Global and Local Entailment Learning for Natural World Imagery

Srikumar Sastry Aayush Dhakal Eric Xing Subash Khanal Nathan Jacobs
Washington University in St. Louis
{s.sastry, a.dhakal, e.xing, k.subash, jacobsn}@wustl.edu

Abstract

Learning the hierarchical structure of data in vision-language models is a significant challenge. Previous works have attempted to address this challenge by employing entailment learning. However, these approaches fail to model the transitive nature of entailment explicitly, which establishes the relationship between order and semantics within a representation space. In this work, we introduce Radial Cross-Modal Embeddings (RCME), a framework that enables the explicit modeling of transitivity-enforced entailment. Our proposed framework optimizes for the partial order of concepts within vision-language models. By leveraging our framework, we develop a hierarchical vision-language foundation model capable of representing the hierarchy in the Tree of Life. Our experiments on hierarchical species classification and hierarchical retrieval tasks demonstrate the enhanced performance of our models compared to the existing state-of-the-art models. Our code and models are open-sourced at https://vishu26.github.io/RCME.

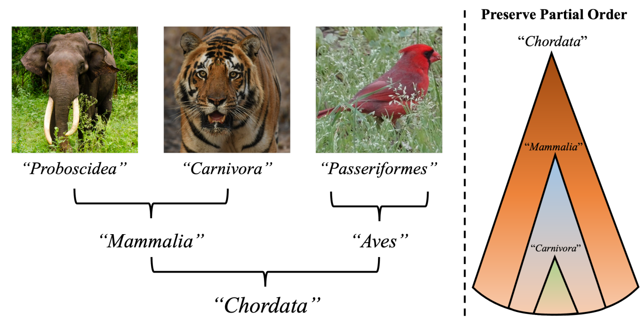


Figure 1. Conceptual overview of our method focusing on preserving the global order of concepts in vision-language models according to their distance from an entailment root. Our method aims to enforce transitivity in entailment.

1. Introduction

Computer Vision has become increasingly valuable in understanding the natural world, thanks to the rise of open citizen science platforms and the abundance of consumer data. These tools, complemented by domain experts, have been instrumental in addressing pressing challenges at scale such as automatic species identification [39, 42], animal behavior understanding [6, 25] and visual geolocalization [16, 43]. Nevertheless, the complex and ever-changing nature of our world poses a significant challenge in constructing models that can generalize and adapt to novel data.

BioCLIP [36] and BioTroveCLIP [45] successfully attempted to build a vision-language foundation model for the Tree of Life. Recently, TaxaBind [34] extended BioCLIP's capabilities to handle additional modalities such as audio and satellite imagery. However, these models fail to fully leverage the hierarchical nature of the label space. This limited capability of these models limits them to reason at the most granular level of the hierarchy (i.e. *species*) using a fixed database of taxonomic labels. Consequently, this restriction prevents the models from accurately representing the actual taxonomic system and the evolution of species in the Tree of Life.

We argue that learning hierarchical representations for the Tree of Life is crucial. A significant portion of species on Earth remain undescribed [29], and labeling specimens up to the species rank is expensive and requires adequate expertise for biologists [15, 32, 37]. Furthermore, the taxonomic classification system and labels are subject to change over time due to mislabeling or the discovery of new species [36]. Hierarchical representations can allow for reasoning about such species at any rank and can eventually be used for grouping and routing specimens to biologists with appropriate expertise [29]. It can also help understand the evolution of certain species in the Tree of Life. For end users with arbitrary expertise, hierarchical representations facilitate classification at any taxonomic rank. Finally, in the paper, we empirically demonstrate the benefits of such structured representations for classification and retrieval tasks.

A popular technique for learning hierarchical representations for vision-language models is entailment learning, which aims to learn concentric cones of embedding sub-regions. In the past, studies on entailment learning relied on explicitly defining aperture angles that defined the structure of these cones [11, 14, 28]. However, these approaches can be limiting since the optimality of the cone structures can vary from application to application. Recently, Alper *et al*. [1] introduced radial embeddings, an approach to fine-tune existing vision-language models that eliminate the dependence of the objective on entailment cones. However, their method fails to enforce the partial order of concepts in their hierarchical embedding space.

In Figure 1, we illustrate transitivity in entailment, imposing a partial ordering of concepts in the embedding space [14]. For instance, if "*Mammalia*" is entailed by "*Chordata*" and "*Carnivora*" is entailed by "*Mammalia*," then "*Chordata*" entails "*Carnivora*". Ideally, this phenomenon should hold for all possible sub-hierarchies in the data. Transitivity is an important property for a representation space as it controls the distance between concepts based on their semantic granularity. For instance, fine-grained concepts are projected farther from coarse-grained concepts.

To this end, we propose a novel framework called Radial Cross-Modal Embeddings (RCME) which enables the learning of hierarchical representations by imposing partial order constraints while eliminating the need to define the structure of the cones. Using our framework, we propose a hierarchical vision-language foundation model for the Tree of Life, outperforming existing state-of-the-art models in hierarchical classification tasks. Notably, our framework is general enough to be adapted for any other domain. Our contributions are as follows:

1. We propose an objective function to optimize for transitivity in textual entailment within vision-language models. We address the issue of Alper *et al*. [1], which overlooks partial order in textual entailment.
2. We propose Radial Cross-Modal Embeddings (RCME), a framework that solves for transitivity-enforced entailment and cross-modal alignment in vision-language models.
3. Experiments show our models outperform the state-of-the-art in hierarchical classification, hierarchical retrieval, and image-to-image retrieval tasks.

