LEARNING PHYLOGENY-CONSTRAINED RELATIVE COMPETITION FOR HIERARCHICAL VISUAL RECOGNITION

Anonymous authors
Paper under double-blind review

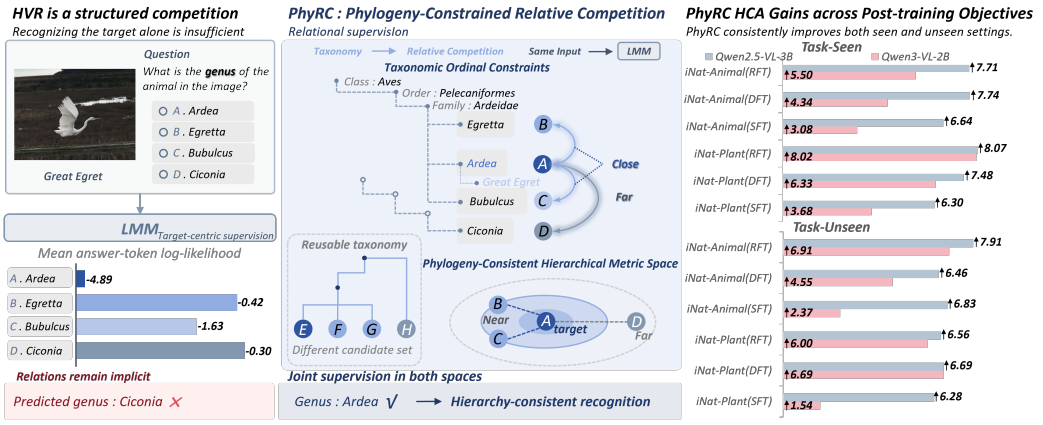


Figure 1: **PhyRC learns relative competition among candidate categories from their taxonomic relations.** It jointly imposes taxonomic ordinal constraints on candidate scores and phylogeny-consistent constraints on visual representations. PhyRC consistently improves HVR across post-training objectives, LMM backbones, and both Task-Seen and Task-Unseen categories.

ABSTRACT

Hierarchical visual recognition (HVR) with large multimodal models (LMMs) aims to learn taxonomy-consistent representations. However, current target-centric post-training of LMMs only supervises the correct candidate, making inter-category discrimination hard to achieve. To bridge this gap, we propose **PhyRC** (**Phy**logeny-Constrained **R**elative **C**ompetition), which recasts the supervision from category identities to relative phylogenetic relations among competing candidates. Specifically, PhyRC introduces taxonomic ordinal constraints that encode discrete taxonomy as multi-level relative separation, requiring stronger separation for larger phylogenetic distances. These constraints further induce a phylogeny-consistent hierarchical metric space, where visual distances preserve taxonomic relatedness. Together, they enable the model to learn a transferable relational structure from phylogeny, allowing generalization to categories unseen during task-specific post-training. Extensive experiments across multiple LMM architectures, post-training objectives, and HVR benchmarks demonstrate consistent gains in hierarchical recognition and taxonomic consistency, including on Task-Unseen categories. Code and configurations are provided in the anonymized supplementary material PhyRC.

1 INTRODUCTION

Hierarchical visual recognition (HVR) with large multimodal models (LMMs) (Bai et al., 2025a;b; Gemma Team, 2025; Chen et al., 2025; Li et al., 2024) concerns the hierarchical understanding of visual categories, requiring models to recognize categories across multiple taxonomic levels while

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.

preserving their hierarchical relationships (Chang et al., 2021; Chen et al., 2022; Jiang et al., 2024; Park et al., 2025). Consider a genus-level query for a Great Egret (*Ardea alba*), where the correct candidate *Ardea* competes with *Egretta*, *Bubulcus*, and *Ciconia*. Although all competitors are incorrect, they differ in their taxonomic proximity to the target, making the candidate set inherently structured rather than flat. Thus, identifying the correct category alone does not capture the relations that organize the competing candidates. HVR requires both category discrimination and taxonomy-aware relational understanding (Silla Jr & Freitas, 2011; Kosmopoulos et al., 2015).

Common LMM post-training objectives such as SFT, DFT, and RFT primarily supervise the target response (Ouyang et al., 2022; Wu et al., 2026; Liu et al., 2025b). Such supervision identifies the correct answer, but does not explicitly use the taxonomic relations among competing candidates to structure the recognition decision. Although hierarchy-aware methods have incorporated taxonomy into learning objectives and representations (Bertinetto et al., 2020; Garg et al., 2022; He et al., 2026b;a), how these relations should directly govern competition among candidates remains underexplored. This motivates us to view HVR through the lens of relative candidate competition, where taxonomic relations determine how strongly the target should be separated from different competitors.

To this end, we propose **PhyRC** (**Phy**logeny-Constrained **R**elative **C**ompetition), a structured post-training framework that recasts supervision from category identities to relative taxonomic relations among competing candidates. Given an image, a hierarchy-specific question, and a candidate set, PhyRC jointly scores all candidates and introduces *taxonomic ordinal constraints* that translate their taxonomic relatedness to the target into different separation requirements, imposing larger margins on more distantly related competitors. PhyRC further extends the same relational principle to visual representation learning by imposing taxonomy-aware relative-distance constraints, inducing a *phylogeny-consistent hierarchical metric space* in which visual distances reflect taxonomic proximity. Because these constraints depend on inter-category relations rather than fixed category identities, the same relational supervision can be instantiated across different category combinations, enabling evaluation beyond categories observed during task-specific post-training. PhyRC can be integrated with SFT, DFT, and RFT as additional structured supervision without changing the underlying recognition task.

We evaluate the effectiveness and generality of PhyRC through extensive experiments across multiple LMM backbones, post-training objectives, and biological hierarchies. Results show consistent improvements in hierarchical recognition and taxonomic consistency under SFT, DFT, and RFT on iNaturalist-2021 (Van Horn et al., 2021), with the gains further extending to Task-Unseen categories and NABirds (Van Horn et al., 2015). These results demonstrate the effectiveness, generality, and transferability of PhyRC for learning taxonomy-aware relational structure in LMMs.

Our contributions are summarized as follows:

- **Category-Transferable Phylogeny-Structured Learning.** We formulate HVR supervision around phylogenetic relations rather than fixed category identities, allowing the same relational structure to be instantiated across different category combinations and to extend beyond categories observed during task-specific post-training.
- **Taxonomic Ordinal Constraints.** We translate the discrete taxonomic hierarchy into multi-level ordinal separation constraints over competing category hypotheses, requiring progressively stronger target–competitor separation as their taxonomic relatedness decreases.
- **Phylogeny-Consistent Hierarchical Metric Space.** We extend the same relational ordering to visual representation learning, encouraging feature distances to reflect taxonomic proximity in a hierarchical metric space.

2 Related Work

Hierarchical Visual Recognition. Hierarchical visual recognition jointly considers fine-grained category identification and predictions across multiple levels of a class taxonomy (Silla Jr & Freitas, 2011; Kosmopoulos et al., 2015; Deng et al., 2014). A substantial body of work exploits category hierarchies through hierarchy-aware objectives, structured labels, and losses that distinguish errors

according to their semantic or taxonomic severity (Bertinetto et al., 2020; Garg et al., 2022; Chen et al., 2022; Jiang et al., 2024; Park et al., 2025). More recent vision–language studies extend hierarchical recognition to flexible semantic granularities and taxonomy-aware evaluation (Snæbjarnarson et al., 2025; Park et al., 2026). With LMMs, recent work further reveals persistent inconsistencies across taxonomic levels and incorporates explicit taxonomic knowledge into multimodal learning (Tan et al., 2026; He et al., 2026b;a). Building on this setting, we study how taxonomic relations among candidate categories can directly structure their competition at a given recognition level.

Hierarchy-Aware Representation Learning. Hierarchical structure has long been incorporated into visual recognition through hierarchy-aware losses, structured label embeddings, and multi-level feature supervision (Bertinetto et al., 2020; Garg et al., 2022; Zhang et al., 2016). Related work models semantic order explicitly through ordered embeddings and hierarchical label–image relations (Vendrov et al., 2016; Dhall et al., 2020). Graph-based methods further propagate relational information over class hierarchies (Xia et al., 2025). Another line of work exploits non-Euclidean geometry to represent tree-like or entailment structure, ranging from Poincaré embeddings and entailment cones to hyperbolic visual and vision–language representations (Nickel & Kiela, 2017; Dhall et al., 2020; Desai et al., 2023; Kwon et al., 2024; Pal et al., 2025; Yoshikawa & Matsubara, 2026). Other approaches explicitly organize visual representations across semantic granularities or category levels (Sani & Anand, 2026; Wang et al., 2025). More recently, hierarchy-aware LMM methods inject taxonomic structure through representation alignment and hierarchical regularization (He et al., 2026b;a). In contrast, PhyRC uses taxonomic relations primarily to structure relative competition among generative candidate scores, with visual relational constraints serving as complementary supervision.

