BIOINTERACT: A Large-Scale Multimodal Dataset for Evaluating Fine-Grained Semantic Understanding of Biotic Interactions

Anonymous Author(s)
Affiliation
Address
email

Abstract

While recent advances in vision–language models (VLMs) have spurred the development of domain-specific datasets and benchmarks, these 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 scientific 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 adversarial 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. Data and code are available at the project website.

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 [34]. 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 [6, 35, 61, 65]. However, most domain-specific benchmarks are treated as monolithic performance targets, their design often emphasizing answer correctness, while fine-grained semantic understanding is typically assessed on datasets centered around common objects and everyday scenes. This disconnect allows models to achieve high performance without robust visual grounding, thereby providing an incomplete picture of multimodal competence. Addressing this evaluation gap is essential for ensuring reliable deployments of such systems in specialized fields [22].

In ecology, advances in vision and machine learning have accelerated biodiversity knowledge through targeted visual tasks ranging from basic visual interpretation (e.g., taxonomic classification, attribute prediction, trait inference) to higher-level ecological context understanding (e.g., disease detection)

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

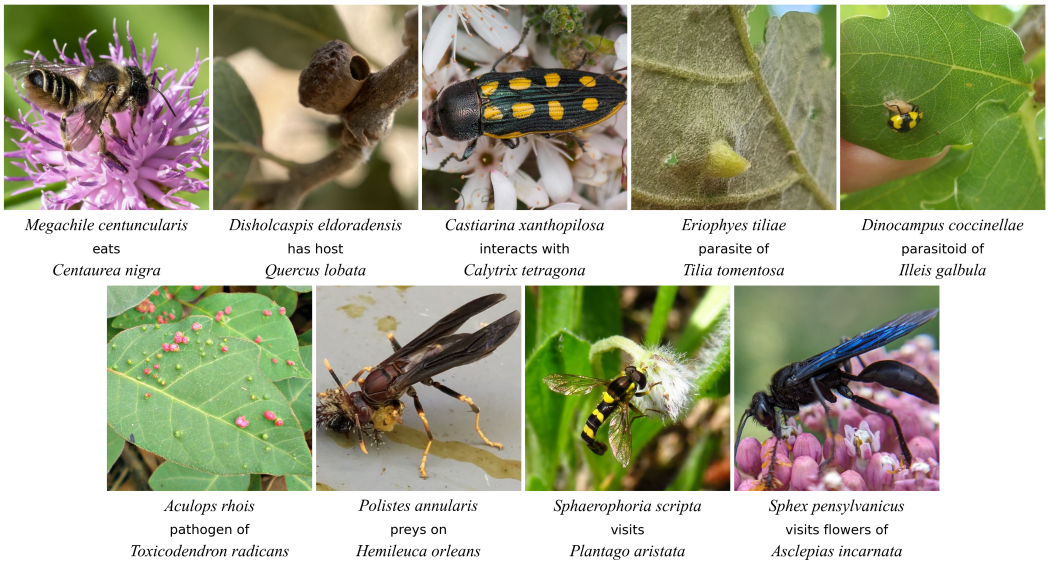

Figure 1: Example of images depicting biotic interactions in the BIOINTERACT dataset. Biotic interactions are annotated triplets—source taxon, interaction type, target taxon (e.g., *Megachile centuncularis* eats *Centaurea nigra*). Interactions are standardized to nine types defined by the Open Biological and Biomedical Ontologies (OBO) Relations Ontology (RO) [39] (see Table 2). Images originate from *research-graded* iNaturalist [28] observation records, where the source taxon occurrence is indexed (i.e., taxonomic identity, location, and time), while interaction type and target taxon are optionally annotated.

[31, 36, 50, 58]. 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 [28] and Observation.org [40] now host hundreds of millions of geo-referenced observations paired with images. Although these datasets are primarily used to document species occurrences, the images are captured *in situ* and frequently contain substantially richer ecological information as a by-product [44], including morphological traits, behavior, habitat, and interactions. For instance, pollinator images may reveal host flower preferences [45], while images of predatory species can provide evidence of trophic interactions [25]. 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 [55]. 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 [1], 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 [15, 23, 37, 54]. 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, as 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, typically evaluating answer accuracy using a single query per image in VQA such as CoralVQA [23] and AgMMU [15], or in text-to-image retrieval such as INQUIRE [54]. 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

Table 1: Current 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 4, enabling adversarial and stress-based evaluation. Studies showcasing the potential of using images to study biotic interactions for advancing ecological research, outside of the vision and machine learning community, are listed in Appendix B.

