritized based on annotation quality, favoring those with available vernacular names for both taxa and *species*-level annotations. Within each triplet group, we promote visual diversity by selecting at most one image sample per observation in the first pass, and subsequently filling remaining slots if necessary (iNaturalist observations frequently contain multiple near-duplicate photographs of the same encounter). This procedure yields BioInteract100, a subset of 28K images, depicting 281 unique interactions of taxa from *Animalia*, *Plantae*, and *Fungi*, balancing data quality, taxonomic specificity, and visual diversity across seven interaction types. We then generate five semantically equivalent statements using the aforementioned perturbations. To stress-test robustness, we also generate two contradictory statements that deliberately invert entity roles and relations. Appendix B lists all query variants for each of the 281 unique interactions.

Table 5: **Performance on different proprietary models grouped per query perturbation type.** First we group semantically equivalent queries. Next we group contradictory queries which deliberately reverse entities and relations. We report the accuracy and flip rate for each group. Lower flip rates indicate greater consistency. **Bold** entries indicate the **best** results. This evaluation excludes *interacts with* interaction type queries, due to its inherent symmetry.

| | Equivalent | | Contradictory | | All | |
|---|---|---|---|---|---|---|
| **Model** | **Acc** | **Flip** | **Acc** | **Flip** | **Acc** | **Flip** |
| *Text-only baselines* | | | | | | |
| GPT 5.2 mini | **0.97 ± 0.02** | **0.10** | **1.0 ± 0.0** | **0.0** | **0.98 ± 0.02** | **0.10** |
| Gemini 3 Flash Preview | 0.05 ± 0.03 | 0.23 | 0.60 ± 0.14 | 0.35 | 0.21 ± 0.26 | 0.81 |
| Claude Sonnet 4.6 | 0.90 ± 0.08 | 0.41 | 0.99 ± 0.0 | 0.71 | 0.93 ± 0.07 | 0.41 |
| *Image-text* | | | | | | |
| GPT 5.2 mini | 0.74 ± 0.12 | 0.54 | 0.52 ± 0.14 | **0.35** | 0.68 ± 0.17 | 0.95 |
| Gemini 3 Flash Preview | 0.77 ± 0.15 | 0.67 | 0.59 ± 0.20 | 0.44 | 0.71 ± 0.20 | 0.97 |
| Claude Sonnet 4.6 | **0.81 ± 0.03** | **0.44** | **0.65 ± 0.26** | 0.60 | **0.77 ± 0.09** | **0.71** |

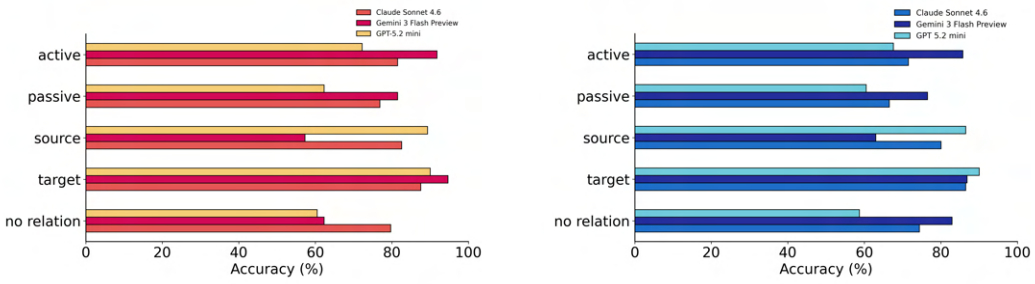


Figure 4: **Performance on natural language understanding of biotic interaction queries grouped by language specificity: (left) scientific terminology; (right) vernacular names.** For each query type we report accuracy (see Appendix E). This evaluation excludes *interacts with* interaction type queries, due to its inherent symmetry.

**Models.** We evaluate CLIP-style models for text-to-image retrieval, such as CLIP (Radford et al., 2021), MetaCLIP (Chuang et al., 2026), SigLIP (Zhai et al., 2023), SigLIP2 (Tschannen et al., 2025), as well as specialized models, such as BioTrove-BioCLIP (Yang et al., 2024), BioCAP (Zhang et al., 2026), BioCLIP (Stevens et al., 2024) and BioCLIP2 (Gu et al., 2026). We adopt proprietary multimodal language models (MLLMs) such as OpenAI GPT-5.2 mini, Anthropic Claude Sonnet 4.6, and Gemini 3 Flash Preview Google for ranking prompting for each image and corresponding queries per image: *Does this image show {some query}? Answer with "Yes" or "No" and nothing else*. Experimental setup is further detailed in Appendix C.

**Metrics.** We report the Average Precision at k (AP@50), Mean Reciprocal Rank (MRR), and Recall at k (Recall@10). We also report accuracy and flip rate over MLLMs performance. Beyond accuracy, we measure whether models answer consistently across query paraphrases: if the model produces inconsistent predictions across this set—i.e., at least one prediction differs—we count this as a flip. The flip rate is the proportion of such instances over the dataset.

**Results.** We report retrieval evaluation of meaning preserving queries on BIOINTERACT100 in Table 4 and Figure 3 (see Appendix E). High-quality training data and scale is crucial for expert-level queries. While models like BioCLIP and BioTrove-BioCLIP were pre-trained on images and taxonomic label queries using TreeOfLife-10M (Stevens et al., 2024), they underperform comparatively to larger general models like SigLIP and SigLIP2. BioCLIP2, the largest specialized model trained on TreeOfLife-200M (Gu et al., 2026), achieves the highest overall retrieval score. Although trained on the same dataset as BioCLIP and BioTrove-BioCLIP, BioCAP's complementary synthetic descriptive caption alignment beyond taxonomic labels achieves competitive performance with BioCLIP2.

Table 6: **Performance on relation-direction reversal queries.** We group meaning-preserving queries with their syntactically similar but semantically contradictory counterparts: **active** queries specifying the full triple $(s, r, t)$, contrasted with $(t, r, s)$; **passive** queries inversing the relation $(t, r', s)$, contrasted with $(s, r', t)$. This evaluation excludes *interacts with* interaction type queries, due to its inherent symmetry (compositionality-tuned encoders evaluation is included in Table 10).

| | By active | | | By passive | | |
|---|---|---|---|---|---|---|
| **Model** | **AP@50** | **MRR** | **Recall@10** | **AP@50** | **MRR** | **Recall@10** |
| SigLIP2 | 0.30 | 0.41 | 0.47 | 0.27 | 0.39 | 0.47 |
| BioCAP | 0.30 | 0.43 | 0.41 | 0.27 | 0.41 | 0.48 |
| BioCLIP2 | **0.32** | **0.44** | **0.49** | **0.28** | **0.41** | **0.48** |

**Different queries present challenges of varying difficulty.** Figure 3 illustrates difference in performance across the five semantically equivalent statements. We observe that, for specialized models, queries containing only the interacting entities—without explicitly specifying the relation—yield the best performance. This is likely because images in BIOINTERACT100 often depict small organisms, with the visual signal dominated by environmental context providing cues of other relevant organisms.