Learning to Rank and Preference Optimization. Learning-to-rank methods model relative preferences among candidates through pairwise and listwise objectives (Burges et al., 2005; Cao et al., 2007; Xia et al., 2008). Recent language-model alignment methods similarly optimize preferences among multiple responses using ranking or preference-based objectives (Rafailov et al., 2023; Yuan et al., 2023; Liu et al., 2025a; Zhao et al., 2023; Gheshlaghi Azar et al., 2024; Meng et al., 2024). These ideas have also been extended to multimodal models through vision-conditioned preference learning and feedback-based alignment (Sun et al., 2024; Zhou et al., 2024; Yu et al., 2025; Pesaran Zadeh et al., 2025). While these methods learn generic response preferences, PhyRC uses the known taxonomic relations among candidates to structure relative competition for hierarchical recognition.

3 METHOD

PhyRC recasts supervision from category identities to relative taxonomic relations among competing candidates. It operationalizes these relations through two complementary objectives: *taxonomic ordinal constraints* that impose relation-dependent separation among candidate scores, and a *phylogeny-consistent hierarchical metric space* that preserves taxonomic relatedness in visual representations through phylogenetic near–far constraints. Together, these objectives encourage the model to internalize a transferable relational structure rather than relying solely on category-specific supervision. Figure 2 provides an overview of the framework.

3.1 PROBLEM FORMULATION

Given an image x, a query q_ℓ at taxonomic level ℓ, and a same-level candidate set $\mathcal{C}_\ell = \{c_1, \dots, c_M\}$ containing the target c_y, the model selects the candidate with the highest generative score, $\hat{c}_\ell = \arg\max_{c_j \in \mathcal{C}_\ell} s_j$. The candidate set is part of the input and may vary across queries. For the answer-token sequence $a_j = (a_{j,1}, \dots, a_{j,T_j})$ of category c_j, we use the length-normalized log likelihood

$$s_j = \frac{1}{T_j} \sum_{t=1}^{T_j} \log p_\theta(a_{j,t} \mid x, q_\ell, \mathcal{C}_\ell, a_{j,<t}) \,. \tag{1}$$

All candidates are scored under the same multimodal context, and length normalization accounts for differences in category-name length.

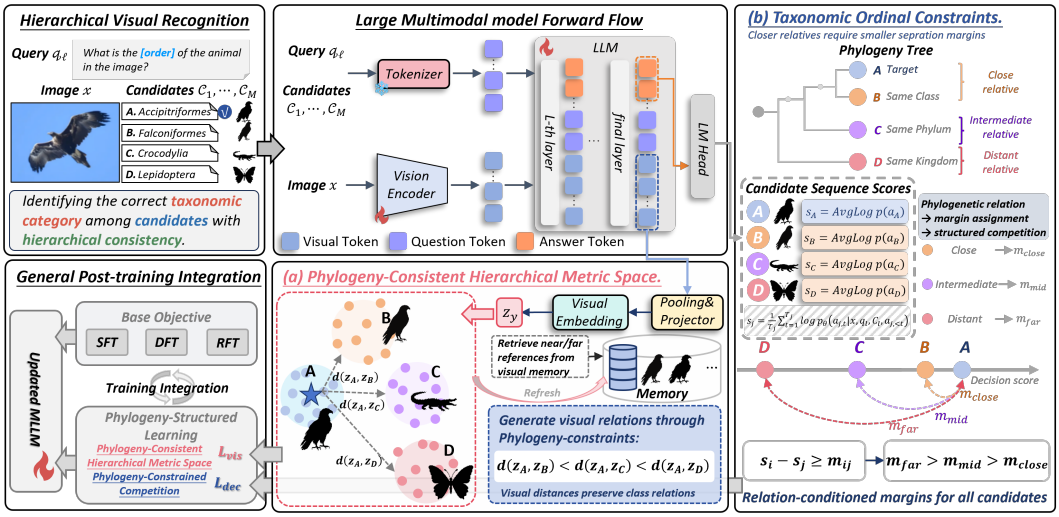


Figure 2: **Overview of PhyRC.** PhyRC uses taxonomic relations to impose ordinal separation among competing candidate hypotheses and to organize visual representations through local near–far constraints. The two objectives share the same relational supervision and can be integrated with SFT, DFT, or RFT for hierarchical visual recognition.

Let $\mathcal{T}$ denote the taxonomy and d_ℓ the depth of level ℓ. We measure the proximity of same-level categories using their lowest common ancestor (LCA):

$$h_{\mathcal{T}}(c_i, c_j) = \frac{\text{depth}(\text{LCA}(c_i, c_j))}{d_\ell}. \tag{2}$$

A larger value indicates deeper shared ancestry and closer taxonomic relatedness. This relation supplies the structural signal for both objectives.

3.2 Taxonomic Ordinal Constraints

PhyRC jointly optimizes the complete candidate set, supervising the target's score relative to every alternative. To incorporate the candidates' taxonomic relationships, we define $r_j = 1 - h_{\mathcal{T}}(c_y, c_j)$ and set

$$m_j = m_0 + \gamma r_j, \tag{3}$$

where m_0 is the minimum separation and γ controls its dependence on taxonomic distance. These *taxonomic ordinal constraints* require larger target–candidate gaps for more distant alternatives. The decision objective is

$$\mathcal{L}_{\text{dec}} = \log\left[1 + \sum_{j \neq y} \exp(s_j - s_y + m_j)\right]. \tag{4}$$

An alternative contributes more strongly when its score gap $s_y - s_j$ falls short of the required margin m_j. The constraint specifies separation from the target for each candidate, without imposing a total ordering among the candidates.

Setting all m_j to zero recovers candidate-level cross-entropy over the answer scores. In contrast, using a constant positive margin imposes uniform separation. PhyRC instead calibrates the required separation by taxonomic relation, operating directly on the LMM's generative scores without a category-specific classifier. Appendix A.4 gives the cross-entropy equivalence and the implemented margin bounds.

3.3 Phylogeny-Consistent Hierarchical Metric Learning

The visual objective encourages local feature distances to preserve the same taxonomic proximity. We mean-pool the final LLM-layer hidden states $\mathbf{H}_x^{(L)}$ at image-token positions, concatenate a

learned hierarchy-level embedding $\mathbf{e}_\ell$, and apply a normalized projection:

$$\mathbf{z}_x = \mathrm{Norm}\Big(g_\phi([\mathrm{MeanPool}(\mathbf{H}_x^{(L)}); \mathbf{e}_\ell])\Big). \tag{5}$$

Here, g_ϕ is a lightweight projection head, so the metric operates on multimodally contextualized visual features.

For each anchor, we select same-level near and far references from the minibatch and a feature queue such that $h_{\mathcal{T}}(c_y, c_{\mathrm{near}}) > h_{\mathcal{T}}(c_y, c_{\mathrm{far}})$. Using cosine distance $d(\cdot,\cdot)$, we optimize

$$\mathcal{L}_{\mathrm{vis}} = \mathrm{softplus}(d(\mathbf{z}_y, \mathbf{z}_{\mathrm{near}}) - d(\mathbf{z}_y, \mathbf{z}_{\mathrm{far}}) + m_{\mathrm{vis}}), \tag{6}$$

where m_{vis} sets the desired relative distance gap. This constraint organizes local neighborhoods by enabling relative competition among multiple categories, rather than requiring explicit global reconstruction of the taxonomy. Through the memory queue, category representations are accumulated across training iterations, providing multi-step relational references that progressively shape the relative distance structure between near and far categories. The historical queue features are detached references, and feature construction and memory details are provided in Appendix A.4.

3.4 Training Objective and Integration

The structured objective combines decision- and representation-level supervision:

$$\mathcal{L}_{\mathrm{struct}} = \lambda_{\mathrm{dec}}\mathcal{L}_{\mathrm{dec}} + \lambda_{\mathrm{vis}}\mathcal{L}_{\mathrm{vis}}. \tag{7}$$

Each iteration performs a structured update followed by an update under the base SFT, DFT, or RFT objective. For RFT, rollouts are collected before the structured update, and their recorded response tokens are re-evaluated under the updated policy before the policy and KL update. The complete training sequence, optimizer settings, and loss weights are provided in Algorithm 1 and Appendix A.4.

4 Experiments

4.1 Experimental Setup

Datasets and Tasks. We evaluate PhyRC on iNaturalist-2021 (iNat21) (Van Horn et al., 2021), using the Plantae and Animalia taxonomies with 4,271 and 5,388 leaf categories, respectively, across six hierarchical levels, which we denote as *iNat-Plant* and *iNat-Animal* hereafter. Following Tan et al. (2026), we formulate hierarchical recognition as candidate-based multiple-choice questions with candidates from the same taxonomic level, using $K = 4$ as the canonical setting. Post-training follows a one-shot setting. To assess transfer beyond leaf categories exposed during post-training, we report results separately for Task-Seen and Task-Unseen leaves. Task-Unseen categories are defined by task labels and are strictly withheld from post-training, such that images from a Task-Unseen species are excluded from training examples at all taxonomic levels. We additionally evaluate on NABirds Van Horn et al. (2015), with 555 native leaf categories, using its official image split and dataset-specific post-training to assess applicability to a different taxonomy.