Tasks	Benchmark	# Images	Query	Description
visual question answering (VQA)	CoralVQA [23]	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[15]	50.2K	single	agriculture knowledge multiple-choice and open-ended questions, including insect/pest identification
image retrieval	INQUIRE [54]	33K	single	200 unique queries extracted from expert interviews and academic literature, including 12 parasitism and symbiosis relationships
multi-task	BIOINTERACT	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

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 ecological 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. Specifically, BIOINTERACT captures who interacts with whom, how, and under what context, by grounding interactions in structured triplets—source taxa, interaction type, target taxa. Examples from BIOINTERACT are shown in Figure 1, while Table 1 summarizes how BIOINTERACT contrasts with existing datasets 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; (2) rich annotations enabling controlled natural language descriptions for fine-grained semantic understanding; (3) BIOINTERACT100, a novel image retrieval task on a fixed set, comprising 281 unique interactions with 100 images; (4) 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

Vision-language models (VLMs) have emerged as a powerful paradigm for learning joint representations across visual and textual modalities. Early contrastive approaches such as CLIP [47] learn a shared embedding space where data from two modalities can be encoded jointly. Building on this foundation, VLMs, including GPT-4o [26], BLIP [26], LLaVA[33], and Qwen2-VL [56], 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 [6, 61, 65]. For instance, GMAI-MMBench [6] and BenchX [65] feature a collection of healthcare and medically relevant questions, while MMMU-Pro [61], comprises visual-questions related to problems from diverse fields such as business, arts, and science. However, the evaluation landscape remains fragmented: domain-specific benchmarks are treated as monolithic performance targets, while fine-grained semantic understanding is typically assessed only in general-domain datasets depicting common objects and everyday scenes (e.g., MS-COCO) [2, 52, 59, 62]. Nearly all domain-specific benchmarks are *static*, with performance gains increasingly reflecting memorization rather than capability in pursuit of state-of-the-art performance [7]. It is crucial to develop evaluation protocols that reflect true capabilities, otherwise we run the risk of deploying models that perform

Table 2: Interaction types in the BIOINTERACT dataset, defined by Open Biological and Biomedical Ontologies (OBO) Relations Ontology (RO) [39]. 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	-
visits flowers of	-

Table 3: Key statistics of the BIOINTERACT dataset. 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

sub-optimally in real-life applications, potentially leading to incorrect decisions, user dissatisfaction, or even serious consequences in high-stakes applications such as healthcare [22].

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 [31, 36, 37, 50, 58], these efforts largely focus on individual organisms or ecological context understanding . In contrast, BIOINTERACT targets biotic interactions, capturing relationships between organisms (e.g., predation, symbiosis, competition), 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 [15, 23] (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 [54] (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 is unambiguous and logic is formal, biotic interactions are particularly challenging, as subtle semantic differences can fundamentally alter their meaning, yielding naturally adversarial interaction instances. To address this, we enable fine-grained semantic evaluation through the systematic generation of semantically equivariant statements from interaction triplets [10]. This approach provides precise control over semantically similar and dissimilar queries, without relying on human or prompt-based biases [9, 14, 32, 48].

3 BIOINTERACT

3.1 Data curation

We retrieve biotic interaction data from GloBI [10], using the archive hosted on Zenodo (version 0.8) [11]. GloBI periodically collects biotic interactions from a wide range of independent datasets, such as iNaturalist [28] and Encyclopedia of Life [13], 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

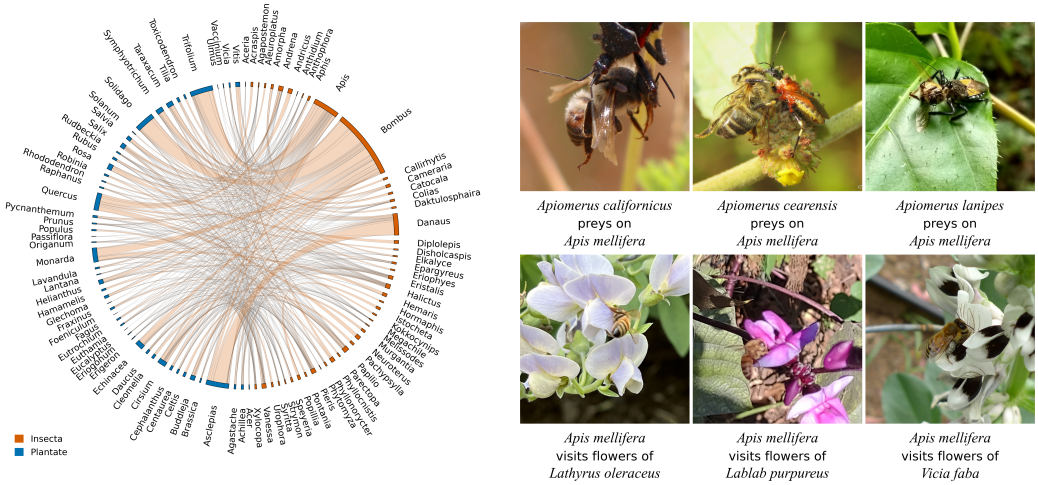

Figure 2: (left) BIOINTERACT top 100 most frequent taxa and their interactions. The dataset is centered around *Insecta-Plantae* interactions. (right) Examples illustrating ecological networks in BIOINTERACT. Top row showcases predatory interactions of the honeybee (*Apis mellifera*) with rove beetles (genus *Apiomerus*) [25]; while bottom row showcases its plant preferences, including associations with families such as *Fabaceae* [4].