2. Related Works

2.1. Representation learning

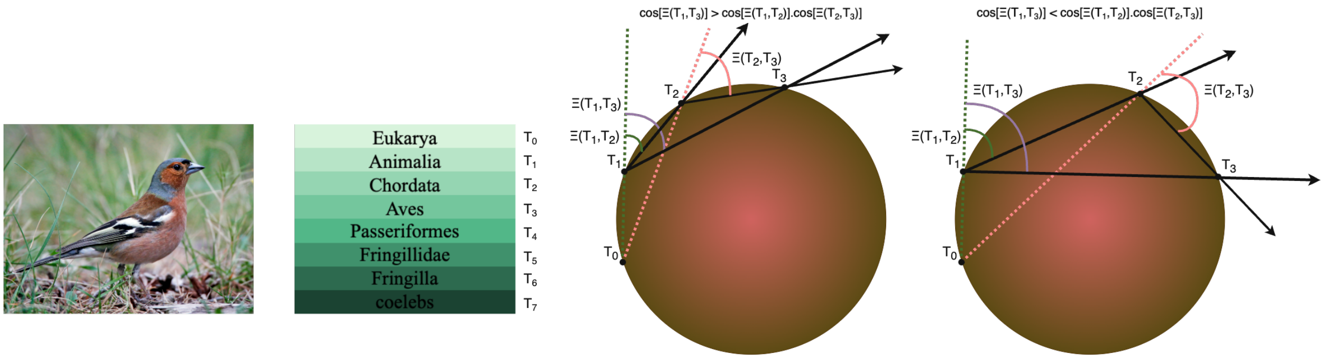
Contrastive learning has enabled training large-scale vision-language models [19, 21, 30] which have generated significant advancement in diverse tasks like image classification and few-shot learning. Recent lines of work have focused on improving these generalist models by either achieving fine-grained alignment [8, 22, 46] or enhancing intra-modal representations [9, 24]. However, such methods are still not ideal for specific domains where there are structures of representations imposed by semantics. Hierarchical representation learning has gained traction, particularly in tasks requiring structured knowledge representation, such as natural language inference (NLI) [12, 20, 38] and knowledge graph embeddings [2, 5]. One of the key challenges in hierarchical representation learning is preserving the partial ordering of concepts in the embedding space while maintaining generalizability across different domains.

2.2. Computer vision for ecology

The intersection of computer vision and ecology has led to significant advances in tasks like fine-grained species classification [15, 36], animal detection using camera traps [3, 35], and animal behavior recognition [6, 25]. Large-scale datasets [23, 36, 41, 45] and citizen science platforms like iNaturalist [39] have enabled the training of deep learning models to solve these tasks. Multimodal representation learning frameworks [7, 10, 17, 33] with wildlife observations and satellite images has shown benefits in solving ecological tasks like species distribution modeling. Recently, vision-language foundation models for the Tree of Life such as BioCLIP [36] and BioTroveCLIP [45] have shown excellent capabilities in zero-shot species identification. TaxaBind [34] extended such vision-language models by incorporating additional modalities such as audio and satellite imagery. However, all such models are limited to fixed taxonomic labels and struggle with classification at arbitrary taxonomic ranks. They lack structured hierarchical representations implied by the hierarchical nature of the Tree of Life.

2.3. Entailment learning

Traditional hierarchical learning approaches often rely on hyperbolic embeddings [4, 26, 27] to enforce hierarchical relationships. Entailment learning is particularly useful for structuring embeddings in a semantic order [40]. Early works on entailment learning explicitly defined cone structures using aperture angles to capture hierarchical dependencies [11, 14, 28, 44, 47]. However, these approaches are often restrictive as the optimal structure of the cones can vary across datasets and applications. Recent works, such as Radial Embeddings [1] and ATMG [31], attempt to relax these constraints by learning hierarchical embeddings without predefined cone structures in the radial and hyperbolic geometry respectively. While these methods improves adaptability, it does not explicitly enforce partial ordering, leading to suboptimal performance in hierarchical retrieval tasks.



(a) Image (b) Taxonomic Hierarchies (c) Transitivity Enforced Entailment (d) Transitivity Violation

Figure 2. **Transitivity in Entailment**. In an ideal transitivity-imposed entailment, textual embeddings satisfy partial order conditions.

3. Preliminaries

We begin with the arguments on the conditions for entailment as proposed in Esteva *et al*. [13] and Ganea *et al*. [14]. For our purposes, we consider the tree of life hierarchy and develop the logic for entailment with respect to it. Let $\mathcal{R}_{\{j=0,1...N\}}$ represent the sets of the domain of textual embeddings at each hierarchical rank. As j increases, the semantic granularity in textual embedding increases. We consider $\mathcal{R}_0$ as the set which contains the entailment root embedding, T_0. Let $T_j^i \in \mathcal{R}_j$, denote the textual embedding for the i^{th} species belonging to some rank j in the hierarchy. We use $T_{j-1}^i \in \mathcal{R}_{j-1}$ to denote the immediate ancestor of T_j^i. Let $T_{j+1}^i \in \mathcal{R}_{j+1}$ represent the child of T_j^i. To define optimal radial cones, we first show the following.