**High accuracy on paraphrasing does not imply consistency.** We report performance of general-purpose proprietary models in Table 5 and Figure 4. First, we evaluate a text-only control condition in which prompts refer to an image despite no image being supplied. Strong performance in this setting reflects correct detection of absent visual input rather than visual recognition capability, with Gemini variant suggesting weaker modality-awareness and lower overall calibration compared to GPT and Claude variants. In image-text settings, MLLMs struggle to maintain stable predictions as reflected by their high flip rates. Claude Sonnet 4.6 seems to achieve the best overall balance between accuracy and consistency, exhibiting both strong performance and lower flip rates compared to other models. In this setting, apparent gains in accuracy can reflect overfitting to benchmark adversarial perturbations rather than genuine semantic understanding. Furthermore, as shown in Figure 4, domain-specific language has no significant impact on the performance of general-purpose proprietary models.

**Biotic interactions are direction-sensitive, and even the best models are affected by reversal.** We report image retrieval evaluation of relation-direction reversals adversarial queries in Table 6. We find that these limitations are particularly pronounced in real-world, direction-sensitive settings, where models frequently fail to distinguish between natural adversarial instances.

**Limitations.** While our labels rely on robust community consensus and literature validation, they remain susceptible to occasional human error (see Appendix D). We report significant zero-shot results, showing that pretrained embeddings can effectively navigate this complex data. However, applying few-shot learning—where models are calibrated using a handful of gold-standard examples—presents a clear path to further enhance performance through task-specific adaptation. Secondly, in-the-wild community science imagery lacks explicit visual bounding boxes, and interacting organisms may be heavily occluded or out of focus. Future work is required to disentangle whether the observed performance gaps stem from a failure in fine-grained multi-entity localization or a genuine deficit in relational reasoning. Consequently, BIOINTERACT100 naturally serves as a challenging, real-world testbed for both few-shot semantic adaptation and future models with localized visual grounding.

## 5 CONCLUSION

We introduce BIOINTERACT, the largest publicly available multimodal dataset designed to better evaluate biotic interactions understanding and ultimately, accelerate trustworthy AI solutions for biodiversity. This dataset focuses on fine-grained semantic understanding, surpassing existing datasets in scale and annotation richness. Our controlled adversarial benchmark over the subset BIOINTERACT100, an image retrieval task on a fixed set, underscores the importance of stress-testing multimodal understanding, particularly for tasks that require inferring meaning from visual evidence and expressing it through diverse, semantically equivalent natural language forms. With BIOINTERACT, we aim to accelerate the development of multimodal AI models for biodiversity that are robust to the nuanced challenges of species interactions in real-world environments.

### AI USE STATEMENT

In this work, we used generative AI tools as models evaluated in our experiments, including LLMs and vision-language models for assessing biotic interaction understanding and robustness to query perturbations (see Section 4 and Appendix C). Generative AI tools were not used to generate ground-truth annotations.

### ETHICS STATEMENT

Our study involves human verification of 2,810 images from BIOINTERACT100 (see Appendix D). Each image was visually inspected and verified whether the labeled biotic interaction was observable in the image. No personal or identifiable information was collected, and the task involved no sensitive data or personal attributes. The verification was conducted solely to assess the visual validity of the dataset annotations and to distinguish potential annotation or visual-signal limitations from model reasoning failures.

### REPRODUCIBILITY STATEMENT

We have released all source code, including data collection and filtering, and evaluation at anonymous.4open.science/r/biointeract-E615. The corresponding dataset, extended from GloBI with metadata from iNaturalist, is available at huggingface.co/datasets/BioInteract/BioInteract. The details regarding dataset collection is presented in Section 3.1. The details to reproduce our evaluation results are included in Appendix E.

## REFERENCES

Luis Abdala-Roberts, Adriana Puentes, Deborah L Finke, Robert J Marquis, Marta Montserrat, Erik H Poelman, Sergio Rasmann, Arnaud Sentis, Celia C Symons, Nicole M van Dam, et al. Connecting the dots: Managing species interaction networks to mitigate the impacts of global change. *eLife*, 14, 2025.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Anthropic. Claude sonnet 4.6. URL `https://www.anthropic.com/claude/sonnet`. Accessed on 2026-05-06.

Fernanda Baena-Díaz and Wesley Dáttilo. Plant traits driving visits by the honeybee apis mellifera across native and introduced ranges: A global analysis. *Ecological Entomology*, 51(1), 2026.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37, 2024.

Zerui Cheng, Stella Wohnig, Ruchika Gupta, Samiul Alam, Tassallah Abdullahi, João Alves Ribeiro, Christian Nielsen-Garcia, Saif Mir, Siran Li, Jason Orender, et al. Position: Benchmarking is broken-don't let ai be its own judge. *Advances in Neural Information Processing Systems*, 38, 2026.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *Advances in Neural Information Processing Systems*, 38, 2026.

Alessio Cocchieri, Luca Ragazzi, Giuseppe Tagliavini, and Gianluca Moro. Remedqa: Are we done with medical multiple-choice benchmarks? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2657–2668. IEEE, 2023.

Encyclopedia of Life (EOL). Encyclopedia of Life (EOL). URL https://eol.org. Accessed on 2026-03-25.

Federico Felizzi, Olivia Riccomi, Michele Ferramola, Francesco Andrea Causio, Manuel Del Medico, De Vita Vittorio, Lorenzo De Mori, Alessandra Piscitelli, Pietro Eric Risuleo, Bianca Destro Castaniti, Antonio Cristiano, Alessia Longo, Luigi De Angelis, Mariapia Vassalli, and Marcello Di Pumpo. Are large vision language models truly grounded in medical images? evidence from italian clinical visual question answering. In *Multimodal Representation Learning for Healthcare (MMRL4H) Workshop at EurIPS*, 2025.

Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram Adve, and Yu-Xiong Wang. Agmmu: A comprehensive agricultural multimodal understanding benchmark. *Advances in Neural Information Processing Systems*, 38, 2026.