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* [27] iNaturalist observation, yielding 1.27M records. To these records we aggregate corresponding iNaturalist observation data from GBIF [17], including taxonomic identity, spatiotemporal information (location, and time) and corresponding images.

Interacting taxa are recorded at heterogeneous taxonomic levels. We align all records to a standardized seven-level Linnaean taxonomy (*kingdom*, *phylum*, *class*, *order*, *family*, *genus*, *species*) using the GBIF Backbone Taxonomy [16] (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 retains 636K records, corresponding to 215K unique biotic interactions. We further enrich the dataset with corresponding English vernacular names from GBIF Backbone Taxonomy [16].

Biotic interaction with image data are documented *a priori* via iNaturalist; however, these annotations are not routinely verified as the *research-grade* data [27]. Consequently, we do not treat iNaturalist as a primary source of valid biotic interactions. Instead, we retain only interactions that are independently supported by at least one additional resource (a separate curated dataset, other than iNaturalist) indexed in the GloBI database [10]. This cross-validation procedure ensures high-confidence annotations and yields 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 OBO RO [39] (see interaction type definitions in 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 dataset key statistics in Table 3). Based on the most recent IUCN Red List assessment, BIOINTERACT covers 87 threatened species [30] and 23 of the world's worst invasive species [18], enabling evaluation on ecologically critical and conservation-relevant taxa (see Appendix A) [20]. 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 [38].

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

Table 4: BIOINTERACT annotations enumerate 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 OBO RO [39].
Source/target rank	The taxonomic rank at which the interacting taxa is identified (*kingdom*, *phylum*, *class*, *order*, *family*, *genus*).
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`.

report strong declines worldwide in both insects and plants, threatening plant diversity, food security and ecosystem stability [42]. Interaction types are hierarchically organized, in which the 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 [29]. 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 4). Controlled perturbation is preferred to ensure semantic equivalence and avoid hallucinations or unintended biases that may arise from generative approaches [57].

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 by with another entity; (3) *relation removal*: the relation is removed, while 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

Table 5: Performance on fine-grained semantic understanding of biotic interaction queries for different models. We report the mean Average Precision at k (mAP@k), mean Reciprocal Rank (mMRR), and mean Recall at k (mRecall@10) computed over semantically equivalent query variants. **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 ± 0.02	0.34 ± 0.06	0.56 ± 0.05
MetaCLIP	ViT-L-14-quickgelu	427	0.25 ± 0.06	0.49 ± 0.10	0.72 ± 0.11
SigLIP	SO400M-14-SigLIP	877	0.35 ± 0.10	0.60 ± 0.11	0.78 ± 0.08
SigLIP2	ViT-L-16-SigLIP2-256	881	0.37 ± 0.10	0.63 ± 0.10	0.83 ± 0.11
BioTrove-BioCLIP	ViT-B-16	149	0.30 ± 0.08	0.57 ± 0.11	0.78 ± 0.09
BioCAP	ViT-B-16	149	0.50 ± 0.06	0.78 ± 0.04	0.93 ± 0.03
BioCLIP	ViT-B/16	149	0.30 ± 0.06	0.63 ± 0.07	0.82 ± 0.07
BioCLIP2	ViT-L/14	427	$\mathbf{0.61 \pm 0.08}$	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.95 \pm 0.02}$

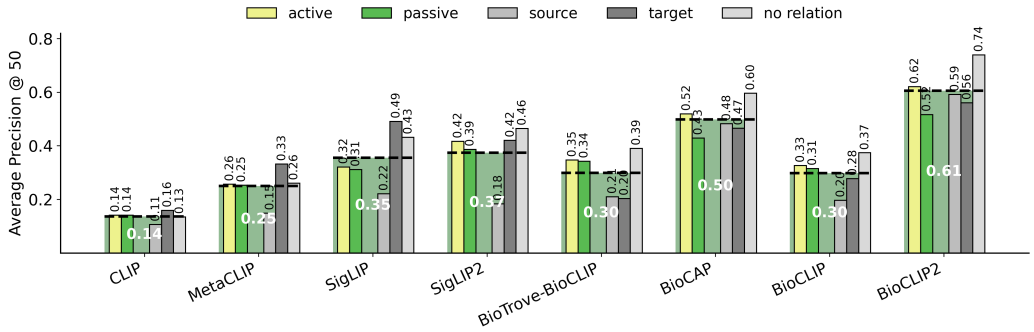

Figure 3: Performance per query reported under AP. Queries are based on interaction triples (s, r, t): **active** queries specify the full triple (s, r, t), **passive** queries invert the interaction (t, r', s), **source** queries condition on (s, r), **target** queries on (t, r'), and **no relation** queries specify only entity pairs without the relation (s, t).

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.

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 Evaluations

Our evaluation considers how predictions change under semantic perturbations, while the visual input remains constant. We introduce a ranking task similar to INQUIRE [54], which fixes images for each query and uses models like GPT-4o [26] to improve over initial text-to-image CLIP-style retrieval. We, thus, measure fine-grained semantic understanding in embedding similarity and binary (yes/no) questions.

