TAXAWALK: SPECIES RECOGNITION AS A WALK DOWN THE TREE OF LIFE

Anonymous authors
Paper under double-blind review

ABSTRACT

Biodiversity has long been organized through a hierarchical system of biological classification. This hierarchy is closely related to visual recognition from broad to fine-grained taxonomic distinctions, narrowing the space of possible identities at each rank. Yet existing methods either classify species directly or predict each taxonomic rank independently, without allowing coarse-rank decisions to guide finer-grained identification. We propose a different perspective, modelling identification as a *traversal* of the taxonomy rather than a classification over its leaves. We introduce **TaxaWalk**, a hierarchical taxonomic model that autoregressively predicts an organism's lineage, one rank at a time, conditioned on preceding ranks. This formulation enables *expert-conditioned* inference where a biologist can specify a taxon at any rank, and the model completes the remaining lineage consistently. We further introduce an adaptive training objective that focuses learning on underperforming ranks, improving fine-grained recognition without allowing easier, coarse-grained predictions to dominate training. TaxaWalk sets a new state of the art in zero-shot open-vocabulary recognition across multiple biodiversity benchmarks, while supporting expert conditioning and zero-shot novelty detection. All code and models will be released upon acceptance.

1 INTRODUCTION

Computer vision has become an essential tool for studying the natural world, driven by the rapidly growing collection of imagery being uploaded to citizen-science platforms, museum digitization and automated sensors (Van Horn et al., 2018; Stevens et al., 2024). These efforts increasingly rely on automatic identification, which is hard because the label space is enormous, fine-grained and open. An estimated 8.7 million eukaryotic species exist and most remain undescribed (Mora et al., 2011; Pollock et al., 2025), while the largest curated recognition benchmarks cover hundreds of thousands of taxa (Gu et al., 2026). Any model deployed on real survey imagery must therefore separate visually similar species, name organisms without labeled training data, and exploit what an expert taxonomist already knows.

Species labels are not a flat set of identifiers. They form an ordered lineage from kingdom to species, and each rank is meaningful only relative to its ancestors. Existing models underutilize this structure. Supervised classifiers encode the hierarchy but close the label space. A flat species classifier that reads its ancestor ranks off a fixed tree of life ties the entire lineage to a single leaf decision, so any error at the leaf propagates upward. Hierarchical classifiers build this order into their prediction heads and keep the ranks consistent (Park et al., 2025; 2026), but their label set is fixed at training time, so adding a species requires retraining.

Vision-language models take the opposite trade-off. They retain an open label set and encode taxonomic structure in their embeddings, either through large-scale hierarchical contrastive training (Gu et al., 2026) or explicit entailment objectives (Sastry et al., 2025a). However, they do not exploit this structure at decoding time. Existing methods either predict a species in a single flat comparison and infer its higher ranks afterward (Park et al., 2025), or predict each rank independently, without carrying earlier decisions into later ones (Sastry et al., 2025a). An expert's identification can therefore filter the candidate set, but it cannot condition the model's subsequent predictions.

We argue that taxonomic identification should be modeled as a sequential decision process. In practice, a taxonomist rules out broad clades before weighing the characteristics that separate visually

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

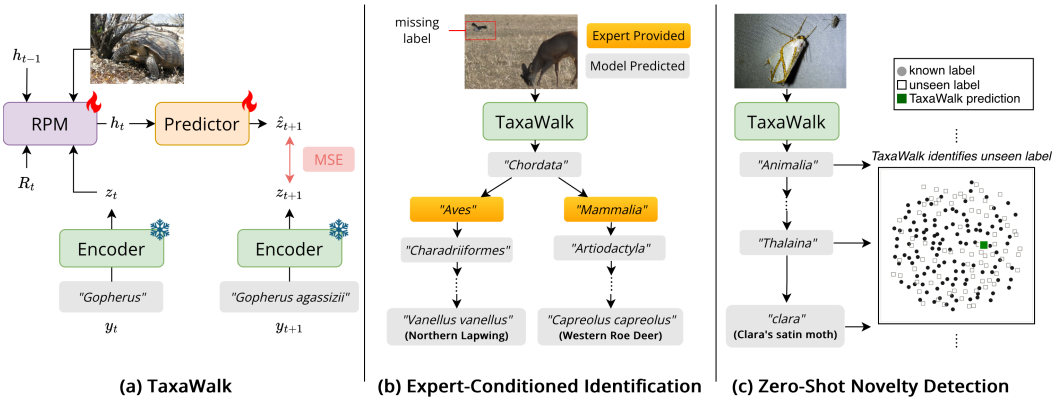

Figure 1: **TaxaWalk.** *(a)* TaxaWalk predicts the next rank's embedding z_{t+1} given the image, its previous prediction z_t, and recurrent state h_{t-1}. *(b)* Expert-provided ranks can steer predictions towards different species in the same image. *(c)* TaxaWalk can predict where a lineage leaves the known taxonomy by utilizing a per-rank distance to known labels.

confusing species, and stops at whichever rank the available evidence supports. This order is not a convention. The cues that place an organism in a class are coarse and low frequency, such as body plan and habitat, and the cues that separate two species in a genus are fine and high frequency, such as textures and markings, so each rank asks a different question of the image. A flat classifier must answer all of them in one comparison, and a distant species with similar high-frequency features can confuse the classifier. Sequential decisions that separate low-frequency features from high-frequency features make each comparison easier, since the candidates at a fine rank already share the coarse structure and differ only in the details that rank is meant to resolve. More importantly, this creates an inference-time interface for human expertise where a biologist can fix any rank they know, and the model conditions its remaining predictions on that decision.

We therefore propose **TaxaWalk** (Figure 1), which predicts a lineage one rank at a time in BioCLIP-2's frozen text embedding space. TaxaWalk is designed to identify organisms autonomously, and expose the taxonomy as an inference-time interface where a biologist can fix any known rank and directly condition the predictions below it. Keeping the text embedding space frozen, we learn a small recurrent module that reads the image and the ranks already decided, regresses the embedding of the next rank, and matches it to the nearest candidate, repeating this along the taxonomic axis.

Our contributions are as follows:

1. **Taxonomy as autoregressive traversal.** We formulate identification as next-rank prediction in a frozen contrastive text embedding space, instantiated as a recurrent predictive module with a mixture of rank experts and a rank-balanced objective. Candidates are supplied at inference, so the label set stays open while every rank is conditioned on those above it (§3). We show that TaxaWalk achieves state-of-the-art species accuracy on multiple open-vocabulary benchmarks.
2. **Expert-conditioned identification.** We show that TaxaWalk obeys constraints given at inference, completing a lineage below what an expert knows. To test this we curate iNat Co-Observed, nearly 6K photographs that each contain two independently identified species, and show that TaxaWalk outperforms existing open-vocabulary models (§4.3, Figure 1b).
3. **Zero-shot novelty detection.** We show that TaxaWalk outperforms existing methods on novel species detection across multiple regions and increasing tiers of novelty on Open-Insect, without fine-tuning or training a detector. In addition, TaxaWalk adds the ability to surface novelty at whichever rank the organism leaves the known set (§4.4, Figure 1c).

2 Related Work

Hierarchical classification. Hierarchical classification models predict a label at every level of a class tree. Classical methods descend the tree, conditioning each level on the one above (Silla & Freitas, 2011; Yan et al., 2015; Redmon & Farhadi, 2017), which lets coarse errors propagate downward (Park et al., 2025). Recent work instead predicts all levels jointly (Chang et al., 2021) and keeps them consistent (Park et al., 2025). Other work learns from labels given at any level (Park et al., 2026), and chooses how deep to predict for each image (Valmadre, 2022; Park et al., 2026). The class hierarchy has also been used to make mistakes less severe (Bertinetto et al., 2020; Karthik et al., 2021; Garg et al., 2022) and to place an object of a novel class under its closest known ancestor (Lee et al., 2018). In all of these methods the tree and its labels are fixed at training time, so adding a species requires retraining. TaxaWalk keeps the top-down conditioning but predicts in an open-vocabulary embedding space, where the candidates at each rank are supplied at inference.

Vision-language models for biology. Contrastive vision-language models (VLMs) replace the classification head with matching images against embedded label text allowing for an open-vocabulary label set (Radford et al., 2021; Ilharco et al., 2021; Tschannen et al., 2025). In biology, this approach has been used to train on taxonomic names at increasing scale (Stevens et al., 2024; Yang et al., 2024; Gu et al., 2026) and extended to additional modalities such as location, audio, satellite imagery, climate and DNA (Sastry et al., 2025b; Gong et al., 2025b). Other work has injected hierarchical information into the embedding geometry through entailment, hyperbolic space or hierarchical regularization (Sastry et al., 2025a; Gong et al., 2025a; Ahmed et al., 2026; Tao et al., 2026), or tunes prompts so that predictions agree across levels (Wu et al., 2024). All of these methods have shown to improve the representation, but inference remains a flat lookup. Current state-of-the-art representations (Stevens et al., 2024; Gu et al., 2026; Sastry et al., 2025a) encode the lineage as a sequence, but decoding remains a flat lookup. Multimodal language models decode a lineage rank by rank as text (Li et al., 2026), but their accuracy collapses at the species rank (Chiranjeevi et al., 2025). TaxaWalk keeps the strong taxonomic text embedding space frozen and learns what they lack: a decoder that moves through them rank by rank.