GBIF Backbone Taxonomy. GBIF Backbone Taxonomy, 2023. URL https://doi.org/10.15468/39omei. Accessed on 2026-08-12.

Global Invasive Species Database. 100 of the world's worst invasive alien species, 2026. URL http://www.iucngisd.org/gisd/100_worst.php. Accessed on 2026-08-12.

GloBI. Global Biotic Interactions: Interpreted Data, 2025. URL https://doi.org/10.5281/zenodo.14640564. Accessed on 2026-08-12.

Google. Gemini 3 flash preview. URL https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview. Accessed on 2026-05-06.

Quentin Groom, Nadja Pernat, Tim Adriaens, Maarten de Groot, Sven D. Jelaska, Diana Marčiulynienė, Angeliki F. Martinou, Jiří Skuhrovec, Elena Tricarico, Ernst C. Wit, and Helen E. Roy. Species interactions: next-level citizen science. *Ecography*, 44(12), 2021.

Jianyang Gu, Sam Stevens, Elizabeth Campolongo, Matthew Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander White, James Balhoff, et al. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. *Advances in Neural Information Processing Systems*, 38, 2026.

Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel CF Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, et al. The illusion of readiness in health ai. *arXiv preprint arXiv:2509.18234*, 2025.

Fang-Shuo Hu, Yun Hsiao, and Alexey Solodovnikov. A global citizen science effort via inaturalist reveals food webs of large predatory rove beetles. *Food Webs*, 43, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

iNaturalist. URL https://www.inaturalist.org. Accessed on 2026-08-12.

iNaturalist. We estimate the accuracy of Research Grade observations to be 95% correct!, January 2024. URL https://www.inaturalist.org/blog/89255-we-estimate-the-accuracy-of-research-grade-observations-to-be-95-correct. Accessed on 2026-01-02.

iNaturalist. What is geoprivacy? what does it mean for an observation to be obscured?, 2025. URL https://help.inaturalist.org/en/support/solutions/articles/151000169938-what-is-geoprivacy-what-does-it-mean-for-an-observation-to-be-obscured-%7D. Accessed on 2026-08-12.

IUCN. The IUCN Red List of Threatened Species, 2026. URL https://www.iucnredlist.org. Accessed on 2026-08-12.

Faizan Farooq Khan, Xiang Li, Andrew J Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

Mario Koddenbrock, Rudolf Hoffmann, David Brodmann, and Erik Rodner. On the domain robustness of contrastive vision-language models. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 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*, 2024.

Fan Liu, Jindong Han, Tengfei Lyu, Weijia Zhang, Zherui Yang, Lu Dai, Cancheng Liu, and Hao Liu. Foundation models for scientific discovery: From paradigm enhancement to paradigm transition. *Advances in Neural Information Processing Systems*, 38, 2026a.

Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. Fgbench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. *Advances in Neural Information Processing Systems*, 38, 2026b.

Georgiana Manolache, Gerard Schouten, and Joaquin Vanschoren. Crypticbio: A large multimodal dataset for visually confusing species. *Advances in Neural Information Processing Systems*, 38, 2026.

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, 2024.

Peter Mikula, Pavel Pipek, Martin Bulla, María L Castillo, Shawan Chowdhury, Łukasz Dylewski, Franz Essl, Josh A Firth, Jérôme MW Gippet, Theresa Henke, et al. Harnessing iecology data to uncover invasive species behaviour. *Methods in Ecology and Evolution*, 2026.

Observation.org. URL https://observation.org. Accessed on 2026-08-12.

Open Biological and Biomedical Ontology Foundry (OBO). Relations Ontology (RO). URL https://obofoundry.org/ontology/ro.html. Accessed on 2026-08-12.

OpenAI. Gpt-5.4 mini. URL https://developers.openai.com/api/docs/models/gpt-5.4-mini. Accessed on 2026-05-06.

Kaixuan Pan, Leon Marshall, Geert R de Snoo, and Jacobus C Biesmeijer. Dutch landscapes have lost insect-pollinated plants over the past 87 years. *Journal of Applied Ecology*, 61(6), 2024.

Nadja Pernat, Susan Canavan, Marina Golivets, Jasmijn Hillaert, Yuval Itescu, Ivan Jarić, Hjalte MR Mann, Pavel Pipek, Cristina Preda, David M Richardson, et al. Overcoming biodiversity blindness: Secondary data in primary citizen science observations. *Ecological Solutions and Evidence*, 5(1), 2024a.

Nadja Pernat, Daniyar Memedemin, Tom August, Cristina Preda, Lien Reyserhove, Jens Schirmel, and Quentin Groom. Extracting secondary data from citizen science images reveals host flower preferences of the mexican grass-carrying wasp isodontia mexicana in its native and introduced ranges. *Ecology and Evolution*, 14(6), 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 2021.

Binesh Sadanandan and Vahid Behzadan. PSF-Med: Measuring and explaining paraphrase sensitivity in medical vision language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 2020.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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, 2024.

Edward Vendrow, Julia Chae, Rupa Kurinchi-Vendhan, Isaac Eckert, Jazlynn Hall, Marta Jarzyna, Reymond Miyajima, Ruth Oliver, Laura Pollock, Lauren Shrack, et al. Inquire-search: Interactive discovery in large-scale biodiversity databases. *Ecological Informatics*, 2026.

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*, 2024.

Wei Wang, Gaowei Zhang, Mingjie Li, Yi Wang, et al. Coralvqa: A large-scale visual question answering dataset for coral reef image understanding. *Advances in Neural Information Processing Systems*, 38, 2026.

Qiyao Wei, Edward R Morrell, Lea Goetz, and Mihaela van der Schaar. Semantic-kg: Using knowledge graphs to construct benchmarks for measuring semantic similarity. *Advances in Neural Information Processing Systems*, 38, 2026.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9), 2018.

Zhenlin Xu, Yi Zhu, Siqi Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Chih-Hsuan Yang, Ben Feuer, Zaki Jubery, Zi K Deng, Andre Nakkab, Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh K Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37, 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 the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

Ziheng Zhang, Xinyue Ma, Arpita Chowdhury, Elizabeth Campolongo, Matthew Thompson, Net Zhang, Samuel Stevens, Hilmar Lapp, Tanya Berger-Wolf, Yu Su, et al. Biocap: exploiting synthetic captions beyond labels in biological foundation models. In *International Conference on Learning Representations*, 2026.

Yang Zhou, Tan L Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu, and Rick S Goh. Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. *Advances in Neural Information Processing Systems*, 37, 2024.