Lemma 1. *In a transitivity-enforced entailment, fine-grained concepts are progressively projected away from the entailment root and into smaller subregion when moving down in the hierarchy.*

Let $\mathfrak{S}_{T_j^i}$ and $\psi(T_j^i)$ denote a cone and its half aperture angle respectively defined at T_j^i with respect to T_0. The transitivity property states that if $T_{j+1}^i \in \mathfrak{S}_{T_j^i}$, then $\mathfrak{S}_{T_{j+1}^i} \subseteq \mathfrak{S}_{T_j^i}$ [14]. The direct consequence of this result is on the aperture angles of nested cones: $\psi(T_j^i) \geq \psi(T_{j+1}^i)$. In other words, if T_{j+1}^i is entailed by T_j^i, then the cone at T_{j+1}^i is completely enclosed by the cone at T_j^i. This means that the distance of the embeddings from the root increases when one goes down in the hierarchy (see Equation 3). Hence, combining the above-stated results, fine-grained concepts (textual embeddings lower in the hierarchy) are contained within smaller cones than coarse-grained concepts. **We provide a mathematical proof in the appendix**. Lemma 1 is a natural property to have in textual entailment because it establishes a direct relationship between the semantic granularity of textual embeddings and their distance from the entailment root.

Ganea *et al*. [14] defined the distance between two embeddings T_j^i and T_l^k in the entailment configuration as the exterior angle (Ξ) between $(T_j^i - T_0)$ and $(T_l^k - T_j^i)$ considering the cone at T_j^i. In the Radial/Euclidean geometry, Ξ is defined as follows:

$$\Xi(T_j^i, T_l^k) = \arccos\left(\frac{\langle (T_j^i - T_0), (T_l^k - T_j^i)\rangle}{||T_j^i - T_0||.||T_l^k - T_j^i||}\right) \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the inner product between the embeddings. Likewise, the similarity measure $\mathcal{S}$ can be defined as:

$$\mathcal{S}(T_j^i, T_l^k) = \cos(\Xi(T_j^i, T_l^k)) \tag{2}$$

Furthermore, they defined the half aperture angle of any cone as a monotonically decreasing function with respect to its distance from the entailment root:

$$\psi(T_j^i) \propto \arcsin(1/r(T_j^i, T_0)) \tag{3}$$

where r is a distance function with range $[\epsilon, 1]$.

Esteva *et al*. [13] proposed entailment configurations to adhere to the transitivity property that defines the partial order of concepts. Inspired from their formulation we devise the transitivity property mathematically as follows:

$$\mathcal{S}(T_{j-1}^i, T_{j+1}^i) \geq \mathcal{S}(T_{j-1}^i, T_j^i).\mathcal{S}(T_j^i, T_{j+1}^i)\ \forall\, j, i \tag{4}$$

where $\mathcal{S} \in [0, 1]$. This constraint establishes the relationship between text embeddings and their higher-level ancestors. In Figure 2, we show two different scenarios in the entailment configuration. Figure 2c) shows a perfect entailment configuration that satisfies transitivity constraints. In Figure 2d) we show a configuration where transitivity is violated. Additionally, for entailment configuration where transitivity holds, Ganea *et al*. [14] showed that $\Xi(T_j^i, T_{j+1}^i) \leq \psi(T_j^i) \leq \pi/2$ is true for any given parent and its child. This means that the cosine similarity between the two embeddings under transitivity constraints is always non-negative with respect to an entailment root.

Alper *et al*. [1] proposed vision-text radial embeddings by minimizing Ξ for positive pairs while maximizing it for negative pairs. They perform text-only fine-tuning while keeping the vision encoder frozen. The main contribution was eliminating the dependence of the objective on aperture angle (Equation 3) and formulating the entailment problem on normalized embeddings. However, this resulted in the objective ignoring the transitivity constraint (equation 4) and providing no guarantee that Lemma 1 holds. Their objective optimizes for the local entailment but does not necessarily solve for the partial order of concepts. Their objective is defined as follows:

$$\mathcal{L}_{LE}(i,j,k) = \Xi(T^i_j, T^i_{j+1}) - \Xi(T^i_j, T^k_{j+1}) \tag{5}$$

where T^k_{j+1} is a negative example for T^i_j. We argue that Alper *et al*. [1] method only optimizes for the *local entailment* objective, i.e. entailment with respect to the immediate ancestor.

4. Method

In this section, we introduce our proposed objective function, which seeks to optimize for transitivity without the requirement of defining an expression for the aperture angles. In addition, we describe our hard negative mining technique for improved performance.

4.1. Global Entailment Learning

We begin by coining the terms local and global entailment. We say local entailment is enforced when T^i_{j+1} is completely entailed by T^i_j up to a reasonable degree, for all possible values of i and j. Global entailment is enforced when Equation 4 holds for all possible sub-hierarchies in addition to local entailment. Mathematically, if $\mathcal{S}(T^i_j, T^i_{j+1}) = \gamma\ (\geq 0)$ and $\mathcal{S}(T^i_{j-1}, T^i_j) = \delta\ (\geq 0)$, then $\mathcal{S}(T^i_{j-1}, T^i_{j+1}) \geq \gamma.\delta$. This ensures that Lemma 1 is satisfied. We enforce this objective using a margin-based loss as follows:

$$\begin{aligned}\mathcal{L}_{GE}(i,j;\alpha) = \max(0, \Xi(T^i_{j-1}, T^i_{j+1}) - \\ \arccos(\mathcal{S}(T^i_j, T^i_{j+1}).\mathcal{S}(T^i_{j-1}, T^i_j)) + \alpha)\end{aligned} \tag{6}$$

where α is the expected margin by which the angles should differ. We set it to the maximum possible value of $\pi/2$. We clip the values of $\mathcal{S}$ between $[0,1]$ for practical implementation. This loss is only calculated for consecutive positive triplets in a given hierarchy.