Predictive representation learning. Recent work in representation learning captures structure by predicting latent targets rather than reconstructing pixels. I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024) predict masked visual features, while LeJEPA (Balestriero & LeCun, 2025) and LeWorldModel (Maes et al., 2026) use distributional regularization to stabilize jointly learned representations. DINO-WM (Zhou et al., 2024) instead learns visual dynamics in a frozen feature space. Cross-modal methods predict language embeddings from images (Chen et al., 2026a; Kuhn et al., 2026) or learn mappings among image, text and audio representations (Lei et al., 2024). TaxaWalk introduces a semantic-granularity axis, using frozen BioCLIP-2 text embeddings as stable targets and predicting them along a taxonomic path conditioned on the image and its ancestors.

3 Method

3.1 Problem Formulation

Given an image $\mathbf{x}$ of an organism, our goal is to predict its taxonomic lineage $y = (y_1, \ldots, y_T)$, the ordered labels from domain to species. The domain y_1 is always EUKARYA and starts the sequence, so the remaining ranks from kingdom to species are predicted ($T = 8$).

We work in the joint embedding space of BioCLIP-2 (Gu et al., 2026), with image encoder $\mathcal{E}_V$ and text encoder $\mathcal{E}_T$. We adapt $\mathcal{E}_V$ using LoRA while keeping $\mathcal{E}_T$ frozen. Let $\mathbf{v} = \mathcal{E}_V(\mathbf{x}) \in \mathbb{R}^d$ be the image embedding and $z_t = \mathcal{E}_T(y_t) \in \mathbb{R}^d$ the embedding of the rank-t label, where each label is embedded on its own. TaxaWalk casts lineage prediction as autoregressive next-rank prediction in this space, with z_t as the latent state and the image as a fixed context. A transition model f_θ predicts the embedding of the next rank from the image and the ranks decided so far:

$$\hat{z}_{t+1} \;=\; f_\theta\big(z_{1:t},\, \mathbf{v}\big), \qquad t = 1, \ldots, T-1. \tag{1}$$

The regression target for $\hat{z}_{t+1}$ is z_{t+1}, the frozen embedding of the true next label. This turns a flat K-way classification into a short sequential problem in a contrastive embedding space.

Inference. A predicted embedding is decoded to a label by nearest-neighbour search over a candidate set:

$$\hat{y}_t \;=\; \arg\min_{\ell \in \mathcal{V}_t(\hat{y}_{<t})} \big\| \hat{z}_t - \mathcal{E}_T(\ell) \big\|_2, \tag{2}$$

where $\mathcal{V}_t(\hat{y}_{<t})$ holds the rank-t candidates consistent with the ranks above. Candidates are supplied at inference and may include taxa never seen in training. In the rollout (AR), the embedding of the decoded label is fed back as z_t and the next rank is predicted. Alternatively, a known label can be supplied at any rank, conditioning all predictions below it. We call this expert-conditioned inference (EC).

Novelty detection. The nearest-neighbor distance in Eq. 2 can also serve as a rank-specific novelty score, $s_t = \min_\ell \|\hat{z}_t - \mathcal{E}_T(\ell)\|_2$. A small value indicates that the predicted embedding is close to a known candidate, whereas a large value indicates that no candidate matches it well. Comparing this score across ranks can therefore identify where the predicted lineage departs from the known taxonomy.

3.2 Architecture

We instantiate f_θ as a small recurrent module on top of BioCLIP-2 (Figure 1a). The frozen label embeddings z_t serve as inputs, regression targets and candidates for decoding. Thus, TaxaWalk learns transitions within the text embedding space without altering its taxonomic geometry.

The Recurrent Predictive Module (RPM) keeps a hidden state h_t that summarises the ranks decided so far, and updates it once per rank,

$$h_t \;=\; \mathrm{RPM}\big(h_{t-1},\, z_t,\, c_t\big), \qquad \hat{z}_{t+1} \;=\; g_{\text{out}}(h_t), \tag{3}$$

where the context c_t combines the image embedding $\mathbf{v}$ with a learned embedding of the rank, and g_{out} is the output head (§3.3). Each RPM cell is a gated recurrent unit (Cho et al., 2014) whose candidate state comes from a transformer block modulated by c_t through adaptive layer normalisation (Peebles & Xie, 2023), and N such cells are stacked. Because the recurrence only moves forward in t, the prediction for rank $t{+}1$ depends only on $z_{\le t}$, so one network supports teacher-forced training in a single pass, autoregressive rollout and expert-conditioned inference.

3.3 Mixture of Rank Experts (MoRE)

The output head must serve ranks with large differences in candidate set size, from a few kingdoms to hundreds of thousands of species. A single dense head would share its capacity across all ranks, so we use a sparse mixture of experts (Shazeer et al., 2017; Fedus et al., 2022) that is routed per rank. With E routed experts $f_1, \ldots, f_E$ and one shared expert f_s, each a two-layer MLP, the head is

$$g_{\text{out}}(h_t) \;=\; f_s(h_t) \;+\; \sum_{e \in S_t} w_{t,e}\, f_e(h_t), \tag{4}$$

where S_t is the set of k experts selected at rank t and $w_{t,e}$ are their normalised gate values. Experts are chosen once per rank from the batch-mean hidden state, so the choice is shared by all images in a batch and the experts specialise by rank. We use $E = 8$ and $k = 2$, with sigmoid gates and bias-based load balancing (DeepSeek-AI et al., 2024; Dai et al., 2024). Further implementation details are provided in Appendix D.

3.4 Rank-Balanced Objective

For a batch of B images, the error of the prediction at step t is the mean squared distance to the target,

$$\bar{e}_t \;=\; \frac{1}{B} \sum_{n=1}^{B} \tfrac{1}{d} \big\| \hat{z}_{n,t+1} - z_{n,t+1} \big\|_2^2, \qquad t = 1, \ldots, T-1, \tag{5}$$

where images whose rank $t{+}1$ is unlabelled are unscored. Naively averaging $\bar{e}_t$ uniformly over ranks gives every rank the same share of the gradient. We find that coarse ranks saturate early in

training and keep their share, while harder ranks like genus and species remain under-fit. We instead aggregate the errors with a temperature-τ LogSumExp (Li et al., 2020),

$$\mathcal{L} \;=\; \tau \,\log \sum_{t=1}^{T-1} \exp\!\big(\bar{e}_t/\tau\big), \tag{6}$$

so ranks with larger error receive more of the gradient. The loss reduces to the worst-rank error as $\tau \to 0$ and weights all ranks equally as $\tau \to \infty$. Appendix A gives a gradient analysis of this objective.

4 Experiments

4.1 Experimental Setup

Training details. TaxaWalk builds on BioCLIP-2 ViT-L (Gu et al., 2026). The text encoder is frozen, while the image encoder receives LoRA adapters (rank 32, $\alpha = 64$) on every block. Additionally, we train an RPM with $N = 6$ cells and $D_H = 1024$ and a MoRE head with $E = 8$ experts. All parameters are trained on TreeOfLife-200M, the corpus BioCLIP-2 was trained on, so TaxaWalk sees no data the underlying encoder backbone has not seen. We use the rank-balanced objective at $\tau = 0.1$. For images with incomplete lineages a missing rank is replaced by an `[unknown]` token, which is left out of the loss, and we replace 1% of known ranks with this token so the model learns to proceed with missing taxonomic context. We use AdamW with a learning rate of 10^{-4} and a global batch size of 4096 over four NVIDIA H100 GPUs.

Benchmarks and baselines. We evaluate TaxaWalk in five settings: open-vocabulary species recognition, cryptic species recognition, comparison against fully supervised classifiers, expert-conditioned identification, and novel species detection. The recognition settings share one evaluation protocol. For each image, the seven ranks from kingdom to species are scored separately using candidate sets built from each benchmark's taxonomy. All naming is normalized to the GBIF Backbone Taxonomy. We report accuracy at the finest rank a benchmark provides and the mean accuracy over its ranks. We compare against general vision-language models as well as state-of-the-art biological vision-language models.

Inference regimes. A deployed identifier may operate autonomously or incorporate a biologist's prior beliefs. Autoregressive rollout (AR) uses only the image, whereas expert-conditioned inference (EC) receives a known taxonomic prefix. In evaluation, we instantiate EC with the ground-truth ancestors of each scored rank, simulating correct expert input. In both, each rank is chosen among the candidate taxa consistent with the ranks above it (Eq. 2). For the baselines, we find that scoring each rank separately is unfavourable, since they are trained on full lineages. We instead classify the species against full lineage strings and read the ancestors off the predicted lineage, which keeps their species accuracy and makes their ranks consistent. When ranks above leave a single candidate, such as monotypic genus, we score the rank on the decoded ancestors.

4.2 Open-Vocabulary Species Recognition

We first test whether traversal improves classification when the label set is open. Table 1 covers seven benchmarks that span camera traps, birds, plankton, insects, plants, fungi and rare species, and Table 2 covers cryptic species (Manolache et al., 2025), which are hard to separate even for experts. No method is fine-tuned on these benchmarks.