## A Dataset Details



Figure 5: **Example images depicting insects visiting plants from BioInteract.**

BioInteract contains 256K ecological images depicting 15.4K unique biotic interactions between organisms across five kingdoms, primarily between *Animalia* and *Plantae* (see Figure 6). Biotic interactions are represented as triplets consisting of a source taxon, an interaction type, and a target taxon, reported at heterogeneous taxonomic levels. Interactions are standardized to nine types defined

Table 7: **Key statistics of BIOINTERACT.** Biotic interactions are recorded between organisms at different taxonomic levels. The table includes the number of unique taxa (source and target taxa) recorded at the *species* and *genus* levels.

| Statistics | eats | has host | interacts with | parasite of | parasitoid of | pathogen of | preys on | visits | visits flowers of | All |
|---|---|---|---|---|---|---|---|---|---|---|
| # Images | 27778 | 6926 | 116483 | 2558 | 13 | 30 | 2878 | 1636 | 98456 | 256758 |
| # Species | 2058 | 961 | 3924 | 219 | 4 | 4 | 370 | 198 | 1628 | 9366 |
| # Genus | 742 | 230 | 1452 | 36 | - | 2 | 109 | 42 | 523 | 3136 |

by Open Biological and Biomedical Ontology Foundry (OBO) (see Figure 7). Biotic interaction triplets are documented *a priori* via the iNaturalist platform and cross-validated with at least one additional GloBI-indexed data source, establishing them as ground-truth annotations. More on annotation quality see Appendix D.

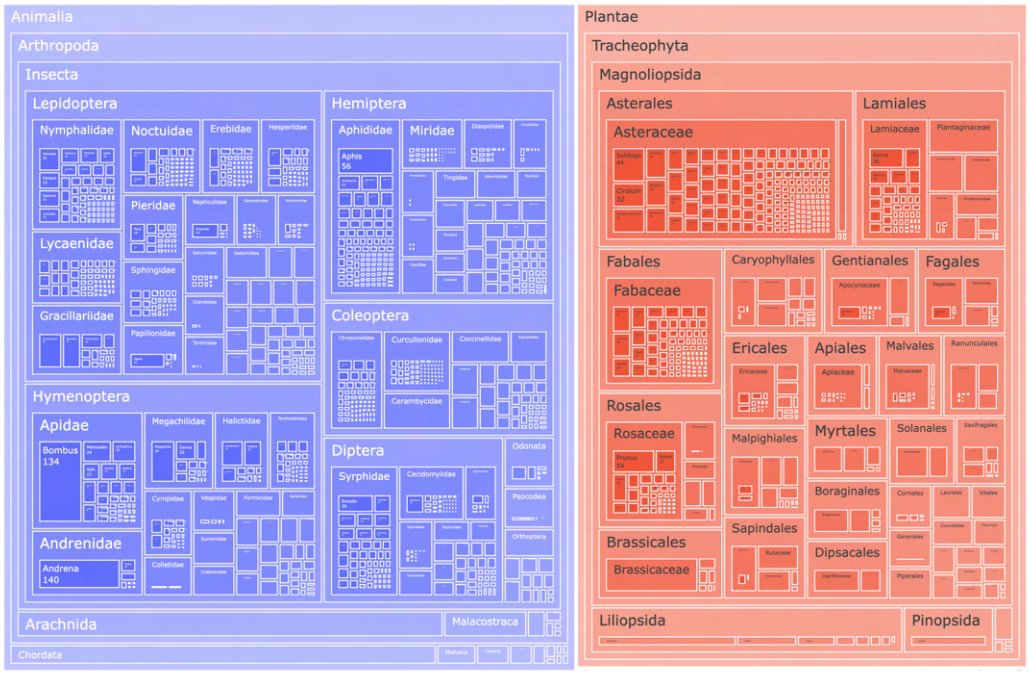


Figure 6: **Treemap visualization of the species covered in BIOINTERACT.**

Spatio-temporal data plays a crucial role in biodiversity datasets because species distributions and appearances are strongly influenced by both geographic location and time (Manolache et al., 2026). Figure 8 illustrates the spatiotemporal distribution. This information is particularly important for monitoring threatened species, whose populations are often restricted to small geographic areas or sensitive ecosystems. Subsequently, it is critically important not to disclose the precise locations of threatened species because doing so can inadvertently put them at even greater risk. Many vulnerable species face threats from poaching, illegal wildlife trade, habitat disturbance, and over-collection. Sharing exact geographic coordinates, especially online or in open databases, can make it possible to locate and exploit these species. To mitigate the risks associated with the disclosure of sensitive biodiversity data, iNaturalist implements automatic geoprivacy measures for taxa listed on the global IUCN.

Current domain-specific evaluations (e.g., CoralVQA, AgMMU, INQUIRE) assess VLMs holistically against a fixed set of queries or labels without testing robustness to alternative formulations of the same underlying concept. In contrast, BIOINTERACT's structured and hierarchical source–relation–target

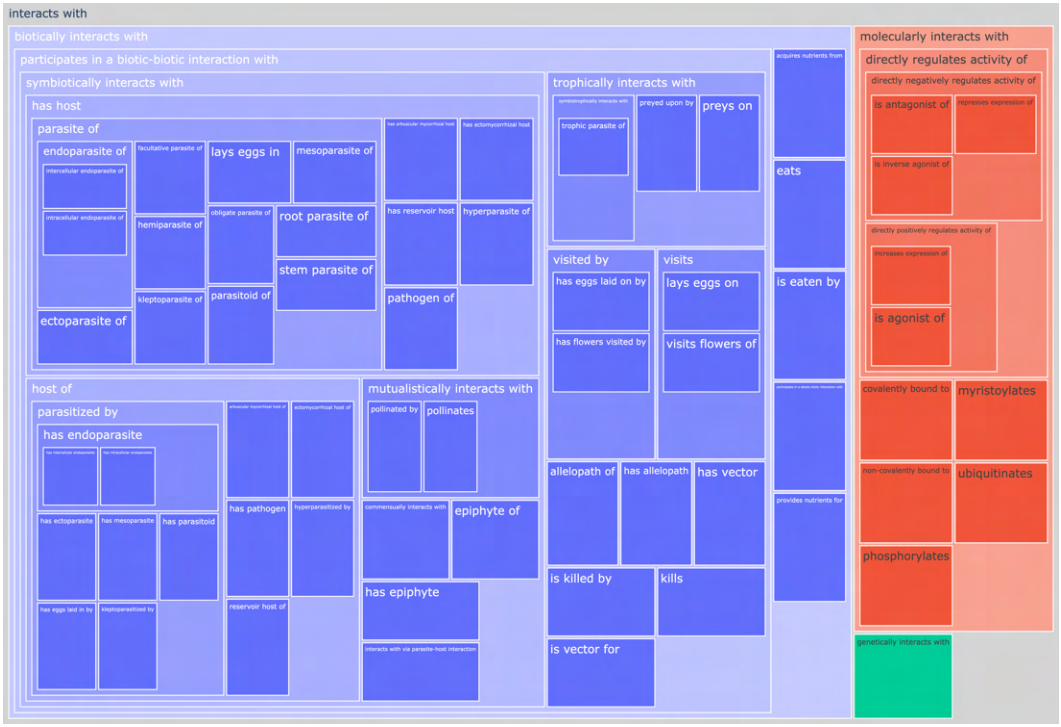


Figure 7: **Treemap visualization of the interaction hierarchy used in Open Biological and Biomedical Ontology Foundry (OBO).**