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 sample per image in the first pass, and subsequently filling remaining slots if necessary. This procedure yields BIOINTERACT100, a

Table 6: Performance on different general-purpose 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.

	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**

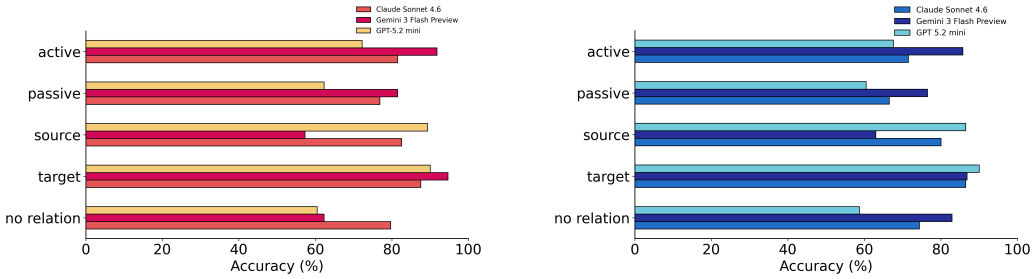

Figure 4: Performance on natural language understanding of biotic interaction queries grouped by language specificity: (left) scientific terminology; (right) vernacular names. For each query type we report accuracy.

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 C lists all query variants for each of the 281 unique interactions.

Models. We evaluate CLIP-style models for text-to-image retrieval, such as OpenAI [47], MetaCLIP [8], SigLIP [63], SigLIP2 [53], as well as specialized models, such as BioTrove-BioCLIP [60], BioCAP [64], BioCLIP [51] and BioCLIP2 [21]. We adopt proprietary multimodal language models (MLLMs) such as GPT-5.4 mini [41], Claude Sonnet 4.6 [3], and Gemini 3 Flash Preview [19] 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*. All proprietary models were evaluated via their APIs with deterministic decoding (`temperature` $= 0.0$).

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 5 and Figure 3 (see also extensive results in Appendix D). 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 [51], they underperform comparatively to larger general models like SigLIP and SigLIP2. BioCLIP2, the largest specialized model (trained on TreeOfLife-200M [21]) 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 7: Performance on adversarial 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). **Bold** entries indicate the **best** results.

	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 about other relevant organisms.

High accuracy on paraphrasing does not imply consistency. We report performance of general-purpose proprietary models in Table 6 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 7. 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. 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 BIOINTERACT100, an image retrieval task on a fixed set, underscore 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.

References

[1] 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:e98899, 2025.

[2] 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*, pages 29612–29622, 2025.

[3] Anthropic. Claude sonnet 4.6. `hhttps://www.anthropic.com/claude/sonnet`, 2026. Accessed 2026-05-06.

[4] 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):191–201, 2026.

[5] Florencia Baudino, Ramiro R Ripa, Jorgelina Franzese, and Victoria Werenkraut. Using citizen science as a research prioritization tool to detect co-occurrences of the invasive species Harmonia axyridis. *Insect Conservation and Diversity*, 2025.

[6] 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*, 37:94327–94427, 2024.

[7] Zerui Cheng, Stella Wohnig, Ruchika Gupta, Samiul Alam, Tassallah Abdullahi, João Alves Ribeiro, Christian Nielsen-Garcia, Saif Mir, Siran Li, Jason Orender, et al. Benchmarking is broken–don't let ai be its own judge. *arXiv preprint arXiv:2510.07575*, 2025.

[8] Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

[9] 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)*, pages 2706–2738, 2026.

[10] GloBI Community. Global Biotic Interactions: Interpreted Data, January 2025. URL `https://doi.org/10.5281/zenodo.14640564`. Accessed via inaturalist.org on 2025-12-13.

[11] GloBI Community. Global Biotic Interactions: Interpreted Data Products hash://md5/e76bf914309ad27dce6ab911d8854590 hash://sha256/ba79836caab5b7ba2d7d659123d27c89f4ad990bd5 0f97ded935edee9fbe9f87 , January 2025. URL `https://doi.org/10.5281/zenodo.14640564`.

[12] Jean-François Doherty, Antoine Filion, Jerusha Bennett, Upendra Raj Bhattarai, Xuhong Chai, Daniela de Angeli Dutra, Erica Donlon, Fátima Jorge, Marin Milotic, Eunji Park, et al. The people vs science: can passively crowdsourced internet data shed light on host–parasite interactions? *Parasitology*, 148(11):1313–1319, 2021.

[13] Encyclopedia of Life. Encyclopedia of life (eol), 2025. URL `https://eol.org`. Accessed: 2026-03-25.

[14] Federico Felizzi, Olivia Riccomi, Michele Ferramola, Francesco Andrea Causio, Manuel Del Medico, Vittorio De Vita, Lorenzo De Mori, Alessandra Piscitelli, Pietro Eric Risuleo, Bianca Destro Castaniti, et al. Are large vision language models truly grounded in medical images? evidence from italian clinical visual question answering. *arXiv preprint arXiv:2511.19220*, 2025.