Evaluation Metrics. We report five metrics following TARA He et al. (2026b). *Hierarchical Consistent Accuracy (HCA)* Wu et al. (2024) measures the fraction of samples whose predictions are correct at every evaluated level, while *Leaf accuracy* measures correctness at the finest level. *Point-Overlap Ratio (POR)* Yi et al. (2022) averages the fraction of correct nodes along each path. *Strict Point-Overlap Ratio (S-POR)* Tan et al. (2026) averages the longest consecutive run of correct predictions normalized by path depth. *Top Overlap Ratio (TOR)* Wu et al. (2024) averages the fraction of adjacent-level pairs with both predictions correct.

Detailed experimental settings, including datasets, task definitions, evaluation protocols, and metric definitions, are provided in Appendix A.1.

4.2 Overall Effectiveness

Table 1: Performance on Task-Seen and Task-Unseen categories under different post-training objectives. Small ↑ annotations denote the absolute gain of PhyRC over the corresponding SFT, DFT, or RFT baseline in percentage points. Boldface denotes the best result under the same experimental setting.

Method	Task-Seen					Task-Unseen				
	HCA↑	Leaf↑	POR↑	S-POR↑	TOR↑	HCA↑	Leaf↑	POR↑	S-POR↑	TOR↑
Qwen3-VL-2B *iNat-Plant*										
SFT	13.79	33.23	70.50	58.78	56.43	14.92	37.80	69.97	56.95	55.03
SFT + PhyRC	**17.47** ↑3.68	**38.24** ↑5.01	**74.31** ↑3.81	**63.22** ↑4.44	**61.58** ↑5.15	**16.46** ↑1.54	**40.86** ↑3.06	**71.87** ↑1.90	**58.86** ↑1.91	**57.46** ↑2.43
DFT	10.47	31.74	65.93	52.21	49.62	10.60	35.01	66.06	51.88	49.93
DFT + PhyRC	**16.80** ↑6.33	**37.79** ↑6.05	**73.40** ↑7.47	**61.27** ↑9.06	**59.97** ↑10.35	**17.29** ↑6.69	**41.14** ↑6.13	**72.27** ↑6.21	**59.30** ↑7.42	**58.13** ↑8.20
RFT	9.88	31.88	65.67	52.28	49.43	9.90	34.45	64.67	50.19	48.34
RFT + TARA	10.72	32.27	65.63	52.36	49.62	11.99	34.31	65.88	52.30	50.35
RFT + PhyRC	**17.90** ↑8.02	**39.08** ↑7.20	**73.30** ↑7.63	**61.38** ↑9.10	**61.54** ↑12.11	**15.90** ↑6.00	**41.98** ↑7.53	**73.08** ↑8.41	**59.25** ↑9.06	**59.08** ↑10.74
Qwen3-VL-2B *iNat-Animal*										
SFT	11.87	33.48	69.96	60.60	57.73	8.64	33.12	68.59	59.51	55.45
SFT + PhyRC	**14.95** ↑3.08	**38.91** ↑5.43	**72.15** ↑2.19	**61.02** ↑0.42	**60.26** ↑2.53	**11.01** ↑2.37	**38.31** ↑5.19	**69.90** ↑1.31	**59.65** ↑0.14	**56.72** ↑1.27
DFT	9.21	31.27	67.68	57.46	54.86	8.10	27.30	66.12	57.95	53.08
DFT + PhyRC	**13.55** ↑4.34	**38.05** ↑6.78	**71.66** ↑3.98	**60.11** ↑2.65	**59.60** ↑4.74	**12.65** ↑4.55	**37.85** ↑10.55	**69.79** ↑3.67	**59.04** ↑1.09	**57.03** ↑3.95
RFT	9.47	29.35	67.73	58.91	55.39	7.10	27.21	65.85	56.48	52.87
RFT + TARA	10.19	30.52	67.52	57.08	55.16	7.92	27.02	66.38	57.87	53.65
RFT + PhyRC	**14.97** ↑5.50	**39.36** ↑10.01	**72.20** ↑4.47	**61.96** ↑3.05	**60.34** ↑4.95	**14.01** ↑6.91	**41.31** ↑14.10	**71.99** ↑6.14	**60.93** ↑4.45	**59.20** ↑6.33
Qwen2.5-VL-3B *iNat-Plant*										
SFT	18.32	44.77	73.02	57.78	57.99	17.29	49.79	72.25	55.04	55.82
SFT + PhyRC	**24.62** ↑6.30	**50.14** ↑5.37	**77.97** ↑4.95	**63.96** ↑6.18	**65.02** ↑7.03	**23.57** ↑6.28	**54.81** ↑5.02	**76.89** ↑4.64	**60.67** ↑5.63	**62.43** ↑6.61
DFT	16.35	42.77	70.94	56.15	55.59	16.32	47.42	71.27	54.32	54.87
DFT + PhyRC	**23.83** ↑7.48	**50.93** ↑8.16	**77.00** ↑6.06	**61.81** ↑5.66	**63.17** ↑7.58	**23.01** ↑6.69	**52.16** ↑4.74	**76.03** ↑4.76	**60.67** ↑6.35	**61.73** ↑6.86
RFT	16.80	43.19	71.64	56.98	56.50	17.29	48.40	71.55	55.16	55.34
RFT + TARA	18.23	45.36	73.20	58.06	58.08	18.27	48.12	72.64	56.62	57.32
RFT + PhyRC	**24.87** ↑8.07	**50.56** ↑7.37	**77.66** ↑6.02	**63.46** ↑6.48	**64.65** ↑8.15	**23.85** ↑6.56	**53.42** ↑5.02	**76.43** ↑4.88	**60.93** ↑5.77	**62.09** ↑6.75
Qwen2.5-VL-3B *iNat-Animal*										
SFT	23.39	47.91	77.00	68.61	65.88	21.20	52.05	75.74	65.88	63.11
SFT + PhyRC	**30.03** ↑6.64	**53.70** ↑5.79	**80.09** ↑3.09	**72.17** ↑3.56	**69.99** ↑4.11	**28.03** ↑6.83	**58.78** ↑6.73	**79.16** ↑3.42	**69.03** ↑3.15	**67.32** ↑4.21
DFT	21.24	45.12	75.85	67.20	64.38	18.11	46.95	74.42	64.56	61.78
DFT + PhyRC	**28.98** ↑7.74	**53.02** ↑7.90	**79.44** ↑3.59	**71.29** ↑4.09	**69.12** ↑4.74	**24.57** ↑6.46	**55.60** ↑8.65	**77.74** ↑3.32	**67.62** ↑3.06	**65.62** ↑3.84
RFT	22.41	45.81	76.21	67.94	65.11	18.20	46.86	74.22	64.42	61.55
RFT + TARA	23.57	47.17	76.95	68.68	65.89	19.93	47.86	74.84	65.36	62.42
RFT + PhyRC	**30.12** ↑7.71	**54.19** ↑8.38	**79.98** ↑3.77	**71.95** ↑4.01	**69.79** ↑4.68	**26.11** ↑7.91	**57.14** ↑10.28	**78.57** ↑4.35	**68.74** ↑4.32	**66.59** ↑5.04

Effectiveness across post-training objectives. Table 1 evaluates PhyRC under SFT Ouyang et al. (2022), DFT Wu et al. (2026), and RFT Liu et al. (2025b) on Qwen3-VL-2B Bai et al. (2025a) and Qwen2.5-VL-3B Bai et al. (2025b) across the Plantae and Animalia taxonomies. PhyRC consistently improves the corresponding base objectives across different post-training paradigms, demonstrating that phylogeny-constrained candidate competition can be integrated with diverse optimization strategies. To further compare with existing taxonomy-aware post-training methods, we evaluate PhyRC against TARA He et al. (2026b) under the RFT setting using the same evaluation protocol, as reported in Table 1. PhyRC consistently improves over TARA across both backbones, taxonomies, *Task-Seen* and *Task-Unseen* categories, and all five evaluation metrics, further supporting the effectiveness of explicitly modeling taxonomic relations among competing candidates.

Generalization across MLLM architectures. Table 2 evaluates PhyRC on Gemma-3-4B Gemma Team (2025), InternVL2.5-4B Chen et al. (2025), LLaVA-OneVision-7B Li et al. (2024), and Qwen3-VL-8B Bai et al. (2025a). Across all evaluated architectures, PhyRC consistently improves the corresponding RFT baseline on both Plantae and Animalia, indicating that its effectiveness is not specific to the primary Qwen backbones. The improvements also persist on Task-Unseen categories, providing further evidence that the learned relational structure can transfer beyond category identities exposed during task-specific post-training. Consistent gains across HCA, Leaf, POR, S-POR, and TOR further show that this benefit extends from fine-grained discrimination to hierarchy-consistent recognition across different model families and scales.

Table 2: Performance on Task-Seen and Task-Unseen categories across different LMM backbones under RFT. Small ↑ annotations denote the absolute improvement of PhyRC over the corresponding RFT baseline in percentage points. Boldface denotes the best result under the same experimental setting.