From the image alone (AR), TaxaWalk has the highest species accuracy on all seven benchmarks in Table 1, and the highest mean accuracy across ranks on six. TaxaWalk has the largest improvements on benchmarks that have been particularly challenging for existing models, such as Camera Trap (90.4 against 44.3), Insects-2 (92.3 against 44.2) and Rare Species (83.9 against 58.7). Figure 2 shows one such image: the flat lookup leaves the true lineage at class, whereas the rollout stays on it and narrows the candidate set at each rank. Figure 3 shows the same pattern across images of *Felis catus*, and Figure 6 in Appendix E.1 across images of *Lynx rufus*. BioCLIP-2 frequently diverges at class toward taxonomically distant species, whereas TaxaWalk retains more images along the true

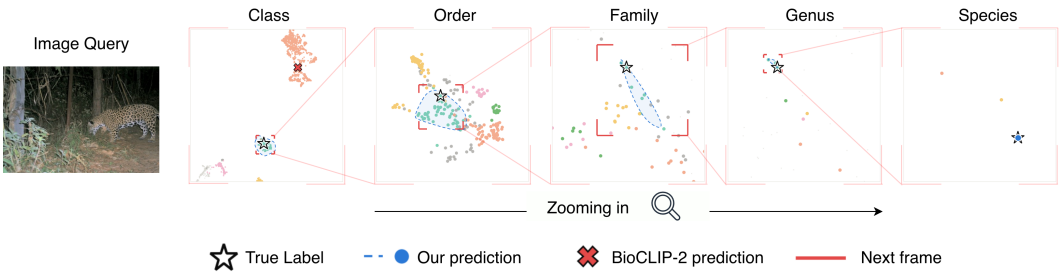

Figure 2: **Rollout in the embedding space.** Each panel projects species embeddings to 2D with UMAP and colours them by their ancestor at the displayed rank. The blue outline/point is TaxaWalk's predicted rank embedding. We zoom in on the relevant region of the embedding space while traversing the hierarchy. TaxaWalk follows the true lineage (star), whereas BioCLIP-2's flat species prediction (red cross) diverges at class.

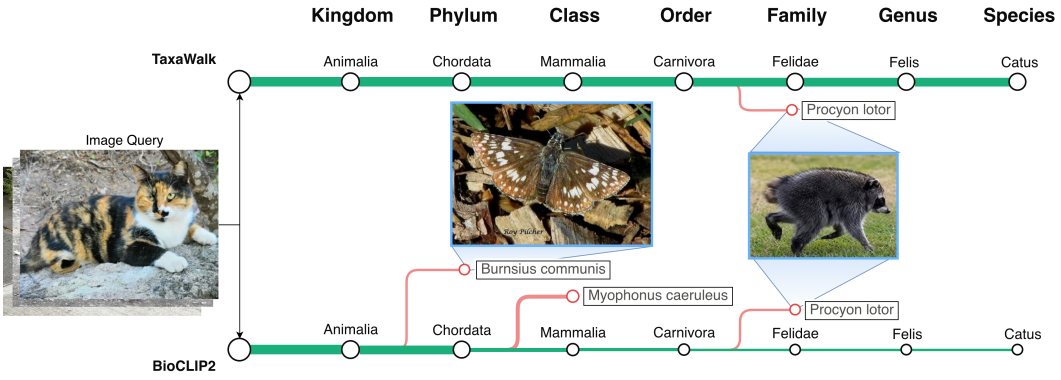

Figure 3: **Traversing the taxonomy.** Predictions for iNat-2021 validation images of *Felis catus*. Green traces the true lineage, with line width indicating how many images remain correct. Red branches mark the rank of the first error and label the resulting species prediction. Insets show an example of the incorrectly predicted species. TaxaWalk (top) retains more images through species, whereas BioCLIP-2 (bottom) frequently diverges at coarser ranks.

lineage and its errors begin much later in the lineage. Supplying the true ancestors (EC) shows some additional gains, but most of the benefit is available without expert input. On CrypticBio, the rollout raises species accuracy on all four splits, by 11 to 29 points, with the largest gain on the Endangered split (70.1 against 40.8), the hardest of the four for every method. On benchmarks where all candidates share a genus or family, such as Heliconius, hard ticks and zebras, the higher ranks have little to rule out, and TaxaWalk performs on par with the best baseline.

Closed-set comparison. Table 3a asks whether candidate pruning alone explains the gains. Applying hierarchical pruning directly to BioCLIP-2 lowers species accuracy from 77.9% to 74.0%. In contrast, pruning improves TaxaWalk under AR from 87.5% to 89.5%, showing that candidate restriction is effective when paired with learned rank conditioning. Supplying true ancestors through EC raises accuracy further to 94.1%.

Panel (b) tests whether the training objective preserves open-vocabulary transfer, using models trained on iNat-2021 only. MSE regresses each prediction toward its frozen text embedding, whereas CE applies cross-entropy to the logits $-\|\hat{z}_t - \mathcal{E}_T(c)\|_2$ over the iNat-2021 name bank at rank t. CE wins in distribution, 93.2% against 88.6% species accuracy, but transfers to Rare Species at 0.9% against 46.1% for regression. Regression to frozen text targets is what carries the model beyond its training label bank.

Table 1: **Open-vocabulary species recognition** on seven benchmarks. Each cell is **species / mean-across-ranks** top-1 accuracy (%). The mean is unweighted over the ranks with at least two distinct labels in the benchmark, and each rank is scored on the images that carry a label at that rank. Ranks use their available annotations rather than a shared image set, so species accuracy can exceed mean-rank accuracy. Best per column in **bold**.

Arch.	Method	Camera Trap	NABirds	Plankton	Insects-2	PlantNet	Fungi	Rare Species
ViT-L	CLIP	19.0 / 39.4	09.1 / 18.8	04.5 / 01.9	22.6 / 39.8	55.3 / 40.6	09.7 / 21.0	14.4 / 39.0
	OpenCLIP	11.4 / 33.2	07.5 / 07.3	03.6 / 04.7	15.7 / 29.7	55.0 / 39.5	08.3 / 12.4	09.7 / 26.2
	SigLIP-2	29.7 / 42.9	30.6 / 49.6	7.6 / 13.6	30.7 / 55.0	82.6 / 61.0	38.7 / 47.7	37.0 / 57.1
ViT-B	BioCLIP	30.0 / 55.1	81.8 / 91.6	14.8 / 33.0	35.7 / 59.2	92.3 / 93.9	49.3 / 72.3	33.7 / 65.1
	BioTrove-CLIP	06.5 / 28.3	53.3 / 77.9	06.1 / 12.3	21.5 / 48.2	62.9 / 70.3	44.1 / 74.4	21.6 / 39.0
	TaxaBind	30.5 / 55.5	82.6 / 92.2	18.9 / 35.0	40.3 / 61.5	93.5 / 95.2	59.9 / 79.4	34.8 / 66.3
	RCME	32.7 / 57.8	83.8 / 93.2	18.1 / 35.0	43.5 / 64.4	93.3 / 95.3	57.5 / 80.2	40.0 / 71.9
ViT-L	BioCLIP-2	44.3 / 65.0	87.9 / 94.9	13.3 / 40.2	44.2 / 64.2	97.0 / 98.0	87.1 / 95.2	58.7 / 80.0
ViT-L	TaxaWalk (AR, **ours**)	90.4 / 81.8	94.9 / 98.0	70.9 / 51.5	92.3 / 84.9	98.3 / 98.1	87.5 / 95.5	83.9 / 78.8
	TaxaWalk (EC, **ours**)	**91.6 / 86.4**	**95.2 / 98.2**	**71.1 / 59.5**	**92.4 / 86.0**	**98.7 / 98.3**	**91.8 / 95.8**	**87.5 / 82.1**

Table 2: **Cryptic species recognition.** Top-1 **species / mean-rank** accuracy (%) on CrypticBio—Common (CB-C), Common-Unseen (CB-CU), Endangered (CB-E), and Invasive (CB-I)—and three fine-grained benchmarks. Dashes suppress redundant means when species is the only nontrivial rank. Best per column in **bold**.

Arch.	Method	CB-C	CB-CU	CB-E	CB-I	Hel	Ixodidae	Zebra
ViT-L	CLIP	10.8 / 67.9	11.9 / 67.4	16.6 / 57.5	13.2 / 69.8	05.9 / –	00.8 / 12.0	25.7 / –
	OpenCLIP	10.5 / 66.5	11.0 / 66.7	19.0 / 56.3	10.0 / 60.6	03.2 / –	01.4 / 06.3	26.2 / –
	SigLIP-2	22.3 / 72.4	25.3 / 72.7	26.7 / 59.6	26.5 / 75.4	10.3 / –	02.2 / 10.6	28.0 / –
ViT-B	BioCLIP	51.2 / 80.3	51.5 / 79.4	37.4 / 63.3	56.4 / 85.5	48.4 / –	14.6 / 34.1	57.2 / –
	BioTrove-CLIP	44.4 / 78.7	43.6 / 78.0	27.0 / 59.5	42.0 / 76.5	14.4 / –	09.2 / 29.4	20.4 / –
	TaxaBind	54.1 / 81.4	53.7 / 80.4	37.5 / 63.9	59.2 / 86.9	44.9 / –	17.2 / 35.6	58.4 / –
	RCME	54.9 / 82.3	55.2 / 81.4	38.0 / 64.3	61.7 / 88.7	45.2 / –	10.7 / 31.0	48.6 / –
ViT-L	BioCLIP-2	73.8 / 88.3	73.4 / 87.6	40.8 / 65.9	75.9 / 93.7	**51.8** / –	18.6 / 39.2	58.2 / –
ViT-L	TaxaWalk (AR, **ours**)	89.6 / 96.6	92.4 / 96.9	70.1 / 91.4	87.2 / 95.9	50.8 / –	17.3 / 41.1	**64.2** / –
	TaxaWalk (EC, **ours**)	**90.6 / 97.1**	**93.5 / 97.3**	**70.8 / 91.9**	**88.6 / 96.2**	50.8 / –	**27.8 / 46.3**	**64.2** / –