[15] Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram S Adve, and Yu-Xiong Wang. Agmmu: A comprehensive agricultural multimodal understanding benchmark. *arXiv preprint arXiv:2504.10568*, 2025.

[16] GBIF Secretariat. GBIF Backbone Taxonomy, 2023. URL https://doi.org/10.15468/39omei. Accessed via GBIF.org on 2025-12-30.

[17] GBIF.org. Global biodiversity information facility. https://www.gbif.org, 2025. Accessed on on 2025-12-25.

[18] Global Invasive Species Database. 100 of the world's worst invasive alien species. http://www.iucngisd.org/gisd/100_worst.php, 2026. Accessed: 26-03-2026.

[19] Google. Gemini 3 flash preview. https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview, 2025. Gemini API model documentation, accessed 2026-05-06.

[20] 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):1781–1789, 2021. doi: https://doi.org/10.1111/ecog.05790. URL https://nsojournals.onlinelibrary.wiley.com/doi/abs/10.1111/ecog.05790.

[21] Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E White, James Balhoff, et al. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. *arXiv preprint arXiv:2505.23883*, 2025.

[22] 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.

[23] Hongyong Han, Wei Wang, Gaowei Zhang, Mingjie Li, and Yi Wang. Coralvqa: A large-scale visual question answering dataset for coral reef image understanding. *arXiv preprint arXiv:2507.10449*, 2025.

[24] Madison Hernandez, Paul Masonick, and Christiane Weirauch. Crowdsourced online images provide insights into predator-prey interactions of putative natural enemies. *Food Webs*, 21: e00126, 2019.

[25] 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:e00399, 2025.

[26] 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.

[27] 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: 2026-01-02.

[28] iNaturalist. https://www.inaturalist.org/, 2025. Accessed via inaturalist.org on 2025-12-13.

[29] iNaturalist. What is geoprivacy? what does it mean for an observation to be obscured? https://help.inaturalist.org/en/support/solutions/articles/151000169938-what-is-geoprivacy-what-does-it-mean-for-an-observation-to-be-obscured-, 2025. iNaturalist Help. Modified on March 12, 2025. Accessed April 6, 2026.

[30] IUCN. The iucn red list of threatened species, 2025. URL https://www.iucnredlist.org. Downloaded on 2025-04-07; accessed via GBIF.org on 2026-03-25.

[31] 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*, pages 20496–20506, 2023.

[32] 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)*, pages 62–76. Springer, 2025.

[33] 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.

[34] Fan Liu, Jindong Han, Tengfei Lyu, Weijia Zhang, Zhe-Rui Yang, Lu Dai, Cancheng Liu, and Hao Liu. Foundation models for scientific discovery: From paradigm enhancement to paradigm transition. *arXiv preprint arXiv:2510.15280*, 2025.

[35] 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. *arXiv preprint arXiv:2508.01055*, 2025.

[36] Georgiana Manolache, Gerard Schouten, and Joaquin Vanschoren. Crypticbio: A large multimodal dataset for visually confusing species. *arXiv preprint arXiv:2505.14707*, 2025.

[37] 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.

[38] 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.

[39] OBO Relations Ontology Consortium. Relations ontology (ro). `https://obofoundry.org/ontology/ro.html`, 2024. Accessed on on 2025-12-15.

[40] Observation International and local partners. Observation.org dataset. `https://observation.org`, 2026. Accessed via observation.org on 2025-12-13.

[41] OpenAI. Gpt-5.4 mini. `https://developers.openai.com/api/docs/models/gpt-5.4-mini`, 2026. OpenAI API model documentation, accessed 2026-05-06.

[42] 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): 1323–1333, 2024.

[43] Connor T Panter and Arjun Amar. Sex and age differences in the diet of the eurasian sparrowhawk (accipiter nisus) using web-sourced photographs: exploring the feasibility of a new citizen science approach. *Ibis*, 163(3):928–947, 2021.

[44] 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):e12295, 2024.

[45] 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):e11537, 2024.

[46] Jorrit H Poelen, James D Simons, and Chris J Mungall. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological informatics*, 24: 148–159, 2014.

[47] 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*, pages 8748–8763. PmLR, 2021.

[48] Binesh Sadanandan and Vahid Behzadan. Psf-med: Measuring and explaining paraphrase sensitivity in medical vision language models. *arXiv preprint arXiv:2602.21428*, 2026.

[49] Manu E Saunders, Emma K Goodwin, Karen CBS Santos, Carolyn A Sonter, and Romina Rader. Cavity occupancy by wild honey bees: need for evidence of ecological impacts. *Frontiers in Ecology and the Environment*, 19(6):349–354, 2021.

[50] 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*, pages 249–253. 2020.

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

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[53] 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.

[54] 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.

[55] 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: A framework for interactive discovery in large-scale biodiversity databases. *arXiv preprint arXiv:2511.15656*, 2025.

[56] 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.

[57] Qiyao Wei, Edward Morrell, Lea Goetz, and Mihaela van der Schaar. Semantic-kg: Using knowledge graphs to construct benchmarks for measuring semantic similarity. *arXiv preprint arXiv:2511.19925*, 2025.