Dataset	Method	Task-Seen					Task-Unseen				
		HCA↑	Leaf↑	POR↑	S-POR↑	TOR↑	HCA↑	Leaf↑	POR↑	S-POR↑	TOR↑
Gemma-3-4B											
iNat-Plant	RFT	4.64	27.72	57.75	40.60	37.78	3.63	28.45	57.07	39.31	37.02
	RFT + PhyRC	**12.07** ↑7.43	**37.90** ↑10.18	**69.09** ↑11.34	**52.13** ↑11.53	**51.89** ↑14.11	**15.76** ↑12.13	**45.75** ↑17.30	**70.80** ↑13.73	**53.53** ↑14.22	**54.42** ↑17.40
iNat-Animal	RFT	6.20	29.03	64.39	54.80	49.49	4.28	28.84	62.24	52.41	46.84
	RFT + PhyRC	**13.64** ↑7.44	**41.22** ↑12.19	**72.56** ↑8.17	**62.99** ↑8.19	**59.45** ↑9.96	**13.92** ↑9.64	**47.13** ↑18.29	**72.11** ↑9.87	**61.30** ↑8.89	**57.83** ↑10.99
InternVL2.5-4B											
iNat-Plant	RFT	6.81	30.39	61.54	45.31	42.70	7.11	33.47	60.51	43.58	41.28
	RFT + PhyRC	**9.71** ↑2.90	**36.18** ↑5.79	**66.64** ↑5.10	**47.41** ↑2.10	**48.61** ↑5.91	**11.02** ↑3.91	**42.68** ↑9.21	**68.81** ↑8.30	**51.14** ↑7.56	**51.02** ↑9.74
iNat-Animal	RFT	9.51	31.27	68.17	59.72	55.33	8.01	29.94	65.74	57.28	52.36
	RFT + PhyRC	**11.73** ↑2.22	**37.96** ↑6.69	**71.12** ↑2.95	**61.90** ↑2.18	**58.03** ↑2.70	**10.46** ↑2.45	**43.04** ↑13.10	**70.26** ↑4.52	**60.30** ↑3.02	**55.67** ↑3.31
LLaVA-OneVision-7B											
iNat-Plant	RFT	9.90	31.26	66.77	52.26	49.89	9.34	34.59	66.90	51.56	49.43
	RFT + PhyRC	**13.96** ↑4.06	**39.48** ↑8.22	**70.63** ↑3.86	**54.20** ↑1.94	**54.23** ↑4.34	**13.53** ↑4.19	**41.98** ↑7.39	**70.41** ↑3.51	**52.79** ↑1.23	**52.58** ↑3.15
iNat-Animal	RFT	8.60	31.78	65.40	50.14	51.55	7.37	31.67	63.80	48.74	49.46
	RFT + PhyRC	**13.41** ↑4.81	**41.99** ↑10.21	**71.66** ↑6.26	**57.36** ↑7.22	**58.49** ↑6.94	**13.01** ↑5.64	**39.13** ↑7.46	**68.14** ↑4.34	**54.75** ↑6.01	**54.50** ↑5.04
Qwen3-VL-8B											
iNat-Plant	RFT	21.41	42.82	74.95	65.41	62.80	21.34	43.65	74.43	62.88	63.40
	RFT + PhyRC	**28.00** ↑6.59	**51.35** ↑8.53	**79.69** ↑4.74	**69.90** ↑4.49	**68.56** ↑5.76	**26.22** ↑4.88	**49.51** ↑5.86	**78.34** ↑3.91	**68.83** ↑5.95	**66.22** ↑2.82
iNat-Animal	RFT	22.80	44.30	77.48	70.82	67.49	20.38	44.86	75.39	67.20	65.81
	RFT + PhyRC	**30.47** ↑7.67	**54.21** ↑9.91	**81.13** ↑3.65	**74.07** ↑3.25	**71.55** ↑4.06	**27.39** ↑7.01	**55.69** ↑10.83	**79.89** ↑4.50	**73.45** ↑6.25	**70.61** ↑4.80

Table 3: Generalization to NABirds under NABirds-native post-training and direct cross-dataset transfer. The left block reports results after post-training on the native NABirds hierarchy, while the right block directly evaluates iNat21-trained checkpoints on NABirds without additional post-training or parameter updates. Small ↑ annotations denote absolute gains of PhyRC over the corresponding RFT baseline in percentage points. Boldface denotes the best result under the same experimental setting.

Method	NABirds-Native Post-Training					Direct iNat21→NABirds Transfer				
	HCA↑	Leaf↑	POR↑	S-POR↑	TOR↑	HCA↑	Leaf↑	POR↑	S-POR↑	TOR↑
Qwen3-VL-2B										
RFT	23.30	43.84	60.91	50.77	41.65	24.16	43.71	60.73	49.50	41.75
RFT + PhyRC	**27.47** ↑4.17	**49.43** ↑5.59	**64.53** ↑3.62	**53.78** ↑3.01	**46.24** ↑4.5	**25.57** ↑1.41	**45.09** ↑1.38	**61.00** ↑0.27	**51.57** ↑2.07	**42.95** ↑1.20
Qwen2.5-VL-3B										
RFT	49.59	63.76	76.55	70.23	64.06	45.21	60.01	73.97	67.42	60.42
RFT + PhyRC	**52.27** ↑2.68	**66.67** ↑2.91	**78.70** ↑2.15	**72.34** ↑2.11	**66.78** ↑2.7	**46.55** ↑1.34	**60.35** ↑0.34	**74.67** ↑0.70	**68.50** ↑1.08	**61.63** ↑1.21

4.3 Generalization Beyond iNaturalist-2021

Generalization to another hierarchy. Table 3 evaluates PhyRC on NABirds (Van Horn et al., 2015) under two complementary settings. With dataset-specific post-training on the native NABirds hierarchy, PhyRC consistently improves the corresponding RFT baseline across both backbones and all five metrics, demonstrating its applicability to a different fine-grained hierarchical benchmark.

Direct cross-dataset transfer. We further evaluate the iNat21-trained checkpoints directly on NABirds without additional post-training, fine-tuning, or parameter updates. PhyRC again improves over the matched RFT checkpoint across all five metrics on both backbones. Although the gains are smaller than under dataset-specific post-training, their consistent direction suggests that part of the learned relational structure transfers across distinct fine-grained category systems.

4.4 Ablation Study

Disentangling supervision designs. Table 4 analyzes PhyRC on Qwen3-VL-2B under both Task-Seen and Task-Unseen evaluation settings on iNat21-Plant and iNat21-Animal. All ablation variants use seed 42 and are independently trained from the same pretrained initialization under the canonical ablation protocol. Adjacent rows therefore represent separate runs rather than cumulative training.

We study the method from two complementary perspectives. First, we decompose PhyRC into target-only supervision (TOS), full-candidate competition (FCC), taxonomy-calibrated separation (TCS), and phylogeny-consistent visual regularization (PCVR), and evaluate their contributions under different category exposure settings. Second, under the same full-candidate competition protocol, we isolate the design of target–competitor separation by comparing uniform candidate separation (UCS), randomized relation assignment (RRA), and TCS.

Table 4: Ablation results on Qwen3-VL-2B under Task-Seen and Task-Unseen evaluation settings. Variants are denoted by the abbreviations introduced in Sec. 4.4. Results are percentages; boldface and underlining denote the best and second-best results, respectively, under the same evaluation setting.

Variant	Task-Seen				Task-Unseen			
	iNat-Plant		*iNat-Animal*		*iNat-Plant*		*iNat-Animal*	
	HCA ↑	Leaf ↑	HCA ↑	Leaf ↑	HCA ↑	Leaf ↑	HCA ↑	Leaf ↑
(a) Component contribution								
RFT	9.88	31.88	9.47	29.35	9.90	34.45	7.10	27.21
+TOS	15.76	35.79	13.31	35.79	15.06	38.08	8.92	31.76
+FCC	16.85	37.65	14.15	37.98	15.06	40.45	<u>12.01</u>	37.49
+PCVR	11.87	31.99	9.84	29.70	11.58	34.87	7.92	29.12
+TCS	<u>17.08</u>	<u>38.29</u>	<u>14.60</u>	<u>39.22</u>	**15.90**	40.86	11.19	<u>38.31</u>
+PhyRC (TCS + PCVR)	**17.90**	**39.08**	**14.97**	**39.36**	**15.90**	**41.98**	**14.01**	**41.31**
(b) Relation calibration under full-candidate competition								
+FCC	16.85	37.65	14.15	37.98	15.06	40.45	<u>12.01</u>	37.49
+UCS	16.91	37.06	14.15	38.19	15.62	40.31	11.19	38.13
+RRA	16.91	37.23	12.87	38.63	<u>15.76</u>	<u>41.84</u>	10.37	37.49
+TCS	<u>17.08</u>	<u>38.29</u>	<u>14.60</u>	<u>39.22</u>	**15.90**	40.86	11.19	<u>38.31</u>

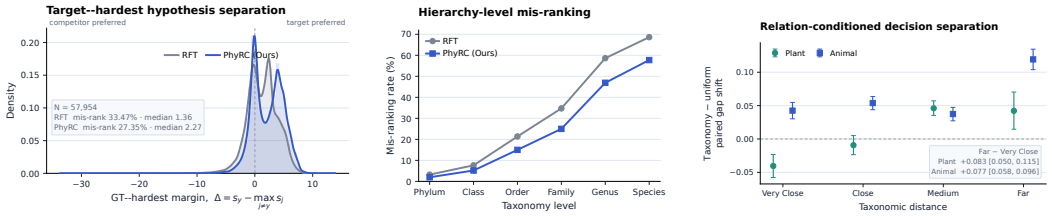
Decision-side supervision. TOS already improves the RFT baseline, showing that additional response-level training itself contributes to recognition performance. However, FCC provides a substantially stronger improvement, indicating that explicitly optimizing the target relative to all competing hypotheses is the more important source of supervision. The separation controls further show that neither UCS nor RRA consistently improves over FCC. In contrast, TCS provides the strongest overall decision-side results, supporting the use of the true taxonomic relation to determine competitor-specific separation.