4.3 Expert-Conditioned Identification on Multi-Species Images

Having established image-only recognition, we next ask whether taxonomic context can steer identification when a photograph contains multiple organisms. Existing multi-species benchmarks focus on particular groups of the tree of life (Goeau et al., 2025; Katija et al., 2022; Swanson et al., 2015) and do not test hints at different taxonomic depths. Flat identifiers can use a hint to restrict their candidates, whereas TaxaWalk also conditions its subsequent predictions on it.

iNat Co-Observed. We curate approximately 6K iNaturalist photographs containing two species independently verified by the research community. No photo identifier overlaps TreeOfLife-200M. Each method receives both organisms' true lineage prefixes through a fixed hint rank and returns two complete lineages. Baselines rank all lineages in the hinted subtree, while TaxaWalk uses beam search with width four to return up to two lineages per hint (Appendix B). A rank is correct only when both true taxa appear among the returned lineages. Applying the same hint rank to every photograph prevents it from revealing where the two lineages diverge. Appendix B gives the construction and the full protocol.

Table 4 varies the hint from kingdom to genus. TaxaWalk achieves the highest average and full-lineage accuracy at every hint depth. With a family hint, for example, it reaches 54.7% average and 45.5% full-lineage accuracy, compared with 48.3% and 39.6% for BioCLIP-2. Because every method uses the same hinted subtrees, the improvement shows the benefit of conditioning predictions on the supplied taxonomy rather than using it only to filter candidates. Appendix E.2 separates photographs whose hints already place the two organisms in different subtrees from those that share one prefix; TaxaWalk leads in both cases.

Table 3: **Contributions of TaxaWalk. (a)** Species top-1 accuracy (%) with and without hierarchical candidate pruning on iNat-2021. **(b)** Embedding regression (MSE) versus closed-set cross-entropy (CE) on in-distribution and open-vocabulary recognition. Mean averages all seven ranks; Full-path accuracy (FPA) requires the complete lineage to be correct.

(a) Rank conditioning and hierarchical pruning

Model	Training	Hier. Modeling	AR Conditioning	No pruning	Hier. pruning
BioCLIP-2 (zero-shot)	ToL-200M	×	×	77.87	74.00
Flat species classifier	iNat-2021	×	×	90.45	–
H-CAST (Park et al., 2025)	iNat-2021	✓	×	87.52	88.30
Taxon-SSL (Park et al., 2026)	iNat-2021	✓	×	91.66	92.00
TaxaWalk-MSE (AR, **ours**)	ToL-200M	✓	✓	87.50	89.50
TaxaWalk-MSE (EC, **ours**)	ToL-200M	✓	✓	**92.12**	**94.14**

(b) Embedding regression vs. closed-set CE

Objective	Training Data	Model	Genus	Species	Mean	FPA
iNat-2021						
MSE	iNat-2021	TaxaWalk (EC)	90.0	88.6	95.0	70.1
CE	iNat-2021	TaxaWalk (EC)	**96.3**	**93.2**	**97.7**	**86.5**
$L_{\mathrm{HV}}+\alpha L_{\mathrm{TK}}$	iNat-2021	H-CAST	93.4	87.5	95.4	81.9
CE	iNat-2021	Flat species classifier	95.0	90.5	96.9	–
iNat-2021 (autoregressive)						
CE	iNat-2021	TaxaWalk (AR)	92.2	86.5	95.4	86.5
Rare Species (open-vocabulary)						
MSE	iNat-2021	TaxaWalk (EC)	**75.6**	**46.1**	**71.9**	–
CE	iNat-2021	TaxaWalk (EC)	46.8	0.9	60.7	–

Table 4: **Multi-species identification on iNat Co-Observed with taxonomic hints.** Average rank / full-lineage accuracy (%) below each hint. Both organisms must be correct. Hint depth is fixed within each column, and columns score different sets of ranks. Dashes omit full-lineage scores identical to the average.

	Rank at which the hint stops					
Method	Kingdom	Phylum	Class	Order	Family	Genus
SigLIP-2 (Tschannen et al., 2025)	28.2/0.7	23.0/0.8	17.8/1.3	16.8/2.7	17.3/8.7	34.1/–
BioCLIP (Stevens et al., 2024)	36.1/6.9	32.2/7.7	28.9/9.1	30.9/14.1	33.8/23.7	49.2/–
BioCLIP-2 (Gu et al., 2026)	50.1/21.2	47.8/22.8	45.2/24.9	46.3/30.3	48.3/39.6	62.2/–
TaxaWalk (**ours**)	**51.3/25.0**	**50.6/27.1**	**49.0/29.8**	**50.7/35.5**	**54.7/45.5**	**66.1/–**

4.4 Hierarchical Novel Species Detection

A model deployed for biodiversity monitoring will meet species it was never trained on, and it should flag them rather than force a known label. We test the novelty score of §3.1 (Figure 1c) on Open-Insect (Chen et al., 2026b), which defines a set of known species for each of three regions (Northeast America, Western Europe and Central America) and three tiers of novel species at increasing taxonomic distance: *local* moths from the same region, *non-local* moths from other regions, and *non-moth* arthropods. The known set enters only as a list of names, and no images or detector training are used. We compare against maximum softmax probability (Hendrycks & Gimpel, 2016) on BioCLIP-2, following (Chen et al., 2026b).

Table 5 reports AUROC. At the species rank, TaxaWalk improves on the baseline in every region and tier, with margins of four to five points on the hardest local tier. Because the score exists at every rank, it also shows how far from the known set an organism lies. A non-moth arthropod is already novel at the family rank and is detected there almost perfectly, whereas a new moth species in a known genus is only detected at the species rank. This tells a curator not only that a specimen is new but at which rank it should be placed, which is where an expert's attention is needed.

Table 5: **Zero-shot novel species detection on OPEN-INSECT**. We compare MSP (Hendrycks & Gimpel, 2016) with BioCLIP-2 backbone against TaxaWalk's per-rank novelty score (§3.1). No training or detector fitting is performed on the benchmark. **Bold** marks TaxaWalk's species-rank score, the level directly comparable to the species-only MSP baseline.

	local			non-local			non-moth		
Method	NE-Am.	W-Eur.	C-Am.	NE-Am.	W-Eur.	C-Am.	NE-Am.	W-Eur.	C-Am.
MSP-BioCLIP-2 [species-only]	84.94	88.74	82.89	94.55	94.87	88.28	87.07	88.55	82.98
TaxaWalk (**ours**)									
Family	64.82	63.67	48.84	78.88	77.41	72.60	99.98	99.99	99.97
Genus	71.12	75.86	77.21	92.90	91.67	94.57	99.78	99.83	99.61
Species	**89.19**	**92.73**	**88.37**	**95.20**	**96.05**	**88.32**	**90.60**	**92.49**	**90.24**

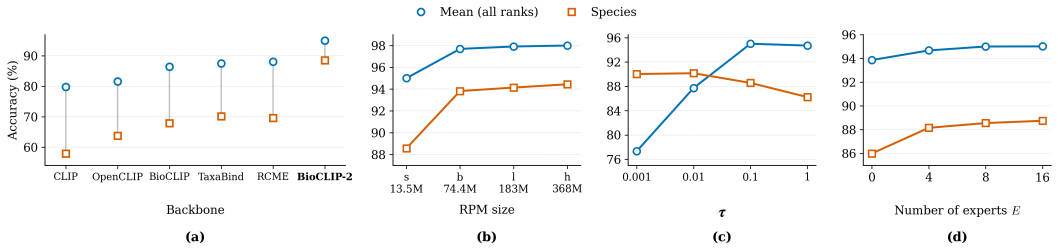

Figure 4: **Ablations** on iNat-2021 under expert-conditioned inference (top-1 %; mean across ranks and species). (a) backbone, (b) RPM size, (c) temperature τ of the rank-balanced objective, (d) MoRE experts E. Anchor: BioCLIP-2, $N = 2$, $D_H = 384$, τ=0.1, E=8.

4.5 ABLATIONS

We ablate TaxaWalk on the closed set of iNat-2021 under expert-conditioned inference, varying one axis at a time around a small anchor model ($N = 2$, $D_H = 384$, $\tau = 0.1$, $E = 8$, BioCLIP-2) and reporting mean and species accuracy (Figure 4). Among six frozen backbones, BioCLIP-2 leads by a wide margin at species, and the web-trained encoders clearly trail the biology-pretrained ones. Scaling the RPM from 13.5M to 368M parameters raises mean accuracy by three points and species accuracy by six, and our main model uses the l preset. As $\tau \to 0$ the objective approaches the worst-rank error and kingdom accuracy collapses, and as τ grows it approaches the uniform mean and species accuracy falls, with mean accuracy peaking at $\tau = 0.1$. MoRE improves on a dense head, with gains levelling off at $E = 8$. We find that the router settles on a fixed pair of experts per rank.