[58] 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):2251–2265, 2018.

[59] 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*, pages 1827–1836, 2024.

[60] Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37:102101–102120, 2024.

[61] 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)*, pages 15134–15186, 2025.

[62] 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? *arXiv preprint arXiv:2210.01936*, 2022.

[63] 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*, pages 11975–11986, 2023.

[64] Ziheng Zhang, Xinyue Ma, Arpita Chowdhury, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Samuel Stevens, Hilmar Lapp, Tanya Berger-Wolf, Yu Su, et al. Biocap: Exploiting synthetic captions beyond labels in biological foundation models. *arXiv preprint arXiv:2510.20095*, 2025.

[65] 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*, 37:6625–6647, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present BioInteract, the largest publicly available multimodal dataset comprising richly annotated images depicting interactions between organsims, or biotic interactions, providing a natural testbed for tasks involving images and unconstrained, free-form natural language, as interacting organisms are discerned from images alone and their relationship can be expressed through multiple linguistic forms.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss Limitations in Section 4.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification:We submit a dataset & evaluation paper.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All assets are reproducible via our curation pipeline available in our git page https://github.com/georgianagmanolache/biointeract.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All experimental results are reproducible, the data is available in https://huggingface.co/datasets/gmanolache/BioInteract; and script in https://github.com/georgianagmanolache/biointeract.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See section 4 and supplementary material D.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See section 4 and supplementary material D.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See section 4 and supplementary material D.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: See Section 3.2 License, Geoprivacy and Responsible use and supplementary material A.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 3 and supplementary material A.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: See Section 3.2 & supplementary material A.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Section 3.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code and dataset are accessible through the project website https://georgianagmanolache.github.io/biointeract.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

A BIOINTERACT Dataset

Figure 5: Sample interactions from BIOINTERACT depicting insects visiting plants.

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 [46]. Interactions are standardized to nine types defined by OBO RO [39] (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 [10], establishing them as reliable ground-truth annotations.

Spatio-temporal data plays a crucial role in biodiversity datasets because species distributions and appearances are strongly influenced by both geographic location and time [36]. Figure 9 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.

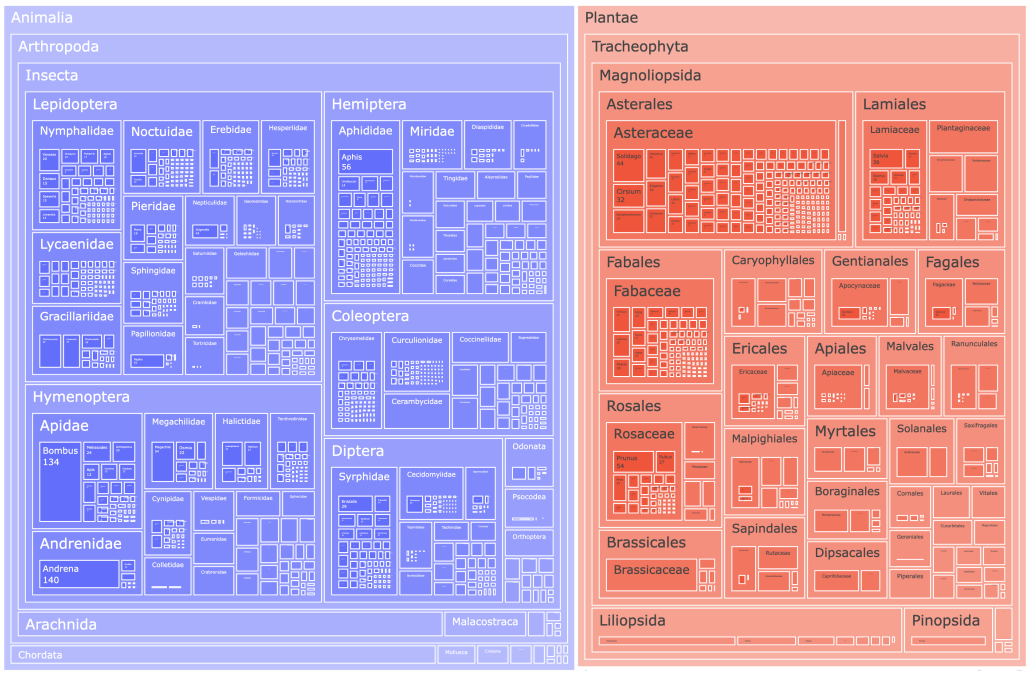

Figure 6: Treemap visualization of the species covered in BIOINTERACT.