Visual-side supervision. PCVR alone yields modest improvements over RFT, but remains substantially weaker than the decision-side objectives. Its contribution becomes clearer when combined with TCS, where the full PhyRC objective further improves nearly all metrics over TCS alone. This pattern supports a complementary role for visual regularization: candidate-level supervision primarily improves discrimination, while PCVR encourages the learned representations to preserve the same taxonomic relations in visual space. We further examine this representation structure in Section 4.5.

Complete ablation results are reported in Appendix B.2 with the summary shown in Table 4. We further provide additional analyses on qualitative examples, hyperparameter sensitivity, and candidate-competition designs in Appendices B.3, B.4, and B.5, respectively.

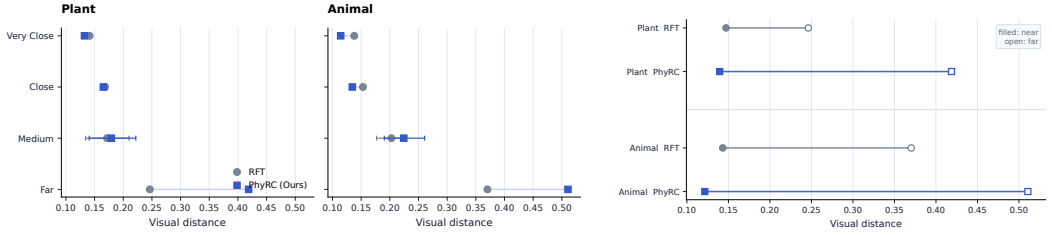
4.5 Mechanism Analysis of Phylogeny-Constrained Learning

Phylogeny-structured hypothesis competition. Figure 3 examines how PhyRC changes competition among candidates. Compared with RFT, PhyRC produces clearer target–competitor separation and reduces mis-ranking across the hierarchy, with particularly visible benefits at finer recognition levels. More importantly, the change in separation is relation-dependent rather than uniform: taxonomically distant competitors are pushed farther from the target than closely related ones. These trends support the intended mechanism of *taxonomic ordinal constraints*, which use phylogenetic relations to structure relative competition among candidate answers. Definitions and paired diagnostic protocols are provided in Appendix A.5.



(a) Target–Hardest Separation (b) Hierarchical Mis-Ranking (c) Relation-conditioned gap shift

Figure 3: **Mechanism analysis of taxonomy-structured candidate competition.** **(a)** compares RFT and PhyRC in target–hardest-competitor separation. **(b)** compares RFT and PhyRC in hierarchy-level mis-ranking rates. **(c)** compares taxonomy-calibrated and uniform margins to examine their effects on relation-dependent target–competitor separation.



(a) Relation-dependent visual reorganization. (b) Near–far visual separation.

Figure 4: **Mechanism analysis of phylogeny-consistent visual representations.** PhyRC reorganizes visual distances according to taxonomic relations, preserving local proximity among close categories while increasing separation from more distant ones.

Phylogeny-consistent visual metric structure. Figure 4 shows that the same relational ordering also emerges in the shared visual representation space. PhyRC preserves or strengthens local proximity among taxonomically related categories while increasing separation from more distant categories, leading to a clearer near–far organization than RFT. This behavior is consistent with the intended *phylogeny-consistent hierarchical metric space*, where local visual geometry reflects taxonomic proximity rather than expanding all feature distances indiscriminately. We interpret this as evidence for local taxonomic ordering rather than global reconstruction of the entire taxonomy. Feature extraction and diagnostic details are provided in Appendix A.5.

We further examine the robustness of PhyRC under various training and evaluation variations, including random initialization, candidate ordering, task-unseen exposure, and hierarchical scale. PhyRC consistently improves over competitive baselines across these settings, indicating that its gains are robust to different experimental conditions and stem from the proposed phylogeny-constrained relative competition mechanism (see Appendix B.7). To further verify that the improvements of PhyRC arise from introducing relation-aware candidate competition rather than additional optimization steps, we conduct a compute-matched training experiment with controlled update budgets (see Appendix B.8).

5 Conclusion

We proposed **PhyRC**, a structured post-training framework that recasts supervision for hierarchical visual recognition from category identities to relative taxonomic relations among competing candidates. PhyRC operationalizes these relations through *taxonomic ordinal constraints* over candidate scores and a *phylogeny-consistent hierarchical metric space* for visual representations, encouraging the model to learn relational structure that can be reused across different category combinations. Across multiple LMM backbones, post-training objectives, and biological hierarchies, PhyRC consistently improves hierarchical recognition and taxonomic consistency, with gains extending to Task-Unseen categories. Overall, our results highlight relative taxonomic structure as an effective source of supervision for moving beyond target-centric post-training in hierarchical visual recognition.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing of the manuscript, including improving grammar, wording, readability, and overall clarity of presentation. They were not used to generate experimental results, design the methodology, conduct analyses, or determine the scientific conclusions of this work. All AI-assisted revisions were carefully reviewed and edited by the authors, who take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work studies hierarchical visual recognition using publicly available research datasets and does not involve human-subject experiments or the collection of personally identifiable or sensitive information. Our method is intended for research on structured visual recognition and taxonomy-aware multimodal learning. As with other visual recognition systems, predictions may inherit biases, coverage limitations, or annotation errors present in the underlying training data and taxonomies, and model outputs should therefore not be interpreted as authoritative biological determinations. We report experimental protocols and limitations to support transparent and responsible interpretation of the results.

REPRODUCIBILITY STATEMENT

We provide detailed information necessary to reproduce the proposed method and experiments throughout the paper and supplementary material. The method formulation, candidate scoring procedure, taxonomic ordinal constraints, visual relational objective, and complete training objective are specified in Sec. 3. Implementation details, model configurations, optimization hyperparameters, data preprocessing, candidate construction, and evaluation protocols are reported in the appendix. We additionally report controlled baselines, multiple random seeds, and protocol-level analyses where applicable to assess the robustness of the reported improvements. Source code, configuration files, and scripts for the main training and evaluation pipelines are provided through the anonymized supplementary material.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12503–12512. IEEE, 2020.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, 2007.

Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your "flamingo" is my "bird": fine-grained, or not. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11471–11480. IEEE, 2021.

Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4848–4857. IEEE, 2022.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European conference on computer vision*, pp. 48–64. Springer, 2014.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3649–3658. IEEE, 2020.

Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In *European Conference on Computer Vision*, pp. 252–267. Springer, 2022.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Hulingxiao He, Zhi Tan, and Yuxin Peng. Learning taxonomic trees with hierarchical representation regularization for large multimodal models. 2026a.

Hulingxiao He, Zhi Tan, and Yuxin Peng. Taxonomy-aware representation alignment for hierarchical visual recognition with large multimodal models. pp. 31124–31134, 2026b.

Juan Jiang, Jingmin Yang, Wenjie Zhang, and Hongbin Zhang. Hierarchical multi-granularity classification based on bidirectional knowledge transfer. *Multimedia Syst.*, 30(4), 2024. ISSN 0942-4962. doi: 10.1007/s00530-024-01412-x. URL https://doi.org/10.1007/s00530-024-01412-x.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015.

Hyeongjun Kwon, Jinhyun Jang, Jin Kim, Kwonyoung Kim, and Kwanghoon Sohn. Improving visual recognition with hyperbolical visual hierarchy mapping. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17364–17374. IEEE, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. Lipo: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2404–2420, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044. IEEE, 2025b.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e099c1c9699814af0be873a175361713-Abstract-Conference.html.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QyCATzIwKb.

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.

Fatemeh Pesaran Zadeh, Yoojin Oh, and Gunhee Kim. LPOI: Listwise preference optimization for vision language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 26830–26844, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Depanshu Sani and Saket Anand. Hier-cos: Making deep features hierarchy-aware via composition of orthogonal subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11263–11272, 2026.

Carlos N Silla Jr and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22(1):31–72, 2011.

Vésteinn Snæbjarnarson, Kevin Du, Niklas Stoehr, Serge Belongie, Ryan Cotterell, Nico Lang, and Stella Frank. Taxonomy-aware evaluation of vision–language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9109–9120. IEEE, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024.