To enable global and local entailment learning, we combine the global and local objective functions. The loss is iteratively computed for each rank given positive and negative examples. The final loss is a combination of Equations 5 and 6:

$$\begin{aligned}\mathcal{L}_{GLE}(i,k;\alpha) = \frac{1}{N-1}\underbrace{\sum_{p=1}^{N-1}\mathcal{L}_{GE}(i,p;\alpha)}_{\text{Global Entailment}} \\ + \frac{1}{N}\underbrace{\sum_{p=0}^{N-1}\mathcal{L}_{LE}(i,p,K[p])}_{\text{Local Entailment}}\end{aligned} \tag{7}$$

where K represents a set of negative examples for each rank of the hierarchy, indexed by p.

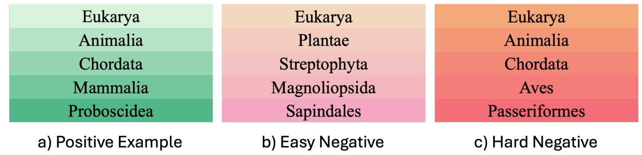


Figure 3. **Hard Negative Examples**. For the local entailment objective, we propose to sample negatives by matching all previous ranks of the positive examples. We recursively sample negatives for each rank separately.

4.2. Radial Cross-Modal Embeddings

In addition to our proposed global entailment objective, we also propose to add a cross-modal alignment loss term to fine-tune the vision encoder along with the text encoder. In [1], they added a prior preservation loss on the text encoder to preserve the original image-text embedding space. Instead, we propose to add a cross-modal alignment term to simultaneously fine-tune the vision and text encoders. This is given by:

$$\mathcal{L}_{\text{CMA}}(i) = -\log\frac{e^{\langle T^i_N, I^i\rangle}}{\sum_{m=1}^{B} e^{\langle T^m_N, I^m\rangle} + e^{\langle T^i_N, I^i\rangle}} \tag{8}$$

where I^i is an image embedding of the same species as represented by T^i_N. Note that the sum in the denominator is over a batch of B negative samples. We only compute the objective for the most granular level of hierarchy which is the *species* level. The final loss is now given by combining Equations 7 and 8:

$$\mathcal{L}_{\text{RCME}}(i,k;\alpha) = \mathcal{L}_{GLE}(i,k;\alpha) + \beta\mathcal{L}_{\text{CMA}}(i) \tag{9}$$

4.3. Hard Negative Mining

We propose a hard negative mining technique to sample negative examples required for the local entailment objective. Recall from equation 5 that the negative example must belong to the same rank as the positive example. To create a hard negative example for a given rank, we propose to sample labels that exactly match the taxonomic labels for all previous ranks of a given positive. In other words, we randomly sample a `sibling` of the parent for a given positive example. We then randomly sample a child of this sibling to create the final negative example. This is done recursively to create negative examples required at each rank. Figure 3 illustrates our hard negative sampling approach. This approach encourages the model to learn fine-grained differences between species of the same ancestry.

5. Experiments

In this section, we present the details of our implementation and the baselines used for comparison. We conduct four experiments to evaluate the effectiveness of the models on: 1) hierarchical retrieval and ordering of taxonomic labels; 2) zero-shot classification at each taxonomic rank; 3) intra-modal image-to-image retrieval at each taxonomic rank; 4) UMAP visualizations of textual embeddings.

5.1. Experimental Setup

Implementation Details. We train two variants of our model. One model is trained using OpenCLIP's initialization and the other model is fine-tuned starting from the BioCLIP's checkpoint (denoted using FT). Both models are based on OpenCLIP's ViT-B/16 architecture. Both models are trained on the TreeofLife-10M dataset using 2 NVIDIA H100 GPUs. We use the word '*Eukarya*' as the entailment root for the Tree of Life. Please refer to the appendix for additional details on the implementation.

Baselines. We compare our models against various vision-language baseline models, including CLIP [30], OpenCLIP [18], BioTroveCLIP [45], BioCLIP [36], TaxaBind [34], Radial Embeddings [1], MERU [11] and ATMG [31]. Each of these models is based on the ViT-B/16 architecture. Since CLIP and OpenCLIP are not specifically trained for the task, we use the common names of species for image classification at the *species* rank [36]. To ensure a fair comparison, we fine-tuned Radial Embeddings, MERU and ATMG on the TreeofLife-10M dataset, starting from BioCLIP's checkpoint. Additionally, we do hard negative mining at each rank for fine-tuning.

Evaluation Datasets. We evaluate our models using the iNaturalist-2021 [39] and BioCLIP-Rare datasets [36]. The iNaturalist-2021 dataset comprises 10,000 unique species of animals, plants, and fungi. It includes a held-out test set with a total of 100,000 images. The BioCLIP-Rare dataset features 400 rare species of animals categorized under the IUCN Red List. Each species in the dataset is represented by 30 images for evaluation.

5.2. Results

Ordering of taxonomic labels. We evaluate the effectiveness of our learned vision-language representations on hierarchical retrieval tasks as defined in Alper *et al*. [1] and Desai *et al*. [11]. Firstly, we check whether the taxonomic labels are correctly ordered according to their distance from the entailment root using Kendall's Tau (τ_d). Secondly, we calculate image-to-text hierarchical retrieval metric relative to each taxonomic label. To ensure a fair evaluation, each unique species in the dataset is represented by a single image, which is selected at random from the test set. More details about the task setup present in the appendix.