5 CONCLUSION

We introduced TaxaWalk, which treats species identification as a *traversal* of the taxonomy rather than classification over its leaves. Built on biological vision-language models, TaxaWalk predicts lineages rank by rank, conditioning each decision on those above it. It achieves state-of-the-art open-vocabulary species recognition, supports expert-conditioned identification by incorporating known taxonomic ranks into its predictions, and improves on the multi-species images of iNAT CO-OBSERVED, a benchmark we introduce. Its per-rank prediction error also enables zero-shot novelty detection by identifying where an organism departs from the known taxonomy. Future work could incorporate additional context such as location, time, DNA, or audio, and use per-rank uncertainty to guide data collection or expert review. Because candidate taxa are supplied at inference, TaxaWalk could also evaluate revised taxonomies without retraining. More broadly, the same formulation may apply to other hierarchically structured label spaces.

AI USE STATEMENT

We have not used generative AI tools for any task that requires disclosure under the ICLR 2027 AI Policy for Authors. Specifically, we have not used generative AI tools to generate synthetic datasets,

develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, or interpret results. We used generative AI tools to polish the writing of this paper, including correcting grammar and improving readability. The authors reviewed every AI-assisted edit to make sure it did not change the technical content or claims of the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, and we collected no new observations. We curate iNAT CO-OBSERVED from existing public iNaturalist records, retaining only images released under CC0, CC-BY or CC-BY-NC licenses with community-verified species labels. TaxaWalk is trained on TreeOfLife-200M and evaluated on publicly available biodiversity benchmarks, each used under its own license. The model takes only images and taxonomic labels as input and does not use geographic location or observer identity.

The model is meant to assist expert identification, not to replace it. Its predictions should not be the sole basis for consequential decisions such as conservation assessments, invasive-species management or regulatory action, and should never be used to judge whether a mushroom or plant is safe to eat or to use medicinally.

REPRODUCIBILITY STATEMENT

TaxaWalk's architecture and inference procedure are specified in Section 3, and its training objective in Section 3.4, with the gradient analysis that motivates it in Section A. Section 4.1 gives the training data and hyperparameters, together with the evaluation protocol shared by all methods: candidate pools, the records scored at each rank, species resolution, and the autoregressive and expert-conditioned settings. All code, checkpoints and the datasets will be released upon acceptance.

REFERENCES

Sk Miraj Ahmed, Xi Yu, Yunqi Li, Yuewei Lin, and Wei Xu. Hierarchy-guided multimodal representation learning for taxonomic inference. *arXiv preprint arXiv:2603.25573*, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your "Flamingo" is my "Bird": Fine-grained, or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*, volume 2026, pp. 89791–89807, 2026a.

Yuyan Chen, Nico Lang, Brian Schmidt, Aditya Jain, Yves Basset, Sara Beery, Maxim Larrivée, and David Rolnick. Open-insect: benchmarking open-set recognition of novel species in biodiversity monitoring. *Advances in Neural Information Processing Systems*, 38, 2026b.

Shivani Chiranjeevi, Hossein Zaremehrjerdi, Zi K Deng, Talukder Z Jubery, Ari Grele, Arti Singh, Asheesh K Singh, Soumik Sarkar, Nirav Merchant, Harold F Greeney, Baskar Ganapathysubramanian, and Chinmay Hegde. TerraIncognita: A dynamic benchmark for species discovery using frontier models. *arXiv preprint arXiv:2506.03182*, 2025.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL https://arxiv.org/abs/2401.06066.

DeepSeek-AI, Aixin Liu, et al. DeepSeek-V3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research (JMLR)*, 23 (120):1–39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In *European Conference on Computer Vision*, 2022.

Herve Goeau, Vincent Espitalier, Pierre Bonnet, and Alexis Joly. Overview of plantclef 2024: multi-species plant identification in vegetation plot images, 2025. URL https://arxiv.org/abs/2509.15768.

ZeMing Gong, Chuanqi Tang, Xiaoliang Huo, Nicholas Pellegrino, Austin T Wang, Graham W Taylor, Angel X Chang, Scott C Lowe, and Joakim Bruslund Haurum. Hyperbolic multimodal representation learning for biological taxonomies. *arXiv preprint arXiv:2508.16744*, 2025a.

ZeMing Gong, Austin Wang, Xiaoliang Huo, Joakim Bruslund Haurum, Scott C Lowe, Graham W Taylor, and Angel Chang. CLIBD: Bridging vision and genomics for biodiversity monitoring at scale. In *International Conference on Learning Representations*, volume 2025, pp. 92306–92336, 2025b.

Jianyang Gu, Sam Stevens, Elizabeth Campolongo, Matthew Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander White, James Balhoff, et al. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. *Advances in Neural Information Processing Systems*, 38:102778–102811, 2026.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Shyamgopal Karthik, Ameya Prabhu, Puneet K Dokania, and Vineet Gandhi. No cost likelihood manipulation at test time for making better mistakes in deep networks. *arXiv preprint arXiv:2104.00795*, 2021.

Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, et al. Fathomnet: A global image database for enabling artificial intelligence in the ocean: K. katija et al. *Scientific reports*, 12(1):15914, 2022.

Lukas Kuhn, Giuseppe Serra, and Florian Buettner. Non-contrastive vision-language learning with predictive embedding alignment. *arXiv preprint arXiv:2602.00653*, 2026.

Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, and Honglak Lee. Hierarchical novelty detection for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Kun Fan, Huazhen Huang, Qingqing Gu, Yetao Wu, Zhonglin Jiang, Yong Chen, et al. M3-jepa: Multimodal alignment via multi-gate moe based on the joint-embedding predictive architecture. *arXiv preprint arXiv:2409.05929*, 2024.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

Zhenghong Li, Kecheng Zheng, and Haibin Ling. Learning consistent taxonomic classification through hierarchical reasoning. *arXiv preprint arXiv:2601.14610*, 2026.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312.

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

Camilo Mora, Derek P Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. How many species are there on earth and in the ocean? *PLoS biology*, 9(8):e1001127, 2011.

Seulki Park, Youren Zhang, Stella Yu, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification. In *International Conference on Learning Representations*, volume 2025, pp. 20808–20830, 2025.

Seulki Park, Zilin Wang, and Stella X Yu. Free-grained hierarchical visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32767–32776, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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*, 1(3): 166–182, 2025.

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*, pp. 8748–8763. PmLR, 2021.

Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Srikumar Sastry, Aayush Dhakal, Eric Xing, Subash Khanal, and Nathan Jacobs. Global and local entailment learning for natural world imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15770–15780, 2025a.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. TaxaBind: A unified embedding space for ecological applications. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1765–1774. IEEE, 2025b.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1701.06538.

Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1–2):31–72, 2011.

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

Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, 2:150026, 2015.

Zhiyuan Tao, Srikumar Sastry, Matthew J Thompson, Elizabeth G Campolongo, Net Zhang, Ziheng Zhang, Hilmar Lapp, Yu Su, Tanya Berger-Wolf, Nathan Jacobs, et al. Beyond flat labels: Level-restricted contrastive learning for hierarchical fine-grained vision classification. *arXiv preprint arXiv:2606.21838*, 2026.

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.

Jack Valmadre. Hierarchical classification at multiple operating points. *Advances in neural information processing systems*, 35:18034–18045, 2022.

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 *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8769–8778. IEEE, 2018.

Tz-Ying Wu, Chih-Hui Ho, and Nuno Vasconcelos. ProTeCt: Prompt tuning for taxonomic open set classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

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

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2024. URL `https://arxiv.org/abs/2411.04983`.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

A THE RANK-BALANCED OBJECTIVE

Setup and notation. The model is trained to autoregress over the taxonomic hierarchy: given the embedding of a node at rank t, it predicts the embedding of its child at rank $t{+}1$. For a batch of B images and T ranks, of which $T{-}1$ are predicted, let $\hat{z}_{n,t+1} \in \mathbb{R}^d$ denote the predicted embedding of rank $t{+}1$ and $z_{n,t+1} \in \mathbb{R}^d$ the (teacher) target, for sample n and rank $t \in \{1, \dots, T-1\}$. Define the per-pair residual and per-pair error

$$r_{n,t} \;=\; \hat{z}_{n,t+1} - z_{n,t+1}, \qquad e_{n,t} \;=\; \tfrac{1}{d}\,\|r_{n,t}\|_2^2, \tag{7}$$

and the per-rank mean error

$$\bar{e}_t \;=\; \frac{1}{B}\sum_{n=1}^{B} e_{n,t}. \tag{8}$$

The vector $\bar{e} = (\bar{e}_1, \dots, \bar{e}_{T-1})$ collects the average prediction error *at each level of the hierarchy*. All objectives below are functions of $\bar{e}$; only the way the ranks are aggregated differs.