Yuwen Tan, Yuan Qing, and Boqing Gong. The llm bottleneck: Why open-source vision llms struggle with hierarchical visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38532–38543, 2026.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 595–604, 2015.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12884–12893, 2021.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06361.

Ziwei Wang, Sameera Ramasinghe, Chenchen Xu, Julien Monteil, Loris Bazzani, and Thalaiyasingam Ajanthan. Learning visual hierarchies in hyperbolic space for image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9924–9934, 2025.

Tz-Ying Wu, Chih-Hui Ho, and Nuno Vasconcelos. Protect: Prompt tuning for taxonomic open set classification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16531–16540. IEEE, 2024.

Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, et al. On the generalization of sft: A reinforcement learning perspective with reward rectification. In *International Conference on Learning Representations*, volume 2026, pp. 27550–27571, 2026.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pp. 1192–1199, 2008.

Peng Xia, Xingtong Yu, Ming Hu, Lie Ju, Zhiyong Wang, Peibo Duan, and Zongyuan Ge. Hgclip: Exploring vision-language models with graph representations for hierarchical understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 269–280, 2025.

Kai Yi, Xiaoqian Shen, Yunhao Gou, and Mohamed Elhoseiny. Exploring hierarchical graph representation for large-scale zero-shot image classification. In *European Conference on Computer Vision*, pp. 116–132. Springer, 2022.

Daiki Yoshikawa and Takashi Matsubara. PHyCLIP: ℓ_1-product of hyperbolic factors unifies hierarchy and compositionality in vision-language representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.08919.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19985–19995. IEEE, 2025.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36:10935–10950, 2023.

Ming Zhang, Xuefei Zhe, Le Ou-Yang, Shifeng Chen, and Hong Yan. Semantic hierarchy preserving deep hashing for large-scale image retrieval. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pp. 1–5. IEEE, 2021.

Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1114–1123, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Zhang_Embedding_Label_Structures_CVPR_2016_paper.html.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Method **3**

3.1 Problem Formulation 3

3.2 Taxonomic Ordinal Constraints 4

3.3 Phylogeny-Consistent Hierarchical Metric Learning 4

3.4 Training Objective and Integration 5

4 Experiments **5**

4.1 Experimental Setup 5

4.2 Overall Effectiveness 5

4.3 Generalization Beyond iNaturalist-2021 7

4.4 Ablation Study 7

4.5 Mechanism Analysis of Phylogeny-Constrained Learning 8

5 Conclusion **9**

A Experimental Protocol and Implementation **16**

A.1 Datasets, Tasks, and Evaluation 16

A.2 Taxonomy Construction and Training Details 18

A.3 Prompt Format and Candidate Scoring Protocol 20

A.4 Implementation Details and Training Procedure 21

A.5 Diagnostic Definitions and Evaluation Protocol 22

B Supplementary Experiments and Analyses **25**

B.1 Reference Comparison with Reported HiR^2 Results 25

B.2 Additional Ablation Study 26

B.3 Qualitative Results 27

B.4 Representation and Hyperparameter Sensitivity 30

B.5 Candidate-Competition Ablations 31

B.6 Additional Representation-Space Diagnostics 31

B.7 Robustness Analysis 32

B.8 Controlled Analysis of Longer RFT Optimization 36

A Experimental Protocol and Implementation

A.1 Datasets, Tasks, and Evaluation

iNaturalist-2021. Our main experiments use the Plantae and Animalia subsets of iNaturalist-2021 (iNat21) (Van Horn et al., 2021), following the hierarchical candidate-recognition protocol of TARA (He et al., 2026b). Rather than constructing a new train–test split, we use the supplied one-shot hierarchical training data and fixed evaluation annotations. The training set contains 9,659 hierarchy-specific QA records, comprising 590 order-level, 1,367 family-level, 2,574 genus-level, and 5,128 species-level examples. Each record consists of an image, a hierarchy-specific question, and four candidate categories from the same taxonomic level, including one ground-truth category and three negatives. The target response is the category name itself rather than an option letter. For controlled comparisons, all methods use the same candidate sets and candidate ordering.

The underlying biological taxonomy follows seven canonical ranks from kingdom to species. Training questions cover order, family, genus, and species, whereas evaluation covers the six non-root ranks from phylum to species. Taxonomic paths and lowest common ancestor (LCA) relations are used only to construct the structured supervision and are not included in the language-model prompt. The fixed evaluation set contains 4,271 Plantae images and 5,388 Animalia images, with identical evaluation records and candidate mappings used across the corresponding method comparisons.

NABirds. We further evaluate PhyRC on NABirds (Van Horn et al., 2015) to examine its applicability to a distinct fine-grained hierarchy. NABirds contains 48,562 images and 555 native leaf categories, with an official split of 23,929 training and 24,633 test images. Following the hierarchy preprocessing adopted by HPDH Zhang et al. (2021), we represent its variable-depth taxonomy using four non-root levels containing 22, 228, 495, and 555 categories from coarse to fine. Missing fine-level entries are completed in the label matrix following this protocol; in our implementation, this completion does not introduce artificial self-edges into the native taxonomy. We preserve the original node identities and taxonomic paths when computing category relations and refer to the resulting hierarchy simply as Level 1–4, rather than assigning Linnaean rank names.

We convert NABirds to the same four-way hierarchical candidate-recognition setting used for iNat21. At each level, an example contains one ground-truth category and three distinct same-level negatives, and evaluation requires predictions at all four levels for each image. For the NABirds-specific experiments, post-training is performed on the official training split and evaluation is conducted on the official test split. This setting therefore tests whether PhyRC remains effective when trained and evaluated under a different dataset and hierarchical organization.

Table 5: Statistics of the *Task-Unseen* split. Unseen species are completely excluded from post-training, including their images and associated hierarchy-level supervision.

Dataset	Total Species	Task-Seen Species	Task-Unseen Species
iNat-Plant	4,271	3,554	717
iNat-Animal	5,388	4,289	1,099

Task-Unseen split verification. We perform an explicit data leakage audit for the *Task-Unseen* protocol over the post-training data construction pipeline. The protocol designates a subset of species as unseen categories and removes all corresponding training instances, including their images and hierarchy-level supervision.

The resulting split contains 717 Task-Unseen species in iNat21-Plant and 1,099 Task-Unseen species in iNat21-Animal, while the remaining Task-Seen training sets contain 3,554 and 4,289 species, respectively (Table 5). The strict training pipeline therefore contains no visual samples belonging to Task-Unseen species, and no exact image overlap exists between training and evaluation sets.

The filtering procedure is applied before constructing the final post-training dataset used by all Task-Unseen experiments. For every QA instance, Task-Unseen species are checked against target labels, candidate options, lineage annotations, prompts, and generated answers. No Task-Unseen leakage is observed in any of these fields. The filtering covers all hierarchy levels present in the original QA

source, including Species, Genus, Family, and Order questions, while Class and Phylum questions are absent.

The final filtered training set is then deterministically expanded to match the canonical optimization schedule, resulting in the same number of training updates used in the reported Task-Unseen experiments. All Task-Unseen results, are obtained from models trained from this strict filtered dataset rather than from the original unrestricted training split.

The PhyRC visual regularization queue follows the same filtering protocol. It is initialized from the strict training samples only and receives entries exclusively from the corresponding filtered training batches. No Task-Unseen species samples are inserted into the queue, resulting in zero visual-queue leakage.

Direct cross-dataset transfer. Separately, we evaluate direct transfer from iNat21 to NABirds. For this setting, the Qwen3-VL-2B and Qwen2.5-VL-3B checkpoints obtained from iNat21 post-training are applied directly to the NABirds evaluation protocol without any NABirds-specific post-training, fine-tuning, or parameter updates. RFT and PhyRC are evaluated using exactly the same NABirds images, hierarchy, candidate sets, and prediction protocol. This experiment is intended to measure transfer beyond the taxonomy used for task-specific post-training and is distinct from the NABirds-specific training setting described above.

Evaluation metrics. We evaluate hierarchical recognition using HCA Wu et al. (2024) , Leaf accuracy, POR and TOR Wu et al. (2024), and S-POR Tan et al. (2026). Let N denote the number of images in the evaluated subset and L_i the number of evaluated hierarchy levels for image i, ordered from coarse to fine. We denote the ground-truth and predicted labels at level j by $y_i^{(j)}$ and $\widehat{y}_i^{(j)}$, respectively, and define the correctness indicator as $c_{ij} = \mathbf{1}[\widehat{y}_i^{(j)} = y_i^{(j)}]$. All metrics below take values in $[0, 1]$ and are reported as percentages.

Hierarchical Consistent Accuracy (HCA). A sample is correct only when all predictions along its evaluated path match the ground truth:

$$\mathrm{HCA} = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{L_i} c_{ij}. \tag{8}$$

This complete-path criterion makes HCA our primary hierarchical recognition metric.