Model	Kendall's τ_d	Precision	Recall	F1
CLIP [30]	0.737	0.047	0.054	0.050
OpenCLIP [18]	<u>0.825</u>	0.149	0.190	0.167
BioTroveCLIP [45]	0.566	0.122	0.173	0.143
BioCLIP [36]	0.012	0.115	0.153	0.131
TaxaBind [34]	0.012	0.116	0.155	0.133
Radial Emb. [1]	0.521	0.147	<u>0.196</u>	0.168
MERU [11]	0.403	<u>0.356</u>	0.133	<u>0.193</u>
ATMG [31]	0.571	0.343	0.130	0.189
RCME^{FT} (ours)	0.963	0.386	0.405	0.395
RCME (ours)	**0.993**	**0.458**	**0.572**	**0.508**

Table 1. **Hierarchical Retrieval Metrics**. We evaluate the ability of different models to encode the partial order of taxonomies in the Tree of Life. Additionally, we evaluate the models on the standard task of hierarchical image-text retrieval.

Table 1 presents the performance of the models on the iNaturalist-2021 dataset. Our learned representations exhibit excellent ordering of the taxonomic labels in the embedding space. This means embeddings corresponding to the *species* rank are projected farthest from the root, while those corresponding to the *kingdom* rank are projected closer to the root. Notably, radial embeddings perform worse than CLIP and OpenCLIP in the ordering task. Both our models show significant improvement, with the model trained using CLIP's checkpoint showing a more substantial performance improvement. We see a minimum absolute gain of +0.168 in correlation as compared to the baseline models. Furthermore, our method is able to outperform the rest of the methods in the hierarchical image-to-text retrieval task. We see a minimum absolute gain of +0.102 and +0.376 precision and recall respectively. Overall, these experiments demonstrate that our proposed objective function successfully imparts partial order to the embedding space.

Zero-shot classification. We perform image classification by using taxonomic labels at each rank of the Tree of Life. The classification of each rank is performed independently to assess the ability of the models in the zero-shot setting. Table 2 and 3 show the performance of different models in this task on iNaturalist-2021 and BioCLIP-Rare datasets

Model	Kingdom	Phylum	Class	Family	Order	Genus	Species	Average
CLIP [30]	79.60	37.45	17.97	17.76	05.77	04.90	52.11	30.79
OpenCLIP [18]	66.72	18.42	15.45	07.80	02.60	04.42	58.55	24.99
BioTroveCLIP [45]	37.43	21.81	19.92	10.61	12.91	59.56	68.00	32.89
BioCLIP [36]	36.96	32.02	19.97	24.31	31.43	61.04	68.24	39.13
TaxaBind [34]	40.45	32.22	19.68	24.38	30.80	62.38	70.08	40.00
Radial Emb. [1]	45.84	35.34	22.23	24.96	32.86	61.07	68.23	41.50
MERU [11]	<u>95.82</u>	<u>94.84</u>	<u>63.12</u>	27.27	3.12	1.30	0.73	40.89
ATMG [31]	**99.12**	**86.79**	**73.03**	**51.83**	33.89	49.59	39.52	<u>61.89</u>
RCME^{FT} (ours)	86.18	68.01	38.27	38.16	<u>38.27</u>	<u>64.31</u>	<u>70.81</u>	57.71
RCME (ours)	88.18	84.81	55.22	<u>46.74</u>	**41.82**	**67.41**	**73.52**	**65.09**

Table 2. Zero-shot classification performance on iNaturalist-2021 dataset at various levels of the taxonomy.

Model	Phylum	Class	Family	Order	Genus	Species	Average
CLIP [30]	77.97	42.54	24.35	11.75	14.18	30.41	33.53
OpenCLIP [18]	20.35	33.56	19.14	04.42	10.54	30.22	19.71
BioTroveCLIP [45]	41.69	37.86	22.85	17.76	31.16	27.82	29.84
BioCLIP [36]	43.55	61.08	53.25	45.32	53.38	34.52	48.51
TaxaBind [34]	46.18	60.59	53.95	46.63	<u>55.09</u>	<u>35.84</u>	<u>49.71</u>
Radial Emb. [1]	43.77	63.03	53.96	45.75	53.43	34.85	49.13
MERU [11]	<u>80.66</u>	58.99	25.92	08.12	05.13	04.53	30.56
ATMG [31]	**82.20**	**80.31**	**72.48**	**53.03**	45.02	35.32	61.39
RCME^{FT} (ours)	62.07	62.25	63.64	47.64	55.33	36.79	54.62
RCME (ours)	79.60	<u>77.34</u>	<u>68.41</u>	<u>50.10</u>	**56.66**	**41.62**	**62.64**

Table 3. Zero-shot classification performance on BioCLIP-Rare dataset at various levels of the taxonomy. Note that this dataset primarily contains Animals.

respectively. In both datasets, our model is able to exhibit excellent classification performance at each taxonomic rank. When averaging performance over each taxonomic rank, our model exhibits gains of +5.17% and +2.03% on iNaturalist-2021 and BioCLIP-Rare respectively. For radial emb., MERU and ATMG, we see a gain in performance at higher ranks of the taxonomy, while a dip in performance at fine ranks such as *genus* and *species*. This empirically demonstrates the usefulness of global entailment learning in preserving performance at fine ranks of the hierarchy.