annotations enable the systematic generation of grounded query variants that preserve or manipulate linguistic form, interaction semantics, directionality, and taxonomic specificity (e.g., *preys on*, *eats*, *interacts with*, see Table 2), allowing new evaluation sets to be generated beyond the current BIOINTERACT100 subset without relying on LLM-generated variations that may introduce hallucinated or semantically incorrect formulations, thereby reducing dependence on fixed query sets and supporting continued evaluation of fine-grained semantic understanding rather than claiming to eliminate the risk of memorization. Rather than claiming to eliminate this problem, we argue for benchmarks that support systematic generation of new, grounded query variants and call for continued evaluation beyond fixed query sets.

## B BIOINTERACT100

The 256K-image BIOINTERACT dataset constitutes the full multimodal resource introduced in this work. BIOINTERACT100 was derived specifically for controlled robustness evaluation, where each interaction is represented by a fixed and sufficiently large number of images (i.e., 100 image samples for 281 interactions). Interactions "parasitoid of" (13 images) and "pathogen of" (30 images) were excluded because they fall short of the 100 minimum images per interaction required for the subset. This balanced construction is important because robustness metrics require comparable evaluation conditions across interactions. Importantly, we do not intend results obtained on BIOINTERACT100 to be interpreted as robustness results over all 256K images. At the same time, by releasing the full 256K-image dataset, we hope to enable subsequent studies to investigate questions beyond the scope of this controlled benchmark. For example, evaluate interaction retrieval and recognition under the naturally occurring long-tailed distribution, or support the development of new sampling strategies, benchmarks, and training approaches.

BIOINTERACT100 query perturbations are based on interaction triples $(s, r, t)$: **active** queries specify the full triple $(s, r, t)$, **passive** queries invert the interaction $(t, r', s)$, **source** queries condition on $(s, r)$, **target** queries on $(t, r')$, and **no relation** queries specify only entity pairs without the relation

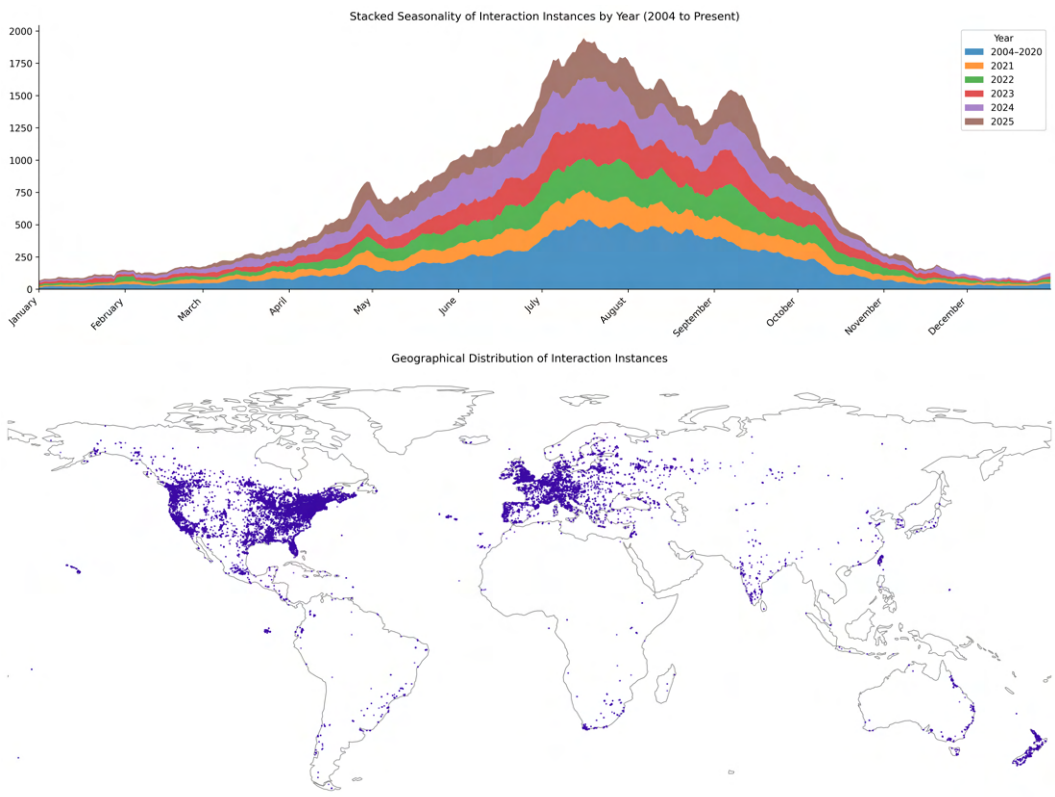


Figure 8: **Spatiotemporal distribution of interaction instances in BIOINTERACT: (top) stacked seasonality; (bottom) geographical distribution.** Majority of records are concentrated in Europe and North America, with a seasonal peak in observations during July.

$(s, t)$, **active reversal** queries swap the source and target from active queries $(t, r, s)$, and **passive reversal** queries swap the source and target from passive queries $(s, r', t)$.

We exclude *interacts with* interaction type from relation-direction reversal evaluations due to its inherent symmetry. The benchmark solely relies on removals and reversals and does not implement hierarchical replacements yet.

## C EXPERIMENTAL DETAILS

We measure fine-grained semantic understanding in embedding similarity using CLIP-style models and binary (yes/no) questions using proprietary multimodal language models (MLLMs) over BIOINTERACT100 (see Table 9). We run our experiments on Nvidia GeForce RTX 3080 GPU using 32 GB of RAM memory.