Leaf accuracy. Accuracy at the finest evaluated level is

$$\mathrm{Acc}_{\mathrm{leaf}} = \frac{1}{N} \sum_{i=1}^{N} c_{iL_i}. \tag{9}$$

A correct leaf prediction need not be accompanied by correct predictions at other levels, so $\mathrm{HCA} \leq \mathrm{Acc}_{\mathrm{leaf}}$.

Point-Overlap Ratio (POR). POR gives partial credit to each correct node, averaging first within each path and then across images:

$$\mathrm{POR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{L_i} \sum_{j=1}^{L_i} c_{ij}. \tag{10}$$

Strict Point-Overlap Ratio (S-POR). S-POR measures the longest uninterrupted sequence of correct predictions, normalized by the evaluated path depth:

$$\text{S-POR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{L_i} \max_{1 \leq a \leq b \leq L_i} \left[(b - a + 1) \prod_{j=a}^{b} c_{ij} \right]. \tag{11}$$

The run may begin at any evaluated level; it is not restricted to a correct prefix from the coarsest level. If all predictions are incorrect, the sample contributes zero.

Top Overlap Ratio (TOR). TOR measures the proportion of adjacent-level pairs whose two predictions are both correct:

$$\mathrm{TOR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_i-1}\sum_{j=1}^{L_i-1} c_{ij}c_{i,j+1}. \tag{12}$$

All evaluated paths have at least two levels. TOR therefore measures local pairwise correctness along the ground-truth path, rather than consistency between predictions irrespective of their correctness.

For Task-Seen and Task-Unseen evaluation, all five metrics use the corresponding image subset defined by ground-truth leaf exposure. We retain the same evaluation implementation and fixed candidate annotations for each method comparison, without test-time taxonomy-aware reranking or path correction. Reported gains are absolute percentage-point differences.

Backbones and comparison methods. Our primary backbones are Qwen3-VL-2B-Instruct Bai et al. (2025a) and Qwen2.5-VL-3B-Instruct Bai et al. (2025b). For both backbones, we evaluate PhyRC under three representative post-training objectives—SFT, DFT, and RFT—and compare each objective with its PhyRC-enhanced counterpart. To assess generality across model families and scales, we further evaluate Gemma-3-4B Gemma Team (2025), InternVL2.5-4B Chen et al. (2025), LLaVA-OneVision-7B Li et al. (2024), and Qwen3-VL-8B Bai et al. (2025a) under RFT. Within each setting, all methods use the same training data, prompts, candidate sets, and evaluation protocol.

SFT applies standard teacher-forced supervised fine-tuning to the ground-truth category response Ouyang et al. (2022), optimizing the token-level negative log-likelihood of the reference answer. It therefore provides direct response-level supervision but does not explicitly model the relative preference between the target and the other candidate categories.

DFT follows Dynamic Fine-Tuning Wu et al. (2026), which retains the teacher-forced training paradigm of SFT but dynamically rescales each target-token objective by the model probability assigned to that token. The probability used for reweighting is detached from gradient computation, so DFT modulates the magnitude of the token-level update without introducing an additional reward model or online sampling. Compared with conventional SFT, this formulation reduces the disproportionate influence of low-probability target tokens and provides a more conservative response-level optimization objective.

RFT adopts reward-based fine-tuning for visual recognition, where responses sampled from the current policy are evaluated with a verifiable correctness reward and used to update the model through policy optimization. We use the no-thinking setting for hierarchical recognition, in which the model directly predicts the requested category without requiring an explicit chain-of-thought response. Unlike SFT and DFT, which optimize a fixed reference response under teacher forcing, RFT optimizes the policy according to sequence-level feedback from its generated predictions.

For all three objectives, PhyRC preserves the corresponding base post-training objective and introduces relation-aware supervision as a complementary signal. Rather than replacing SFT, DFT, or RFT, it explicitly structures competition among the candidate hypotheses according to their taxonomic relations and regularizes the associated visual representations.

Under RFT, we additionally compare against TARA He et al. (2026b), a recent taxonomy-aware method for hierarchical visual recognition with LMMs. TARA incorporates hierarchical structure through representation alignment, providing a relevant comparison for taxonomy-aware post-training. Unless otherwise specified, comparison methods do not use the relation-calibrated candidate competition or phylogeny-consistent visual regularization introduced by PhyRC.

A.2 Taxonomy Construction and Training Details

Taxonomy source. Taxonomic relations are derived from the hierarchical category annotations associated with each dataset rather than retrieved from an external knowledge base during training. For each category, we use its annotated lineage to construct the corresponding taxonomy.

Relation extraction. Given a target category c_y and a competing candidate c_j, we measure their taxonomic relatedness using the depth of their lowest common ancestor (LCA). For a query at level

ℓ, we define

$$r_j = 1 - h_{\mathcal{T}}(c_y, c_j), \tag{13}$$

where $h_{\mathcal{T}}(c_y, c_j)$ denotes the normalized LCA-depth proximity defined in Eq. 2. A smaller r_j indicates a closer taxonomic relation, whereas a larger r_j corresponds to a more distant candidate.

Decision-space margin. PhyRC maps the target–competitor relation to a candidate-specific separation margin,

$$m_j = \mathrm{clip}\left(0.05 + 0.10 r_j,\ 0.05,\ 0.20\right). \tag{14}$$

For the normalized relation values used in our experiments, the margins are bounded within $[0.05, 0.15]$, with the attainable values determined by the discrete taxonomy structure. The structured decision objective is weighted by $\lambda_{\mathrm{dec}} = 0.20$.

Visual-space margin. For the visual objective, we select same-level near and far references according to their taxonomic relations to the anchor. Using cosine distance $d(\mathbf{z}_i, \mathbf{z}_j) = 1 - \cos(\mathbf{z}_i, \mathbf{z}_j)$, the visual margin is

$$m_{\mathrm{vis}} = \mathrm{clip}\left(0.05 + 0.10\left(r_{\mathrm{far}} - r_{\mathrm{near}}\right), 0.05, 0.20\right), \tag{15}$$

and the corresponding loss is

$$\mathcal{L}_{\mathrm{vis}} = \mathrm{softplus}\left(d_{\mathrm{near}} - d_{\mathrm{far}} + m_{\mathrm{vis}}\right). \tag{16}$$

We use $\lambda_{\mathrm{vis}} = 0.05$ in the canonical configuration.

PCVR extracts image-token hidden states from the final language-transformer layer, applies mean pooling and a learned hierarchy-level embedding, and maps them through a projection head with a 1,024-dimensional hidden layer and a 768-dimensional normalized output. Near and far references are drawn from the current distributed minibatch and a detached FIFO feature queue with capacity 4,096.

Controlled separation variants. All variants use the same full-candidate decision objective and differ only in how the target–competitor separation margins are constructed. FCC removes explicit separation requirements by setting

$$m_j = 0, \qquad j \neq y. \tag{17}$$

UCS retains the same full-candidate competition but assigns every negative candidate an identical positive margin,

$$m_j = m_{\mathrm{UCS}} = 0.0883984945, \qquad j \neq y. \tag{18}$$

This value is the exact mean of the taxonomy-conditioned margins over all 28,815 valid target–negative pairs in the training set. UCS is therefore mean-margin matched to TCS while removing any dependence on the target–competitor taxonomic relation. The margin is fixed throughout training and is not further clipped or scaled.

TCS instead uses the true target–competitor relation,

$$m_j^{\mathrm{TCS}} = \mathrm{clip}\left(0.05 + 0.10 r_j,\ 0.05,\ 0.20\right). \tag{19}$$

RRA preserves the same taxonomy-derived margin values as TCS but disrupts their correspondence to individual competitors. For each query, we first compute the three TCS margins and then apply a deterministic permutation over the negative candidates,

$$m_j^{\mathrm{RRA}} = m_{\pi(j)}^{\mathrm{TCS}}, \qquad j \neq y. \tag{20}$$

FCC, UCS, RRA, and TCS otherwise share the same candidate scorer, full-candidate coverage, decision-loss weight, and optimization budget, and none includes the visual regularization branch.

Optimization. We use separate ShardedFP32AdamW optimizers for the structured and policy updates, with independent Adam moments. The policy optimizer uses an initial learning rate of 1×10^{-6} and a linear schedule over 604 policy updates with no warm-up. The structured update uses a constant learning rate of 1×10^{-6} for the Qwen parameters and 1×10^{-5} for the visual projection module, without a learning-rate scheduler. Training uses bfloat16 precision. The RFT objective uses gradient clipping at 1.0, whereas the structured update does not apply gradient clipping.

For RFT rollouts, we use a group size of four, sampling temperature 0.7, top-$p = 0.8$, top-$k = 20$, and a KL coefficient of 0.04. The maximum prompt and completion lengths are both 1,024 tokens.

Trainable parameters. The structured objective propagates gradients through the multimodal model parameters involved in visual-token representations as well as the learned visual projector and hierarchy-level embedding. In particular, the visual objective propagates gradients through the vision encoder, multimodal merger, and language-model representations, while historical features stored in the FIFO queue are detached and receive no gradient.