Two aggregations. The naive objective averages the error uniformly over all (sample, rank) pairs,

$$\mathcal{L}_{\text{mean}} \;=\; \frac{1}{B(T-1)}\sum_{n=1}^{B}\sum_{t=1}^{T-1} e_{n,t} \;=\; \frac{1}{T-1}\sum_{t=1}^{T-1}\bar{e}_t, \tag{9}$$

i.e. a fixed, uniform convex combination of the per-rank errors that weights every rank equally regardless of its difficulty. We instead optimize the temperature-τ LogSumExp over ranks, the rank-balanced objective of §3.4,

$$\mathcal{L}_{\text{rb}} \;=\; \tau\,\log\sum_{t=1}^{T-1}\exp\!\big(\bar{e}_t/\tau\big). \tag{10}$$

This objective smoothly interpolates between the two natural extremes,

$$\lim_{\tau\to 0^+}\mathcal{L}_{\text{rb}} = \max_t \bar{e}_t, \qquad \mathcal{L}_{\text{rb}} = \tfrac{1}{T-1}\textstyle\sum_t \bar{e}_t \;+\; \tau\log(T-1) \;+\; \mathcal{O}(\tau^{-1}) \quad \text{as } \tau\to\infty, \tag{11}$$

recovering the (uniform) mean over ranks up to an additive constant. Crucially, $\mathcal{L}_{\text{rb}}$ is a tight, everywhere-differentiable upper bound on the worst-rank error,

$$\max_t \bar{e}_t \;\le\; \mathcal{L}_{\text{rb}} \;\le\; \max_t \bar{e}_t + \tau\log(T-1), \tag{12}$$

so minimizing equation 10 minimizes the largest per-rank error to within a controllable slack $\tau\log(T-1)$.

Gradient analysis. Both objectives are functions of $\bar{e}$, so by the chain rule the parameter gradient is a weighted sum of the per-rank error gradients,

$$\nabla_\theta\mathcal{L} \;=\; \sum_{t=1}^{T-1}\alpha_t\,\nabla_\theta\bar{e}_t, \qquad \nabla_\theta\bar{e}_t \;=\; \frac{1}{B}\sum_n \frac{2}{d}\,J_{n,t}^\top r_{n,t}, \tag{13}$$

where $J_{n,t} = \partial\hat{z}_{n,t+1}/\partial\theta$ is the Jacobian of the prediction head and shared backbone. The two objectives differ *only* in the rank-mixing coefficients α_t:

$$\alpha_t^{\text{mean}} \;=\; \frac{1}{T-1}, \qquad \alpha_t^{\text{rb}} \;=\; \frac{\exp(\bar{e}_t/\tau)}{\sum_{t'}\exp(\bar{e}_{t'}/\tau)} \;=\; \big[\mathrm{softmax}(\bar{e}/\tau)\big]_t. \tag{14}$$

This is the crux of the analysis. Under the mean objective the gradient mass allocated to rank t is pinned to a uniform $1/(T-1)$ and is *independent of how hard that rank currently is*. Under the rank-balanced objective the gradient mass is the softmax of the current per-rank errors: the rank with the largest residual receives the largest share, and the allocation *adapts as training proceeds*. The relative emphasis the two objectives place on the hardest versus easiest rank is, respectively,

$$\frac{\alpha_{\max}^{\text{mean}}}{\alpha_{\min}^{\text{mean}}} = 1, \qquad \frac{\alpha_{\max}^{\text{rb}}}{\alpha_{\min}^{\text{rb}}} = \exp\!\Big(\tfrac{\bar{e}_{\max}-\bar{e}_{\min}}{\tau}\Big), \tag{15}$$

so the rank-balanced objective amplifies the gradient on the hardest rank *exponentially* in its error gap, with τ setting the sharpness.

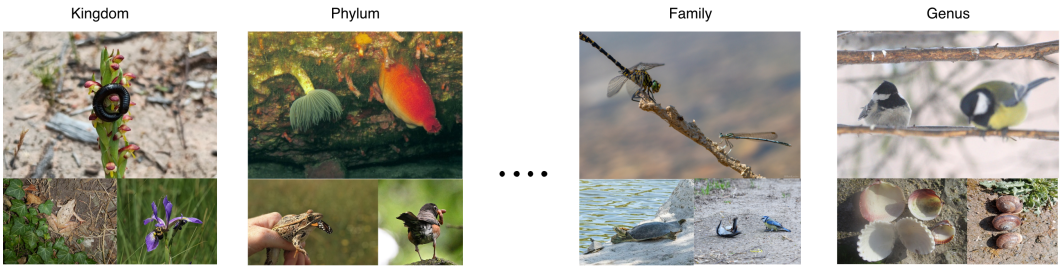

Figure 5: **iNat Co-Observed examples.** Photographs whose paired species lineages first diverge at kingdom, phylum, family or genus. The benchmark is stratified across divergence ranks from kingdom to genus.

Why this matters: gradient starvation across the hierarchy. The taxonomic ranks are far from equally difficult. Shallow ranks (kingdom, phylum) are coarse, low-entropy, and quickly driven to near-zero error; deep ranks (genus, species) are fine-grained, high-entropy, and carry most of the residual error throughout training. Three consequences follow from equation 14.

(i) The mean objective starves the hard ranks. Because every rank is weighted by the same $1/(T-1)$ regardless of difficulty, the easy ranks that saturate early ($\bar{e}_t \downarrow 0$) still occupy a fixed share of equation 13, and the shared backbone is updated by a gradient in which the few hard ranks are merely a $1/(T-1)$ share of an *equally weighted* sum. Hard ranks never receive emphasis proportional to their error, so they are under-optimized—a per-rank instance of gradient starvation in multi-task learning.

(ii) The rank-balanced objective is self-balancing. By equation 14, α_t^{rb} rises automatically for any rank whose error grows relative to the others and decays once that rank catches up; at the fixed point where all per-rank errors are equal, the softmax weights collapse to the uniform $1/(T-1)$ and equation 10 reduces to the mean equation 11. The objective thus spends gradient budget precisely where the model is currently weakest and stops as soon as the ranks equalize, without any hand-tuned per-rank loss weights.

(iii) Smoothness avoids the pathologies of the hard max. The non-smooth alternative $\max_t \bar{e}_t$ routes the entire gradient to a single rank per step equation 11, producing sparse, oscillatory updates that flip between ranks. The LogSumExp in equation 10 keeps the gradient dense—every rank retains a strictly positive weight—while still concentrating mass on the hardest ranks, and τ smoothly trades off this concentration against uniformity. This worst-case-aware but smooth behavior is exactly what a hierarchical model requires: it must reason coherently at *every* level of the taxonomy, since a single poorly-modeled rank breaks the autoregressive chain that links coarse to fine predictions.

B iNat Co-Observed: Construction and Protocol

Source labels. iNaturalist stores one taxon per observation. When a photograph contains two organisms, an observer may create two observations that share the same photo identifier, and the community identifies each independently. We retain photographs for which both observations reach research grade, giving two community-verified species labels for the same pixels. To our knowledge, this structure has not previously been used to construct a multi-species recognition benchmark across divergent species lineages.

Construction. We identify iNaturalist photos linked to two independently research-grade species observations with complete lineages, then retain openly licensed images from Animalia, Plantae and Fungi. We remove every photo identifier found in TreeOfLife-200M, so no benchmark image appears in the training corpus. To balance task difficulty, we sample approximately $1{,}000$ photos at each divergence rank from kingdom through genus, round-robin over kingdom and class pairs and with caps per species and observer. Species-level divergence is excluded because the hint would reveal the target. The resulting benchmark contains $5{,}996$ photographs, each with two identified organisms. Table 6 gives the full composition and filtering details.

Table 6: **iNat Co-Observed composition and provenance.** Counts after licensing, lineage-completeness filtering, TreeOfLife-200M overlap removal and stratified sampling.

Photographs	5,996
Test pairs (two per photograph)	11,992
Per divergence rank (kingdom … genus)	1000 / 1000 / 998 / 999 / 1000 / 999
Distinct species	6,557
Distinct class pairs	561
Candidate gallery	14,162 lineages (iNat-2021 ∪ benchmark)

Evaluation protocol. Every method receives both organisms' true lineages through a fixed hint rank, the same for every photograph so that the hint cannot reveal where the two lineages diverge, and returns two complete lineages. All methods draw candidates from one gallery, formed by the iNat-2021 training species and the benchmark species, restricted to the hinted subtrees. TaxaWalk conditions on each hint and rolls out the remaining ranks, while the baselines score complete lineage strings within the restricted gallery. A rank is correct only when both true taxa appear among the returned lineages. Each rank is scored only where it lies below the hint, and we report the unweighted mean over those ranks together with full-lineage accuracy, which counts a photograph only when every rank below the hint is correct. All methods use six views per image, the full image, four quadrants and a centre crop, merged by the maximum score per lineage. Hint depth varies from kingdom to genus (Table 4).

Decoding. We use each model's native decoder. Contrastive baselines rank all gallery lineages exactly with one image–text matrix product. Because TaxaWalk scores lineages autoregressively, exhaustive search is intractable and we use beam search with width four. Beam search is also required to return two lineages from a shared hint; with two hints, each contributes its best lineage. We score the baselines on complete lineage strings, their stronger native setting. Applying width-four beam search to truncated prefixes lowers BioCLIP-2's full-lineage accuracy by 20–27 points (Appendix E.3).

Limitations. Labels reflect iNaturalist's research-grade consensus at the export date and may change. We do not automatically verify that every photograph contains two distinct individuals, so rare pairs may instead reflect conflicting identifications of one organism. The second organism is often small, distant or occluded, which is part of the task. Finally, 1,390 photographs share an observation pair with another photograph. The release identifies these groups so users can restrict evaluation to one photo per pair.

C Detecting and Localising Misidentified Records

Taxonomic records may be correct at coarse ranks but wrong below a particular point in the lineage. We evaluate whether a model can both detect an inconsistent species claim and identify the first rank where the claimed lineage conflicts with the image. TaxaWalk scores each claimed rank conditioned on the image and the preceding claimed ancestors, while the contrastive baselines score ranks independently.