### C.1 IMAGE-RETRIEVAL PROCEDURE

**CLIP-style models.** For each model, we encode all images in BIOINTERACT100 using its image encoder and encode each interaction query using its text encoder. We note that image embeddings are computed once and reused across query variants, while only the text embeddings change. We then compute the cosine similarity between the query embedding and all image embeddings and rank the images from highest to lowest similarity. Images associated with the queried interaction are treated as relevant, and retrieval performance is evaluated using AP@50, MRR, and Recall@10. The same image embeddings are used across all query perturbations, while only the textual query is modified, allowing us to directly assess how changes in query formulation affect retrieval.

Table 8: **BIOINTERACT100 query examples for each interaction type.**

| Interaction type | Perturbations | Query example |
|---|---|---|
| eats | active | *Tetranychus lintearius* eats *Ulex europaeus* |
| | passive | *Ulex europaeus* is eaten by *Tetranychus lintearius* |
| | source | *Tetranychus lintearius* eats another organism |
| | target | *Ulex europaeus* is eaten by another organism |
| | no relation | *Tetranychus lintearius* and *Ulex europaeus* |
| | active reverse | *Ulex europaeus* eats *Tetranychus lintearius* |
| | passive reverse | *Tetranychus lintearius* is eaten by *Ulex europaeus* |
| has host | active | *Trichilogaster acaciaelongifoliae* has host *Acacia longifolia* |
| | passive | *Acacia longifolia* is hosted by *Trichilogaster acaciaelongifoliae* |
| | source | *Trichilogaster acaciaelongifoliae* has host another organism |
| | target | *Acacia longifolia* is hosted by another organism |
| | no relation | *Trichilogaster acaciaelongifoliae* and *Acacia longifolia* |
| | active reverse | *Acacia longifolia* has host *Trichilogaster acaciaelongifoliae* |
| | passive reverse | *Trichilogaster acaciaelongifoliae* is hosted by *Acacia longifolia* |
| interacts with | active | *Aceria erinea* interacts with *Juglans regia* |
| | passive | *Juglans regia* interacts with *Aceria erinea* |
| | source | *Aceria erinea* interacts with another organism |
| | target | *Juglans regia* interacts with another organism |
| | no relation | *Aceria erinea* and *Juglans regia* |
| | active reverse | *Juglans regia* interacts with *Aceria erinea* |
| | passive reverse | *Aceria erinea* interacts with *Juglans regia* |
| parasite of | active | *Cotesia glomerata* parasite of *Pieris brassicae* |
| | passive | *Pieris brassicae* is parasitized by *Cotesia glomerata* |
| | source | *Cotesia glomerata* parasite of another organism |
| | target | *Pieris brassicae* is parasitized by another organism |
| | no relation | *Cotesia glomerata* and *Pieris brassicae* |
| | active reverse | *Pieris brassicae* parasite of *Cotesia glomerata* |
| | passive reverse | *Cotesia glomerata* is parasitized by *Pieris brassicae* |
| prey on | active | *Philanthus triangulum* preys on *Apis mellifera* |
| | passive | *Apis mellifera* is preyed on by *Philanthus triangulum* |
| | source | *Philanthus triangulum* preys on another organism |
| | target | *Apis mellifera* is preyed on by another organism |
| | no relation | *Philanthus triangulum* and *Apis mellifera* |
| | active reverse | *Apis mellifera* preys on *Philanthus triangulum* |
| | passive reverse | *Philanthus triangulum* is preyed on by *Apis mellifera* |
| visits | active | *Bombus pensylvanicus* visits *Helianthus annuus* |
| | passive | *Helianthus annuus* is visited by *Bombus pensylvanicus* |
| | source | *Bombus pensylvanicus* visits another organism |
| | target | *Helianthus annuus* is visited by another organism |
| | no relation | Bombus pensylvanicus and Helianthus annuus |
| | active reverse | *Helianthus annuus* visits *Bombus pensylvanicus* |
| | passive reverse | *Bombus pensylvanicus* is visited by *Helianthus annuus* |
| visits flowers of | active | *Agapostemon virescens* visits flowers of *Heliopsis helianthoides* |
| | passive | *Heliopsis helianthoides* is visited by *Agapostemon virescens* |
| | source | *Agapostemon virescens* visits flowers of another organism |
| | target | *Heliopsis helianthoides* is visited by another organism |
| | no relation | *Agapostemon virescens* and *Heliopsis helianthoides* |
| | active reverse | *Heliopsis helianthoides* visits flowers of *Agapostemon virescens* |
| | passive reverse | Agapostemon virescens is visited by *Heliopsis helianthoides* |

Table 9: **Model stack.**

| Model | URL |
|---|---|
| CLIP ViT-L-14 (Radford et al., 2021) | huggingface.co/openai/clip-vit-large-patch14 |
| MetaCLIP ViT-L-14-quickgelu (Chuang et al., 2026) | huggingface.co/facebook/metaclip-2-worldwide-huge-quickgelu |
| SigLIP SO400M-14-SigLIP (Zhai et al., 2023) | huggingface.co/timm/ViT-SO400M-14-SigLIP |
| SigLIP2 ViT-L-16-SigLIP2-256 (Tschannen et al., 2025) | huggingface.co/timm/ViT-L-16-SigLIP2-256 |
| BioTrove-BioCLIP ViT-B-16 (Yang et al., 2024) | huggingface.co/BGLab/BioTrove-CLIP |
| BioCAP ViT-B-16 (Zhang et al., 2026) | huggingface.co/imageomics/biocap |
| BioCLIP ViT-B-16 (Stevens et al., 2024) | huggingface.co/imageomics/bioclip |
| BioCLIP2 ViT-L-14 (Gu et al., 2026) | huggingface.co/imageomics/bioclip-2 |
| GPT-5.4 mini (OpenAI) | developers.openai.com/api/docs/models/gpt-5.4-mini |
| Gemini 3 Flash Preview (Google) | ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview |
| Claude Sonnet 4.6 (Anthropic) | anthropic.com/claude/sonnet |

**Proprietary models.** For each model, we present an image together with a natural-language query asking whether the specified biotic interaction is depicted and constrain the model to answer "Yes" or "No". The same image is evaluated repeatedly across the different query perturbations, while only the textual formulation is changed. For semantically equivalent variants, the expected answer remains unchanged, whereas for contradictory variants (e.g., reversing the interaction direction or removing the relation), the expected answer changes accordingly. We use the same prompt over all

image/query/model variants: *Does this image show {some query}? Answer with "Yes" or "No" and nothing else*. We evaluate performance using accuracy and quantify sensitivity to query formulation using the flip rate, which measures how often a model's prediction changes across query variants. This controlled setup isolates the effect of linguistic and semantic perturbations while keeping the visual input fixed.