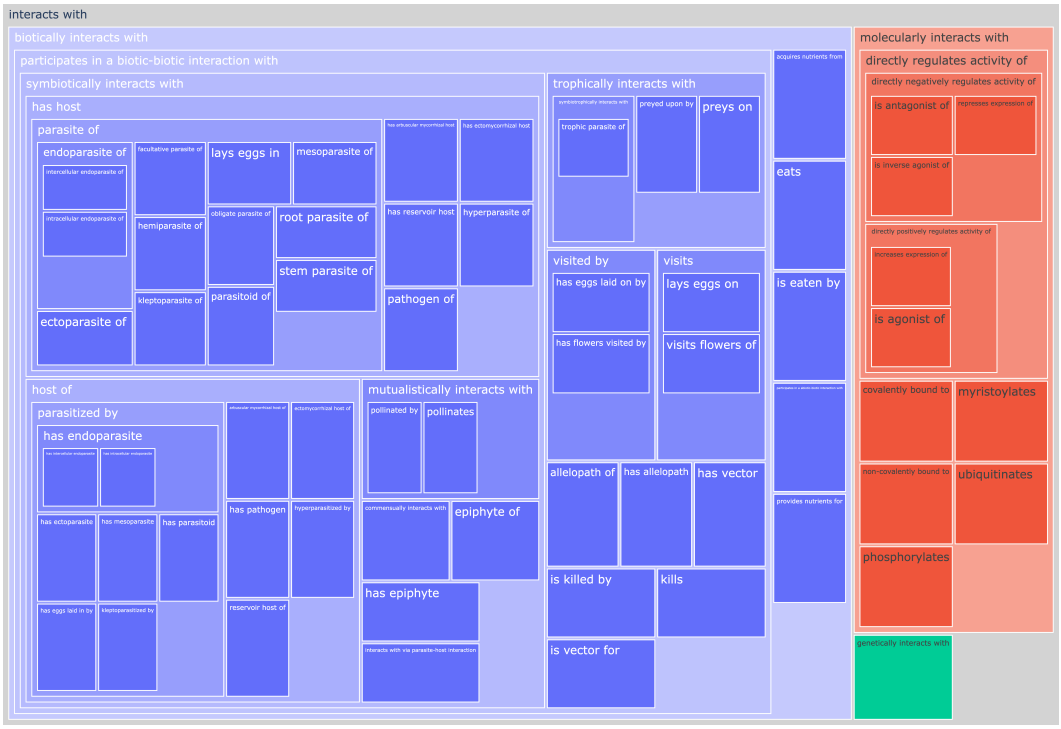

Figure 7: Treemap visualization of the interaction hierarchy used in OBO RO [39].

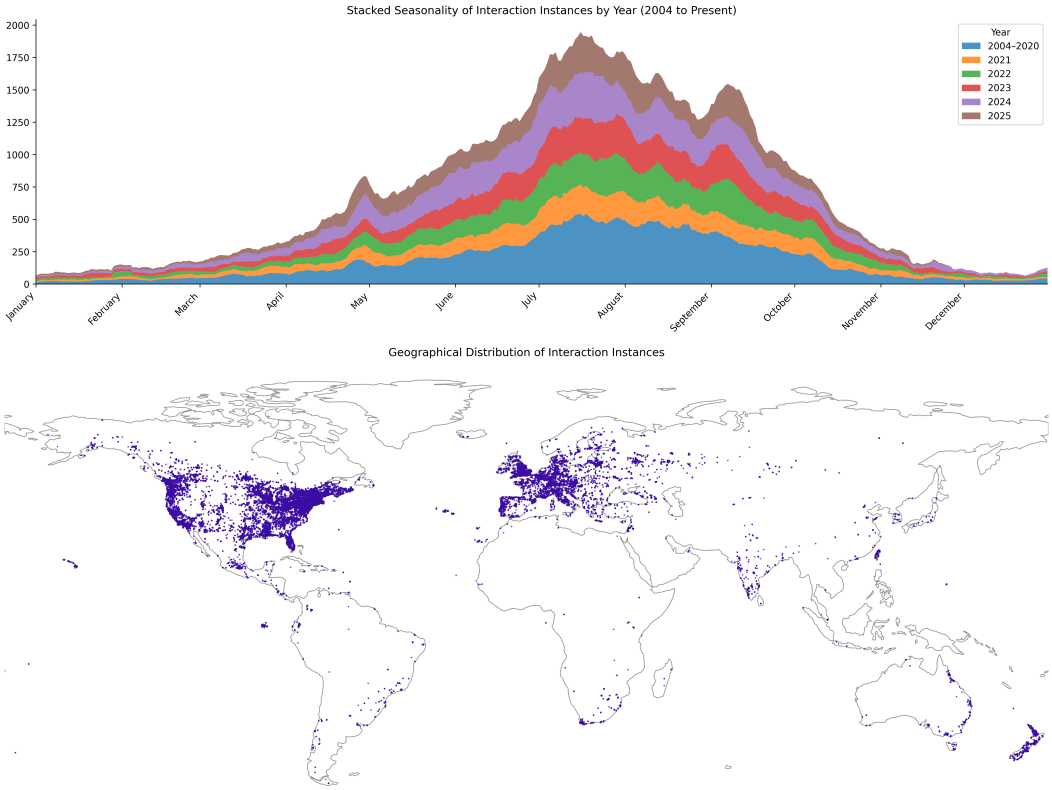

Figure 8: Spatiotemporal distribution of interaction instances in BIOINTERACT: (top) stacked seasonality; (bottom) geographical distribution. Majority of records are concentrated in Europe and North America, with a seasonal peak in observations during July.