Protocol. We create synthetic label errors by corrupting half of the 11,983 Rare Species records. A corrupted record is assigned another species and that species' complete lineage. The divergence rank is the first rank where the assigned and true lineages differ. In the CONGENER setting, the replacement belongs to the true genus, so only the species is wrong. In the RANDOM setting, the replacement can come from anywhere in the taxonomy. We report detection AUROC over all records and divergence-rank accuracy over corrupted records. There are six possible divergence ranks, giving 16.7% chance accuracy. Results are averaged over three corruption seeds, with AUROC varying by at most ± 0.007. These synthetic corruptions do not necessarily represent the true frequency or distribution of real identification errors. Since Rare Species and TreeOfLife-200M draw from the same data sources, this experiment measures localisation under controlled inconsistencies.

Table 7: **Misidentified-record detection and localisation** on Rare Species (11,983 records, half corrupted, mean of three seeds). CONGENER reassigns within the true genus; RANDOM reassigns anywhere. Localisation is the accuracy of the predicted divergence rank over corrupted records (chance 16.7%); the whole-lineage variant scores one cosine against the entire claimed path and cannot localise. Best per column in **bold**.

		Detection AUROC		Localisation (%)		Whole-lineage AUROC	
Arch.	Method	Congener	Random	Congener	Random	Congener	Random
ViT-L	CLIP	0.778	0.931	18.4	1.9	0.763	0.909
	OpenCLIP	0.746	0.887	17.0	5.0	0.700	0.821
	SigLIP-2	0.819	0.970	19.0	0.8	0.807	0.969
ViT-B	BioCLIP	0.794	0.961	15.6	1.7	0.788	0.960
	BioTrove-CLIP	0.650	0.696	17.3	2.5	0.626	0.657
	TaxaBind	0.794	0.962	16.1	1.5	0.787	0.959
	RCME	0.812	**0.973**	20.5	0.3	0.802	0.977
ViT-L	BioCLIP-2	0.773	0.913	6.9	3.2	0.842	0.972
ViT-L	TaxaWalk (AR, unconditioned)	0.809	0.916	31.3	18.2	–	–
	TaxaWalk (claim-conditioned, **ours**)	**0.828**	0.921	**70.0**	**70.9**	–	–

For a claimed label c_r at rank r, we measure its inconsistency as

$$m_r = d(\hat{e}_r, t_{c_r}) - \min_{c \in \mathcal{V}_r} d(\hat{e}_r, t_c), \tag{16}$$

where $\mathcal{V}_r$ is the candidate set at that rank. The margin is zero when the claimed label is the nearest candidate and increases as a different label becomes more plausible. We use $\max_r m_r$ for detection and $\arg\max_r m_r$ for localisation. For TaxaWalk, $\hat{e}_r$ is predicted from the image and the preceding claimed ancestors. For the contrastive baselines, it is the image embedding at every rank. We also evaluate a whole-lineage score for detection, which produces no rank-level localisation.

Detection. TaxaWalk is competitive at detecting corrupted records but is not the strongest method. On RANDOM corruptions, RCME reaches 0.977 whole-lineage AUROC compared with TaxaWalk's 0.921. On CONGENER corruptions, BioCLIP-2 reaches 0.842 whole-lineage AUROC compared with TaxaWalk's 0.828. These results show that rank-wise conditioning is not necessary for detecting an inconsistent lineage. Its benefit appears in locating where the inconsistency begins.

Localisation. TaxaWalk localises the divergence rank with 70.0% accuracy on CONGENER corruptions and 70.9% on RANDOM corruptions. No contrastive baseline exceeds 20.5% and their accuracy on RANDOM corruptions ranges from 0.3% to 5.0%, below the 16.7% chance rate. Once a claimed lineage diverges, every lower rank is also incorrect. Independent rank scores therefore tend to select a later mismatch rather than the first one.

Conditioning effect. Using TaxaWalk's unconditioned rollout reduces localisation from 70.0% to 31.3% on CONGENER corruptions and from 70.9% to 18.2% on RANDOM corruptions. Detection AUROC changes only from 0.828 and 0.921 to 0.809 and 0.916. With conditioning, the first incorrect rank is predicted from the correct prefix and receives a large margin. At subsequent ranks, the claimed ancestors move the prediction into the claimed subtree, preventing the initial error from being counted again at every lower rank. The largest margin therefore identifies the first divergence more reliably.

D ARCHITECTURE DETAILS

Recurrent Predictive Module. Two-layer MLPs project the label and image embeddings to dimension D_H. We define $x_t = g_{\text{in}}(z_t)$ and $c_t = g_{\text{cond}}(\mathbf{v}) + R_t$, where R_t is a learned rank embedding. For one RPM cell, with the cell index omitted, the hidden state is initialised to zero and updated as

$$\begin{aligned} u_t &= \sigma\big(W_x^u x_t + W_h^u h_{t-1}\big), & r_t &= \sigma\big(W_x^r x_t + W_h^r h_{t-1}\big), \\ \tilde{h}_t &= \mathrm{TBlock}\big(x_t,\ r_t \odot h_{t-1},\ c_t\big), & h_t &= (1 - u_t)\, h_{t-1} + u_t\, \tilde{h}_t. \end{aligned} \tag{17}$$

Table 8: RPM size presets. The main results use the l preset.

preset	cells N	hidden D_H	heads
s	2	384	8
b	4	768	12
l	6	1024	16
h	8	1280	16

The transformer block computes the candidate state from the current input, gated memory and image-rank context. We stack N cells, passing the hidden-state sequence from each cell to the next. The output head reads the final cell's state. Table 8 lists the four model sizes.

Mixture of rank experts. Let $h_{b,t} \in \mathbb{R}^{D_H}$ be the last cell's state for image b at rank t. A shared router averages these states across the batch separately at each rank. With parameters $W_g \in \mathbb{R}^{E \times D_H}$ and $b_g \in \mathbb{R}^E$,

$$\bar{h}_t = \tfrac{1}{B} \textstyle\sum_{b=1}^{B} h_{b,t}, \tag{18}$$

$$a_t = W_g \bar{h}_t + b_g, \tag{19}$$

$$g_{t,e} = \sigma(a_{t,e}), \tag{20}$$

Each gate is an independent sigmoid (DeepSeek-AI et al., 2024), and the router makes one decision per rank. We maintain a balancing bias $\beta \in \mathbb{R}^E$ that receives no gradient. Experts are selected using the biased gates but weighted using the unbiased gates,

$$S_t = \text{Top-}k\big(g_{t,\cdot} + \beta\big), \qquad |S_t| = k, \tag{21}$$

$$w_{t,e} = g_{t,e} \Big/ \textstyle\sum_{e' \in S_t} g_{t,e'} \quad \text{for } e \in S_t. \tag{22}$$

Thus, β affects expert selection but not their output weights. After each training step, it is updated toward uniform load. With $n_e = \sum_t \mathbb{1}[e \in S_t]$ and $\bar{n} = kT/E$,

$$\beta_e \leftarrow \beta_e - \eta \,\text{sign}(n_e - \bar{n}), \tag{23}$$

where $\eta = 10^{-3}$. No balancing term enters the training loss. We also apply a router z-loss (Zoph et al., 2022) with weight 10^{-3} to limit the router logits under mixed-precision training.

E Additional Experiments

E.1 Additional Qualitative Results

Figure 6 repeats the analysis of Figure 3 for iNat-2021 validation images of *Lynx rufus*.

E.2 Distinct versus Shared Hints on iNat Co-Observed

Section 4.3 applies the same hint rank to every photograph, so whether the two organisms receive different hints depends on where their lineages diverge. At a fixed hint depth, the two organisms may receive distinct prefixes if their lineages have already diverged, or share one prefix if they diverge below the hint. Table 9 separates these cases. TaxaWalk leads under both conditions at every available hint depth, showing that its advantage does not depend on the hints already separating the organisms.

E.3 Baselines under Beam Search

TaxaWalk's width-four beam explores multiple autoregressive lineage continuations, allowing it to return two lineages from a shared hint. Contrastive baselines instead retrieve by ranking complete lineage strings. To test whether giving them the same search procedure changes the comparison, we decode BioCLIP-2 with a width-four beam over truncated prefixes while holding the hint, gallery,

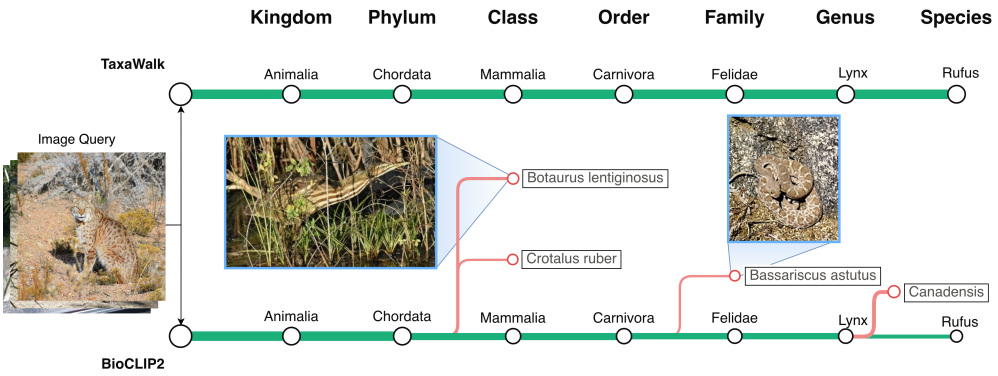

Figure 6: **Lineage predictions for *Lynx rufus*.** Predictions on iNat-2021 validation images. Green follows the true lineage, with line width indicating the number of images that remain correct. Red branches mark the first incorrect rank and are labelled with the resulting species prediction. Insets show the incorrectly predicted species. TaxaWalk retains the true lineage through species, while BioCLIP-2 diverges at class, family or species.