### C.2 METRICS

**Average Precision at k=50 (AP@50).** We measure retrieval quality within the top 50 ranked images, accounting for both whether relevant images are retrieved and how highly they are ranked. Higher values indicate better retrieval performance.

**Mean Reciprocal Rank (MRR).** We measure the rank at which the first relevant image is retrieved, assigning higher scores when a relevant image appears earlier in the ranking.

**Recall at k=10 (Recall@10).** We measure the proportion of relevant images retrieved within the top 10 ranked results. Higher values indicate that more relevant images are retrieved among the highest-ranked results.

**Flip rate.** Let $p_i^q$ denote the model output for image $i$ under query formulation $q \in Q$, where $Q$ contains the similar outcome query variants (i.e., where the outcome is either "Yes" for all $q \in Q$ or "No" for all $q \in Q$) and $p_i$ the ground-truth. For each image $i$, we define a flip as:

$$F_i = \mathbf{1}\left[\exists q \in Q : p_i \neq p_i^q\right]$$

where $F_i = 1$ if at least one query variant changes the model's prediction relative to the active query, and $F_i = 0$ otherwise. The flip rate across $N$ images is then:

$$\text{FlipRate} = \frac{1}{N}\sum_{i=1}^{N} F_i$$

Thus, the flip rate measures the proportion of images for which at least one query reformulation changes the model prediction relative to the original active formulation. Lower values indicate greater consistency across query variants. For example, as shown in Table 5 for equivalent queries, Claude Sonnet 4.6's flip rate of 0.44 indicates that although the model has a relatively high accuracy, it is sensitive to query formulation (i.e., at least one query reformulation changes the model's prediction relative to the original query).

## D VISUAL VERIFICATION

To estimate image-level annotation noise, we randomly sampled 10 images from each of the 281 unique interactions in BIOINTERACT100 (i.e., 2,810 images in total). We identified 39 images (1.4%) in which the annotated interaction was not visually observable. Most cases resulted from multi-image iNaturalist observations, where some photographs depict only the source or target organism rather than the interaction itself. In a few cases, the interaction could not be reliably distinguished because it was too distant or lacked sufficient visual detail (see examples: inaturalist.org/observations/183678987, inaturalist.org/observations/272273734, inaturalist.org/observations/130839856). This analysis suggests that the level of image-level annotation noise is relatively low. Nevertheless, a comprehensive manual review of all images could further reduce the remaining annotation errors and improve overall image-level quality. Such an undertaking would require substantial effort, as BIOINTERACT comprises 109,668 observations containing 256,758 images. It is also important to note that our dataset is constructed from iNaturalist research-grade observations, following the common practice adopted by numerous biodiversity vision datasets (Yang et al., 2024; Manolache et al., 2026). While research-grade observations provide a high-quality foundation, they are not guaranteed to be free of image- or observation-level errors (iNaturalist, 2024). Consequently, a small residual level of annotation noise is an expected characteristic of datasets derived from large-scale community science repositories rather than a limitation unique to our dataset.

# E RESULTS & DISCUSSION

BioCLIP2 performs best overall, but all three models show lower AP@50 under passive formulations, while Recall@10 remains relatively stable (see Table 11 and Figure 9). This suggests that relevant images are still retrieved but ranked less effectively when interaction direction is reformulated, highlighting limitations in robust relational understanding. As indicated in our limitations, we will further investigate this behavior using fine-tuned models to assess whether targeted training on interaction semantics and direction improves robustness. More broadly, this work aims to highlight current limitations of state-of-the-art models and motivate benchmarks that support the systematic generation of new, grounded query variants, enabling continued evaluation beyond fixed query sets.

Gemini 3 Flash Preview's low text-only accuracy reflects a strong tendency to answer "Yes" regardless of query formulation, despite using the same prompt and protocol across models (see Table 5). This may reflect model-specific training or task interpretation. Conversely, GPT-5.2 mini performs near-perfectly text-only, consistently recognizing the absence of visual input, but drops substantially with images, suggesting that visual evidence introduces uncertainty or conflicts with its linguistic/ecological priors. This pattern extends across all three models, highlighting limitations in fine-grained visual grounding. Claude Sonnet 4.6 nevertheless performs best overall in the image-text setting, although its substantial flip rates indicate continued sensitivity to query formulation.

To assess whether models explicitly designed to improve compositional sensitivity also capture the directionality of biotic interactions, we evaluate compositionality-tuned retrieval encoders under source-target reversal and compare their behavior with proprietary multimodal language models (Table 10). The compositionality-tuned (i.e., NegCLIP (Yuksekgonul et al., 2023), SVLC (Doveh et al., 2023)) encoders remain sensitive to source-target reversal despite being explicitly designed to improve compositional understanding. This suggests that improved general compositional sensitivity does not necessarily translate into robust understanding of directional biotic relations.

Our benchmark BIOINTERACT100 is heavily represented by insect–plant interactions, where the source is often the active insect and the target is a plant, and, although images are typically close-ups of the interaction, the insect often occupies substantially fewer pixels than the flower or plant, making the target visually more prominent. The higher performance for active compared with passive queries may reflect the dominant representation (or canonical representation) of biotic interactions, where the source is the active organism and the target receives the interaction (e.g., "a bee visits a flower" versus "a flower is visited by a bee") encountered potentially during pre-training (we note that these specialized models are evaluated using their existing fine-tuned checkpoints; we do not fine-tune them ourselves). However, the differences are relatively small, so we interpret this pattern cautiously. Similarly, the stronger performance of target-only queries may reflect the visual prominence and contextual signal of plants and flowers, suggesting that models may rely on target/environmental cues rather than fully grounding the interaction between both organisms. Future work could test these hypotheses by controlling source/target visual prominence, evaluating more balanced interaction types and taxonomic groups, and fine-tuning models on controlled active/passive formulations to disentangle visual bias from linguistic and pre-training effects.