B Related Work

A small body of work leverages online image repositories to manually extract biotic interaction data; however, these datasets are typically interaction- or taxon-specific and are not released as ML-ready datasets at scale (see Table 8).

Table 8: BIOINTERACT comparable datasets. Data is manually extracted from sources and is interaction- and taxon-specific.

Interaction Type	Description	# Images	Sources	Reference
Predation	Prey in relation to age and sex of sparrowhawk (*Accipiter nisus*).	843	iNaturalist, Google Images, Macaulay Library, BirdGuides, Facebook, Twitter	[43]
Predation	Prey spectrum of assassin bugs (*Hemiptera*: *Reduviidae*)	832	Flickr, iSpot Nature, BugGuide, NatureWatch, Google Images	[24]
Predation / parasitism	Detect co-occurrences and potential novel interactions between the invasive ladybird (*Harmonia axyridis*) and local biota (plants, arthropods and fungi).	783	iNaturalist, ArgentinNat, WhatsApp	[5]
Competition	Competition for nest cavities between bee species and wild honey bee colonies in native and introduced ranges.	326	iNaturalist	[49]
Flower visitation	Host flowers of the introduced wasp *Isodontia mexicana*.	250	iNaturalist	[45]
Host preferences	Host preferences and phenological peak of hairworms inferred from infected hosts.	20	iNaturalist	[12]

C Controlled diagnostic benchmarks

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), **active reversal** queries swap the source and target from active queries (t, r, s), and **passive reversal** queries swap the source and target from active passive (t, r, s).

Table 9: 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*

D Evaluation details

Table 10: Performance on natural language understanding of biotic interaction queries across BIOINTERACT100 on scientific terminology queries.

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	$\mathbf{0.72 \pm 0.07}$	$\mathbf{0.88 \pm 0.03}$
BioCLIP	ViT-B/16	149	0.16 ± 0.07	0.41 ± 0.12	0.60 ± 0.14
BioCLIP2	ViT-L/14	427	$\mathbf{0.47 \pm 0.06}$	0.70 ± 0.06	0.85 ± 0.07

Table 11: Performance on natural language understanding of biotic interaction queries across BIOINTERACT100 on vernacular terminology queries.

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	$\mathbf{0.43 \pm 0.12}$	$\mathbf{0.70 \pm 0.14}$	$\mathbf{0.88 \pm 0.14}$

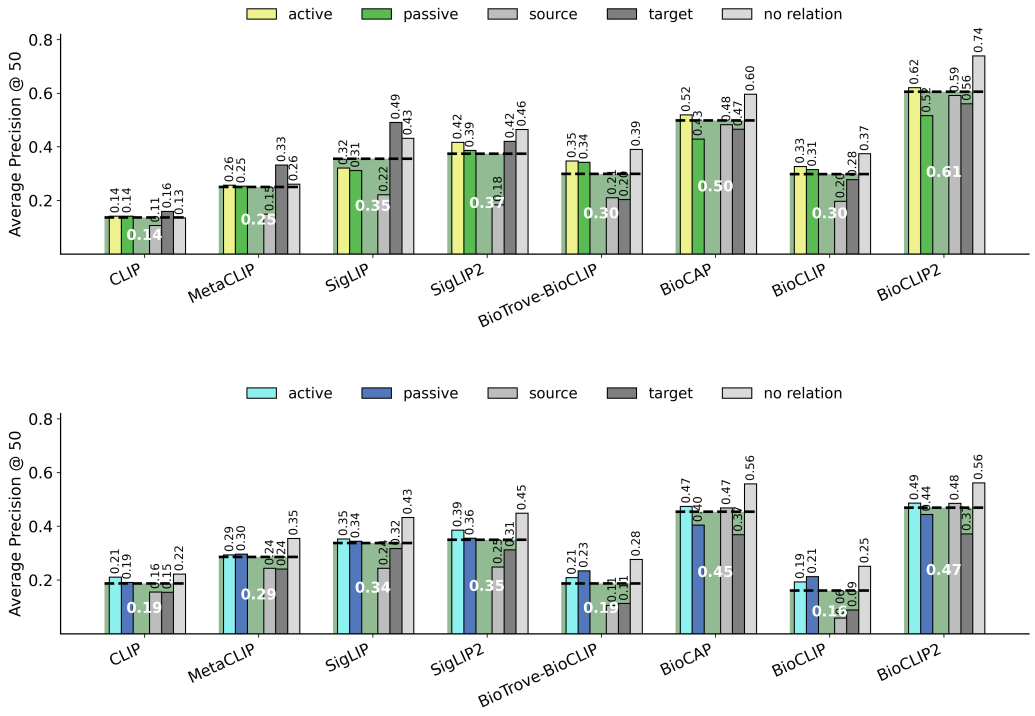

Figure 9: Retrieval performance for each semantic variation query for (top) scientific terminology (bottom) vernacular terminology. Queries are based on interaction triples (s, r, t): **active** queries specify the full triple (s, r, t), **passive** queries invert the interaction (s, r', t), **source**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