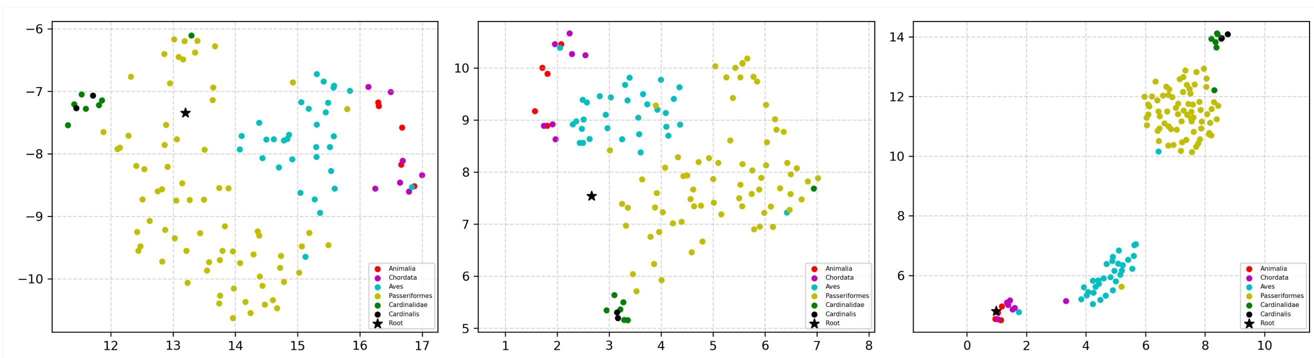
Interestingly, each of the models show lower performance for ranks *class*, *family*, and *order* than for ranks *genus* and *species* in the iNaturalist dataset. This is because of the lower performance of the models for plants as compared to animals in these classes. Notably, this behavior is not seen in Table 3 since the dataset only contains animals. We suspect this happens as plants usually exhibit convergent traits and are usually mislabeled A combination of factors including morphological variability, hybridization, genetic complexity, and evolving methodologies usually complicates plant taxonomy. We show the kingdom-wise performance of our model on this dataset in the appendix to further analyze this behavior. This demonstrates the importance of learning hierarchical representations to detect such behavior and improve the taxonomic classification system.

Image-to-image retrieval. In this experiment, we explore whether our method enhances intra-modal representations. We conduct image-to-image retrieval at each taxonomic rank. Given an image of a particular species and its corresponding taxonomic label at a rank, our objective is to retrieve images of species with the same taxonomic label at the given rank. For example, given an image of an *elephant* with the label *Mammalia*, our goal is to retrieve images of *Mammalia* using solely the image of the given *elephant*.

Table 4 presents the results of this experiment on the iNaturalist-2021 dataset. We compute the recall metric (R@1) for this task. Our method outperforms all other models under consideration. We see a gain of +3.33% when performance is averaged over all taxonomic ranks. Notably, the radial embedding method's performance is limited to the frozen image encoder model used, as it does not fine-tune the image encoder. There is a noticeable gap in the performance of the models at the *species* rank as compared to our model showing that our model is effective at extracting fine-grained visual features. CLIP and OpenCLIP exhibit poor performance, particularly in the *genus* and *species* categories, indicating that these models struggle to extract fine-grained visual features. This suggests the potential of our

Model	Kingdom	Phylum	Class	Family	Order	Genus	Species	Average
CLIP [30]	95.09	91.38	80.04	49.29	29.50	14.64	10.60	52.94
OpenCLIP [18]	96.29	93.33	85.42	59.85	41.85	26.11	16.78	59.95
BioTroveCLIP [45]	98.60	98.07	95.31	85.06	78.89	68.97	55.27	82.88
BioCLIP [36]	98.50	97.60	94.94	84.98	78.63	67.22	51.62	81.92
TaxaBind [34]	98.66	98.07	95.71	85.64	78.90	68.84	54.47	82.90
Radial Emb.* [1]	98.50	97.60	94.94	84.98	78.63	67.22	51.62	81.92
MERU [11]	98.20	96.02	85.04	58.81	41.96	26.56	16.40	60.43
ATMG [31]	99.16	98.45	96.88	89.02	82.30	70.09	52.69	84.08
RCME$^{\text{FT}}$ (ours)	98.72	98.08	95.76	86.95	80.43	70.27	57.10	83.91
RCME (ours)	**99.65**	**99.04**	**97.11**	**90.61**	**85.38**	**75.09**	**61.27**	**86.88**

Table 4. **Image-to-Image Retrieval**. We evaluate the effectiveness of the intra-modal image representations learned by the models on the task of image-to-image retrieval at each taxonomic rank. *Radial Emb. performs identically to BioCLIP since the vision encoder is not fine-tuned during its training.



(a) BioCLIP (b) Radial Embeddings (c) RCME (Ours)

Figure 4. **UMAP Visualization of Textual Embeddings**. We visualize the textual embeddings using 2-D UMAP to show our model learns to preserve the partial order of taxonomic labels based on their distance from the entailment root.

model in future applications such as fine-grained retrieval augmented generation.