Table 10: **Performance under directional query perturbations.** We report accuracy and flip scores over **active** queries $(s, r, t)$ and **passive** queries $(t, r', s)$ from BIOINTERACT100. **Bold** and underlined entries indicate the **best** and the second best results, respectively.

| Model | Active (Acc) | Passive (Acc) | Flip | Active-reverse (Acc) | Passive-reverse (Acc) | Flip |
|---|---|---|---|---|---|---|
| *Compositionality-tuned retrieval encoders* | | | | | | |
| NegCLIP | 0.214 | 0.237 | 0.69 | 0.106 | 0.118 | 0.91 |
| SVLC | 0.320 | 0.326 | 0.55 | 0.160 | 0.163 | 0.93 |
| *Proprietary multimodal language models* | | | | | | |
| GPT 5.2 mini | **0.974** | **0.948** | 0.02 | **0.923** | 0.948 | **0.07** |
| Gemini 3 Flash Preview | 0.769 | 0.769 | **0.00** | 0.769 | **0.974** | 0.20 |
| Claude Sonnet 4.6 | 0.769 | 0.820 | 0.05 | 0.871 | 0.970 | 0.10 |

Table 11: **Performance on natural language understanding of biotic interaction queries across BIOINTERACT100 on (top) scientific terminology and (bottom) vernacular terminology queries. Bold** and underlined entries indicate the **best** and the second best results, respectively.

| Model | Method | Params (M) | mAP@50 | mMRR | mRecall@10 |
|---|---|---|---|---|---|
| CLIP | ViT-L-14 | 427 | 0.19 ± 0.03 | 0.40 ± 0.04 | 0.59 ± 0.05 |
| MetaCLIP | ViT-L-14-quickgelu | 427 | 0.29 ± 0.04 | 0.51 ± 0.05 | 0.68 ± 0.06 |
| SigLIP | SO400M-14-SigLIP | 877 | 0.34 ± 0.06 | 0.57 ± 0.06 | 0.76 ± 0.05 |
| SigLIP2 | ViT-L-16-SigLIP2-256 | 881 | 0.35 ± 0.07 | 0.59 ± 0.08 | 0.78 ± 0.07 |
| BioTrove-BioCLIP | ViT-B-16 | 149 | 0.19 ± 0.07 | 0.38 ± 0.12 | 0.58 ± 0.14 |
| BioCAP | ViT-B-16 | 149 | 0.45 ± 0.07 | **0.72 ± 0.07** | **0.88 ± 0.03** |
| BioCLIP | ViT-B/16 | 149 | 0.16 ± 0.07 | 0.41 ± 0.12 | 0.60 ± 0.14 |
| BioCLIP2 | ViT-L/14 | 427 | **0.47 ± 0.06** | 0.70 ± 0.06 | 0.85 ± 0.07 |
| **Model** | **Method** | **Params (M)** | **mAP@50** | **mMRR** | **mRecall@10** |
| CLIP | ViT-L-14 | 427 | 0.07 ± 0.02 | 0.24 ± 0.06 | 0.42 ± 0.09 |
| MetaCLIP | ViT-L-14-quickgelu | 427 | 0.14 ± 0.05 | 0.36 ± 0.11 | 0.60 ± 0.13 |
| SigLIP | SO400M-14-SigLIP | 877 | 0.21 ± 0.10 | 0.46 ± 0.14 | 0.67 ± 0.15 |
| SigLIP2 | ViT-L-16-SigLIP2-256 | 881 | 0.23 ± 0.09 | 0.48 ± 0.13 | 0.71 ± 0.13 |
| BioTrove-BioCLIP | ViT-B-16 | 149 | 0.20 ± 0.06 | 0.46 ± 0.09 | 0.69 ± 0.12 |
| BioCAP | ViT-B-16 | 149 | 0.35 ± 0.10 | 0.67 ± 0.14 | 0.87 ± 0.13 |
| BioCLIP | ViT-B/16 | 149 | 0.21 ± 0.06 | 0.53 ± 0.09 | 0.75 ± 0.12 |
| BioCLIP2 | ViT-L/14 | 427 | **0.43 ± 0.12** | **0.70 ± 0.14** | **0.88 ± 0.14** |

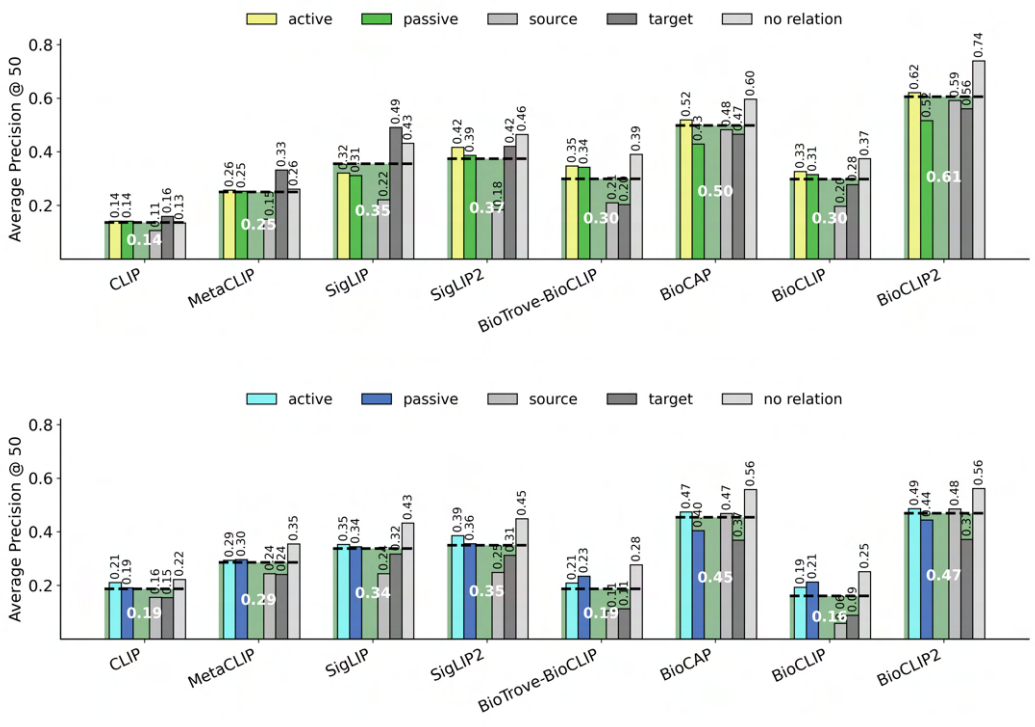


Figure 9: **Retrieval performance for each semantic variation query for (top) scientific terminology (bottom) vernacular terminology.** Queries are based on interaction triples $(s, r, t)$: **active** queries specify the full triple $(s, r, t)$, **passive** queries invert the interaction $(t, r', s)$, **source** queries condition on $(s, r)$, **target** queries on $(t, r')$, and **no relation** queries specify only entity pairs without the relation $(s, t)$.