Table 9: **Effect of whether the hint separates the two organisms.** Average rank / full-lineage accuracy (%) below each hint. Distinct hints place the organisms in different subtrees, while a shared hint requires one prefix to yield both lineages. Dashes mark unavailable or duplicate scores.

	Rank at which the hint stops					
Method	Kingdom	Phylum	Class	Order	Family	Genus
Distinct hints (different subtrees)						
SigLIP-2 (Tschannen et al., 2025)	16.2/0.4	14.8/0.5	13.6/1.1	15.5/2.8	19.0/9.2	34.1/–
BioCLIP (Stevens et al., 2024)	27.7/8.5	25.2/7.9	25.0/9.1	30.9/15.0	36.0/24.8	49.2/–
BioCLIP-2 (Gu et al., 2026)	48.9/28.2	42.9/24.2	43.9/26.9	46.8/31.5	50.7/41.2	62.2/–
TaxaWalk (**ours**)	**53.5/32.2**	**46.6/26.7**	**46.0/29.6**	**50.9/36.3**	**57.0/47.1**	**66.1/–**
Shared hint (same subtree)						
SigLIP-2 (Tschannen et al., 2025)	30.8/0.8	27.4/0.9	22.1/1.5	19.7/2.4	8.6/6.0	–
BioCLIP (Stevens et al., 2024)	38.0/6.5	35.9/7.6	33.1/9.0	30.9/12.0	23.0/18.3	–
BioCLIP-2 (Gu et al., 2026)	50.3/19.7	50.4/22.0	46.5/22.9	45.4/27.9	36.2/31.9	–
TaxaWalk (**ours**)	**50.8/23.4**	**52.7/27.3**	**52.0/30.0**	**50.3/33.7**	**43.4/38.0**	–

views and output budget fixed. Table 10 shows that complete-lineage retrieval remains stronger by 20 to 27 points of full-lineage accuracy from kingdom through family. The two decoders agree at genus, where only species remains. We therefore use each model family's stronger decoding procedure in the main comparison.

E.4 Few-Shot Learning

TaxaWalk classifies each rank by nearest neighbour against a candidate bank supplied at inference, allowing labelled support images to be incorporated without additional training. For every taxon at every rank, we average TaxaWalk's predicted embeddings of its support images, conditioned on their known lineages, rescale the resulting prototype p to the norm of the corresponding text embedding t, and blend the two as $c = (1-w)t + wp$. Queries are then classified image-only using the standard autoregressive rollout. The contrastive baselines use the analogous prototype adapter, where p_s is the mean support-image embedding for each species; $w = 0$ therefore recovers the zero-shot model. We draw $k \in \{1, 5\}$ support images per species, evaluate on the remaining images, and average over three support/query splits across the seven benchmarks in Table 2. All methods use identical species, support sets, and query sets, and w is selected once for each method and value of k according to its average performance across the seven benchmarks.

Few-shot adaptation improves nearly all models (Table 11), with the largest gains occurring for models with weak zero-shot text priors. CLIP and OpenCLIP improve by approximately 14 points with

Table 10: **BioCLIP-2 decoding on iNat Co-Observed.** Average rank / full-lineage accuracy (%) using exhaustive retrieval of complete lineages or width-four beam search over truncated prefixes. All other settings match Table 4. Dashes omit duplicate genus-hint scores.

	Rank at which the hint stops					
Decoder	Kingdom	Phylum	Class	Order	Family	Genus
Complete-lineage retrieval	**50.1/21.2**	**47.8/22.8**	**45.2/24.9**	**46.3/30.3**	**48.3/39.6**	**62.2/–**
Prefix beam search	0.4/0.0	2.2/0.1	3.1/0.3	8.4/3.0	23.4/19.5	**62.2/–**

Table 11: **Few-shot species recognition** (species top-1 accuracy, %, mean over three support/query splits shared by all methods). Baselines blend each species' text embedding with the mean image embedding of its k supports, $c_s = (1-w)\,t_s + w\,p_s$; TaxaWalk applies the same blend at every rank to its own predicted embeddings of the supports and classifies queries image-only by its rollout. w is chosen once per method and k by the seven-dataset average, Δ is the gain over zero-shot ($w = 0$), bold marks the best image-only result, † receives the true ancestors, and FGC is Fine-Grained Classification.

			CrypticBio				FGC				
Arch.	Method	w	Common	Unseen	Endang.	Invasive	Ixodidae	Zebra	Heliconius	Avg.	Δ
$k = 1$ (1 support image per species)											
ViT-L	CLIP	0.50	18.4	19.7	43.1	21.7	16.0	46.4	30.8	28.0	+14.4
	OpenCLIP	0.50	19.3	20.2	40.7	21.1	16.6	44.9	33.5	28.0	+14.2
	SigLIP-2	0.25	27.0	29.4	47.4	29.5	15.6	44.1	24.0	31.0	+7.3
ViT-B	BioCLIP	0.00	58.6	59.8	63.2	57.2	14.6	57.0	48.7	51.3	+0.0
	BioTrove-CLIP	0.50	53.2	51.6	53.7	49.6	19.4	36.1	37.2	43.0	+10.0
	TaxaBind	0.25	56.9	57.5	66.2	55.8	26.8	60.1	50.4	53.3	+0.9
	RCME	0.25	57.2	57.5	66.1	56.7	24.5	52.2	49.6	52.0	+0.9
ViT-L	BioCLIP-2	0.25	85.4	84.9	75.4	78.0	29.0	62.0	58.0	67.5	+4.7
ViT-L	TaxaWalk (AR, **ours**)	0.50	**86.5**	**86.5**	**78.0**	**79.5**	**32.0**	**75.0**	**60.0**	**71.1**	+5.5
	TaxaWalk (EC, **ours**)†	0.50	91.5	94.0	82.0	88.0	35.0	75.0	60.0	75.1	+4.7
$k = 5$ (5 support images per species)											
ViT-L	CLIP	1.00	31.1	33.2	58.2	36.5	21.4	65.0	54.7	42.9	+29.3
	OpenCLIP	1.00	32.5	34.1	55.4	35.8	23.0	62.5	57.4	43.0	+29.2
	SigLIP-2	0.50	37.5	40.0	59.9	41.9	21.9	67.7	48.4	45.3	+21.6
ViT-B	BioCLIP	0.50	61.2	61.7	71.9	60.3	33.5	62.2	64.2	59.3	+7.8
	BioTrove-CLIP	1.00	66.2	66.6	67.9	63.9	30.8	47.7	59.4	57.5	+24.3
	TaxaBind	0.50	64.1	64.6	73.9	63.3	35.5	68.9	64.9	62.2	+9.5
	RCME	0.50	65.2	65.7	73.7	65.0	33.8	62.4	62.9	61.2	+10.1
ViT-L	BioCLIP-2	0.50	89.0	88.8	85.1	84.0	42.5	73.9	68.0	75.9	+12.9
ViT-L	TaxaWalk (AR, **ours**)	0.50	**89.5**	**89.5**	**86.0**	**84.5**	**43.0**	**80.0**	**69.0**	**77.4**	+11.6
	TaxaWalk (EC, **ours**)†	0.75	92.5	94.5	88.0	89.0	46.0	80.0	69.0	79.9	+9.2

one support image and 29 points with five. TaxaWalk also benefits substantially while starting from a much stronger zero-shot representation. Its autoregressive variant reaches 71.1% average accuracy at k=1 and 77.4% at k=5, compared with 67.5% and 75.9% for BioCLIP-2. This corresponds to gains of $+5.5$ and $+11.6$ points over TaxaWalk's own zero-shot performance. TaxaWalk exceeds BioCLIP-2 on all seven benchmarks at both k=1 and k=5, with particularly large improvements on Zebra (75.0 vs. 62.0 at k=1) and consistent gains on the four CrypticBio splits.

Providing the true ancestors further improves few-shot recognition. TaxaWalk (EC) reaches 75.1% and 79.9% average accuracy at k=1 and k=5, respectively, corresponding to gains of $+4.7$ and $+9.2$ points over its zero-shot counterpart. The improvement under both AR and EC indicates that the few-shot gains arise from adapting the candidate representations themselves rather than simply correcting errors accumulated during autoregressive rollout. Overall, these results show that TaxaWalk can incorporate a small number of labelled examples at inference time while retaining its hierarchical prediction mechanism and without updating model parameters.

F LIMITATIONS

TaxaWalk operates in a frozen text embedding space and inherits its weaknesses. Where BioCLIP-2 struggles, as on plankton, so does TaxaWalk. Training uses teacher forcing while inference is autoregressive, so the model is subject to exposure bias, and because each decoded rank restricts the candidates below it, errors at coarse ranks propagate downward and cannot be revised. The gap between AR and EC indicates the model's sensitivity to errors in the predicted ancestors. Finally, the novelty score ranks specimens for expert review rather than deciding on its own, and a large distance can reflect a poor image as well as a new taxon.