UMAP visualization of textual embeddings. In Figure 4, we present a 2-D Uniform Manifold Approximation and Projection (UMAP) visualization of textual embeddings obtained from BioCLIP, Radial Emb., and RCME. For a given *species*, we plot the embedding for all the ancestors in the taxonomic hierarchy and their siblings. For instance, if we are given the specie *Cardinalis cardinalis*, we begin with kingdom label *Animalia* and plot all the *Phyla* that belong to the kingdom *Animalia*. This is repeated for all the subsequent ranks. From the plot, we anticipate two properties: 1) siblings share similar embeddings; 2) the embeddings are ordered in a coarse-to-fine manner (from kingdom to species), based on their distance from the entailment root. From the figure, it is evident that our model has successfully preserved the partial order of taxonomic labels based on their distance from the entailment root. For instance, the embeddings corresponding to the *species* rank are projected farthest away from the entailment root. However, BioCLIP and Radial Emb. are unable to effectively enforce transitivity. We provide additional UMAPs in the appendix.

5.3. Ablations

We conduct an ablation study to analyze the effect of each loss component on the performance of our models. For the losses not using our cross-modal alignment term ($\mathcal{L}_{CMA}$), we include the $\mathcal{L}_{prior}$ from Alper *et al.* [1] to preserve the original vision-language alignment. It is given as: $\mathcal{L}_{prior} = -\langle T_j^i, T_j^{i*} \rangle$, where T_j^{i*} is an embedding from a frozen pretrained text encoder. For losses using $\mathcal{L}_{prior}$, we perform fine-tuning starting from BioCLIP's checkpoint. Table 5 presents the performance of these losses for the image classification task on the iNaturalist-2021 dataset. Notably, in-

$\mathcal{L}_{LE}$	$\mathcal{L}_{GE}$	$\mathcal{L}_{prior}$	$\mathcal{L}_{CMA}$	Kingdom	Phylum	Class	Family	Order	Genus	Species	Average
			✓	36.96	32.02	19.97	24.31	31.43	61.04	68.24	39.13
✓		✓		45.84	35.34	22.23	24.96	32.86	61.07	68.23	41.50
✓			✓	47.34	38.14	25.78	26.88	35.34	62.67	69.43	43.65
✓	✓	✓		85.13	81.11	53.21	44.33	39.82	65.90	71.28	62.97
✓	✓		✓	**88.18**	**84.81**	**55.22**	**46.74**	**41.82**	**67.41**	**73.52**	**65.09**

Table 5. **Loss Ablation**. We evaluate the performance of our model when trained with various combinations of the loss terms proposed in our objective function.

RCME	iNaturalist-2021	BioCLIP-Rare
w/o negative mining	62.22	59.12
with negative mining	**65.09**	**62.64**

Table 6. **Benefits of hard negative mining**. We evaluate the effectiveness of our hard negative mining approach for hierarchical representation learning of Tree of Life.

Model	Kendall's τ_d	Precision	Recall	F1
CLIP^B	0.883	**0.335**	0.142	0.199
CLIP^B (MERU)	0.855	0.122	0.401	0.187
CLIP^B (ATMG)	0.981	0.134	0.422	0.203
CLIP^B (HyCoCLIP)	0.892	0.124	0.451	0.194
CLIP^B (Radial Emb.)	0.988	0.155	0.441	0.229
CLIP^B (RCME)	**0.991**	0.162	**0.467**	**0.241**
CLIP^L	0.881	0.151	0.343	0.209
CLIP^L (Radial Emb.)	0.973	0.145	0.415	0.215
CLIP^L (RCME)	**0.992**	**0.158**	**0.452**	**0.234**

Table 7. Hierarchical retrieval metrics on HierarCaps dataset. Our objective function results in improved ordering and image-to-text retrieval performance.

corporating our proposed global entailment objective function significantly enhances models' performance compared to using only the local entailment objective. We notice a gain of +51.73% in performance when adding our global entailment objective function to Alper *et al.*'s [1] objective function. Furthermore, our cross-modal alignment term outperforms the prior preservation loss. We notice a minimum gain of +3.36% when replacing the prior preservation loss with our cross-modal alignment loss to train the vision and text encoder simultaneously.

We additionally investigate whether our proposed hard negative mining approach outperforms the random sampling approach. We evaluate the models trained using our negative mining and random sampling approaches on the iNaturalist-2021 and BioCLIP-Rare datasets. Table 6 presents the comparison. We notice that we get a performance improvement of +4.61% and +5.62% on both datasets respectively. See appendix for additional ablations.

5.4. Generalization to HierarCaps

To demonstrate the generalizability of our proposed objective function across other domains, we conducted experiments on the HierarCaps dataset [1]. This dataset comprises a subset of images from the Conceptual Captions (CC) dataset, each accompanied by captions at four different levels of granularity. We fine-tune the ViT-B/16 and ViT-L/14 variants of the CLIP model on this dataset using our proposed objective function in equation 9 without hard negative mining. Once trained, we compute hierarchical retrieval metrics on the held-out test set of HierarCaps. As evident from Table 7, our model outperforms radial embeddings in both ordering and hierarchical image-to-text retrieval tasks. These results demonstrate the successful application of our objective function in other application domains, enabling the imposition of a partial ordering along with entailment in an embedding space.

6. Conclusion

In this work, we presented Radial Cross-Modal Embeddings (RCME), a framework for learning transitivity-enforced entailment in vision-language models. We proposed a novel objective function to enable global learning of entailment, which aids in preserving the partial order of concepts. Our framework not only improves cross-modal representations but also intra-modal representations. By leveraging our framework, we proposed a hierarchical foundation model for the Tree of Life, outperforming the state-of-the-art. We showed how hierarchical representations can improve the taxonomic classification of species and reveal unusual patterns in the taxonomic classification system, especially in plants. Our future works will focus on using the learned hierarchical representations to understand and comprehend species evolution, identify distinctive anomalies within the Tree of Life, and devise strategies to enhance the taxonomic classification system.

7. Acknowledgements

This research used the TGI RAILs advanced compute and data resource which is supported by the National Science Foundation (award OAC-2232860) and the Taylor Geospatial Institute.

References

[1] Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations. *European Conference on Computer Vision*, 2024. 2, 4, 5, 6, 7, 8

[2] Ivana Balazevic, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 2

[4] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019. 2

[5] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. *arXiv preprint arXiv:2005.00545*, 2020. 2

[6] Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13061, 2023. 1, 2

[7] Theresa Chen and Yao-Yi Chiang. Mitree: Multi-input transformer ecoregion encoder for species distribution modelling. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 110–120, 2024. 2

[8] Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15095–15104, 2023. 2

[9] Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. *arXiv preprint arXiv:2410.18857*, 2024. 2

[10] Rangel Daroya, Elijah Cole, Oisin Mac Aodha, Grant Van Horn, and Subhransu Maji. Wildsat: Learning satellite image representations from wildlife observations. *arXiv preprint arXiv:2412.14428*, 2024. 2

[11] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023. 2, 5, 6, 7

[12] Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 59–69, New Orleans, Louisiana, USA, 2018. Association for Computational Linguistics. 2

[13] Francesc Esteva, Lluís Godo, Ricardo O Rodríguez, and Thomas Vetterlein. Logics for approximate and strong entailments. *Fuzzy Sets and Systems*, 197:59–70, 2012. 3

[14] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*, pages 1646–1655. PMLR, 2018. 2, 3

[15] ZeMing Gong, Austin Wang, Xiaoliang Huo, Joakim Bruslund Haurum, Scott C. Lowe, Graham W. Taylor, and Angel X Chang. CLIBD: Bridging vision and genomics for biodiversity monitoring at scale. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2

[16] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12893–12902, 2024. 1

[17] Andy V Huynh, Lauren E Gillespie, Jael Lopez-Saucedo, Claire Tang, Rohan Sikand, and Moisés Expósito-Alonso. Contrastive ground-level image and remote sensing pre-training improves representation learning for natural world imagery. In *European Conference on Computer Vision*, pages 173–190. Springer, 2024. 2

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 5, 6, 7

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[20] Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. *arXiv preprint arXiv:1902.00913*, 2019. 2

[21] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[22] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 2

[23] M. Maruf, Arka Daw, Kazi Sajeed Mehrab, Harish Babu Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P. Balhoff, Yasin Bakis, Bahadir Altintas, Matthew J. Thompson, Elizabeth G. Campolongo, Josef C.

Uyeda, Hilmar Lapp, Henry L. Bart, Paula M. Mabee, Yu Su, Wei-Lun Chao, Charles Stewart, Tanya Berger-Wolf, Wasila Dahdul, and Anuj Karpatne. Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images, 2024. 2

[24] Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. *arXiv preprint arXiv:2502.04263*, 2025. 2

[25] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19023–19034, 2022. 1, 2

[26] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 2

[27] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018. 2

[28] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *The Thirteenth International Conference on Learning Representations*, 2025. 2

[29] Laura J Pollock, Justin Kitzes, Sara Beery, Kaitlyn M Gaynor, Marta A Jarzyna, Oisin Mac Aodha, Bernd Meyer, David Rolnick, Graham W Taylor, Devis Tuia, et al. Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nature Reviews Biodiversity*, pages 1–17, 2025. 1

[30] 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. 2, 5, 6, 7

[31] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27263–27272, 2024. 2, 5, 6, 7

[32] Alex David Rogers, Hannah Appiah-Madson, Jeff A Ardron, Nicholas J Bax, Punyasloke Bhadury, Angelika Brandt, Pier-Luigi Buttigieg, Olivier De Clerck, Claudia Delgado, Daniel L Distel, et al. Accelerating ocean species discovery and laying the foundations for the future of marine biodiversity research and monitoring. *Frontiers in Marine Science*, 10:1224471, 2023. 1

[33] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, and Nathan Jacobs. Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7136–7145, 2024. 2

[34] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *Winter Conference on Applications of Computer Vision*. IEEE/CVF, 2025. 1, 2, 5, 6, 7

[35] Fanny Simões, Charles Bouveyron, and Frédéric Precioso. Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning. *Ecological Informatics*, 75:102095, 2023. 2

[36] 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. 1, 2, 5, 6, 7

[37] Jong-Chyi Su and Subhransu Maji. Semi-supervised learning with taxonomic labels. *arXiv preprint arXiv:2111.11595*, 2021. 1

[38] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 2

[39] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 1, 2, 5

[40] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 2

[41] 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, 2024. 2

[42] Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate 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, 2025. 1

[43] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2023. 1

[44] Ziwei Wang, Sameera Ramasinghe, Chenchen Xu, Julien Monteil, Loris Bazzani, and Thalaiyasingam Ajanthan. Learning visual hierarchies with hyperbolic embeddings. *arXiv preprint arXiv:2411.17490*, 2024. 2

[45] 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, 2025. 1, 2, 5, 6, 7

[46] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[47] Tao Yu, Toni JB Liu, Albert Tseng, and Christopher De Sa. Shadow cones: A generalized framework for partial order embeddings. *arXiv preprint arXiv:2305.15215*, 2023. 2