Separation from textual prompting. The taxonomy tree, LCA depths, and derived relation values are not provided to the LMM as textual input. They are used only during post-training to construct the structured objectives.

A.3 Prompt Format and Candidate Scoring Protocol

Hierarchical recognition prompt. For all candidate-conditioned HVR experiments, the model receives an image, a hierarchy-specific question, and the complete candidate set at the queried taxonomic level. All candidates therefore compete under the same multimodal context. Unless otherwise specified, we use $K = 4$ candidates as the canonical setting. The generic prompt template is:

```
<image>
What is the [LEVEL] of the object in the image?
Candidates:  [C1], [C2], ..., [CK]
Please directly output the answer.
```

For example, an order-level query can be instantiated as:

```
<image>
What is the order of the animal in the image?
Candidates:  Accipitriformes, Falconiformes,
Strigiformes, Crocodylia
Please directly output the answer.
```

The hierarchy level is specified directly in the question, while the candidate set contains only categories from that level. The same input format is used across the evaluated post-training objectives; their differences lie in the learning objective rather than in the task prompt.

Teacher-forced candidate scoring. PhyRC evaluates all candidates under the same image, question, and candidate context. For each candidate c_j, the shared context is kept fixed and only the assistant continuation is replaced by the corresponding category name. For the example above, the four candidate-specific continuations are:

```
Assistant:  Accipitriformes
Assistant:  Falconiformes
Assistant:  Strigiformes
Assistant:  Crocodylia
```

Thus, the model scores category names directly rather than predicting an option letter. Let $a_j = (a_{j,1}, \ldots, a_{j,T_j})$ denote the tokenized answer corresponding to candidate c_j. Its score is the length-normalized teacher-forced log-likelihood

$$s_j = \frac{1}{T_j} \sum_{t=1}^{T_j} \log p_\theta \left(a_{j,t} \mid I, q, \mathcal{C}, a_{j,<t} \right), \tag{21}$$

where I is the input image, q is the hierarchy-specific question, and $\mathcal{C} = \{c_1, \ldots, c_K\}$ is the complete candidate set. Only assistant-answer tokens contribute to s_j; prompt tokens, image tokens, padding tokens, and other context tokens are excluded. Length normalization reduces the bias toward category names with fewer tokens.

In the canonical $K = 4$ setting, the four candidate continuations are evaluated under the same multimodal context to obtain $\{s_1, s_2, s_3, s_4\}$. The predicted category is

$$\hat{c} = \arg\max_{c_j \in \mathcal{C}} s_j. \tag{22}$$

These candidate scores are also the quantities used by the taxonomic ordinal constraints in the decision-space objective.

Taxonomic information is not included in the prompt. No LCA information, taxonomic distance, relative-relation label (e.g., close or distant), or near–far annotation is provided to the model through the textual prompt. Such information is used only to construct the structured training objectives. Consequently, PhyRC injects taxonomic structure through supervision rather than through explicit taxonomy hints at inference time.

Evaluation and cross-dataset usage. Evaluation follows the same candidate-conditioned input format and candidate scoring procedure. The same template is also used on NABirds, with the queried hierarchy level and candidate categories replaced according to its native taxonomy. No additional prompt engineering is introduced for cross-dataset evaluation.

A.4 Implementation Details and Training Procedure

Candidate competition and margin calibration. With zero margins, Eq. (4) reduces to cross-entropy on the generative candidate scores:

$$\mathcal{L}_{\text{set}} = -\log \frac{\exp(s_y)}{\sum_{j=1}^{M} \exp(s_j)} = \log \left[1 + \sum_{j \neq y} \exp(s_j - s_y) \right]. \tag{23}$$

Taxonomy-conditioned margins shift each candidate's contribution according to its relation to the target. The implementation uses $m_j = \text{clip}(m_0 + \gamma r_j, m_{\min}, m_{\max})$. For the coefficients below, clipping leaves the linear margin of Eq. (3) unchanged throughout the normalized relation range.

Training configuration. For the main Qwen3-VL-2B experiments, PhyRC is integrated into the canonical rollout-based RFT pipeline. Each training epoch contains 604 structured updates and 604 policy updates, whereas the vanilla RFT baseline contains only the 604 policy updates. Thus, the two methods share the policy-update budget but are not matched in total optimizer updates; compute-controlled variants are analyzed separately in Appendix B.8.

RFT uses a rollout group size of four with temperature 0.7, top-p 0.8, and top-k 20. The KL coefficient is 0.04. For structured supervision, we set $\lambda_{\text{dec}} = 0.20$ and $\lambda_{\text{vis}} = 0.05$. The taxonomy margin is $m_j = \text{clip}(0.05 + 0.10\, r_j, 0.05, 0.20)$. For Qwen3-VL-2B, images are resized to 384×384; other backbones use their corresponding preprocessing configurations. All experiments were conducted on a single server equipped with eight NVIDIA RTX PRO 6000 GPUs.

Visual representation and memory. PCVR uses image-token hidden states from the final language-transformer layer, followed by mean pooling, a learned hierarchy-level embedding, and a projection head with a 1,024-dimensional hidden layer and a 768-dimensional normalized output. Same-level near and far references are selected according to their LCA relations from the current minibatch and a detached FIFO feature queue of size 4,096. Queued features are not back-propagated through and are refreshed online with recent visual embeddings.

Integration with base objectives. PhyRC adds one structured update before each corresponding base-objective update. For SFT and DFT, the structured update is followed directly by the respective supervised update. For RFT, rollouts are first collected under the current policy; after the structured update, the recorded response tokens are re-evaluated under the updated policy for the policy and KL objectives, without resampling new rollouts.

Algorithm 1 PhyRC-Integrated Post-Training

Require: Minibatch $\mathcal{B} = \{(x, q_\ell, \mathcal{C}_\ell, y)\}$, taxonomy $\mathcal{T}$, feature queue $\mathcal{Q}$, and base objective $\mathcal{L}_{\text{base}}$
Ensure: Updated LMM parameters θ and visual parameters ϕ

if $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{RFT}}$ **then**
  Collect rollout responses $\mathcal{R}$ under the current policy θ.
end if
Compute the length-normalized score of every candidate:

$$s_j = \frac{1}{T_j} \sum_{t=1}^{T_j} \log p_\theta \left(a_{j,t} \mid x, q_\ell, \mathcal{C}_\ell, a_{j,<t} \right).$$

Compute taxonomy-conditioned margins for all $j \neq y$:

$$r_j = 1 - h_{\mathcal{T}}(c_y, c_j), \qquad m_j = \text{clip}\left(m_0 + \gamma r_j, m_{\min}, m_{\max}\right).$$

Compute candidate-competition loss:

$$\mathcal{L}_{\text{dec}} = \log \left[1 + \sum_{j \neq y} \exp(s_j - s_y + m_j) \right].$$

Extract visual embeddings

$$\mathbf{z}_x = \text{Norm}\left(g_\phi \left[\text{MeanPool}(\mathbf{H}_x^{(L)}); \, \mathbf{e}_\ell \right] \right).$$

Select same-level near/far references from $\mathcal{B} \cup \mathcal{Q}$ according to taxonomic proximity and compute

$$\mathcal{L}_{\text{vis}} = \text{softplus}\left(d(\mathbf{z}_y, \mathbf{z}_{\text{near}}) - d(\mathbf{z}_y, \mathbf{z}_{\text{far}}) + m_{\text{vis}} \right).$$

$\mathcal{L}_{\text{struct}} = \lambda_{\text{dec}} \mathcal{L}_{\text{dec}} + \lambda_{\text{vis}} \mathcal{L}_{\text{vis}}$.
Update (θ, ϕ) using $\mathcal{L}_{\text{struct}}$.
if $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{RFT}}$ **then**
  Re-evaluate the recorded rollout tokens under the updated policy and apply the RFT policy/KL update.
else
  Apply the corresponding SFT or DFT update under the updated policy.
end if
Refresh $\mathcal{Q}$ with detached recent visual embeddings.
return θ, ϕ

A.5 Diagnostic Definitions and Evaluation Protocol

Target–competitor separation. We first examine whether the recognition gains of PhyRC correspond to a reorganization of competition among category candidates. As shown in Figure 3a, we compare matched RFT and PhyRC checkpoints of Qwen3-VL-2B-Instruct trained with seed 42 using the native length-normalized generative scores under the same image, question, and candidate set. The analysis is conducted over all $57{,}954$ candidate-conditioned evaluation queries, including $25{,}626$ from iNat-Plant and $32{,}328$ from iNat-Animal, without further partitioning into Task-Seen and Task-Unseen subsets.

For each query, we define the target–hardest-competitor margin as

$$\Delta = s_y - \max_{j \neq y} s_j, \tag{24}$$

where s_y is the length-normalized teacher-forced log-likelihood of the target candidate and s_j denotes that of competitor c_j. A negative margin, $\Delta < 0$, indicates that at least one competing candidate receives a higher score than the target. Figure 3a compares the distribution of Δ between RFT and PhyRC, together with its summary statistics and the corresponding mis-ranking rate. The displayed density curves are estimated using Gaussian kernel density estimation with Scott's band-

width rule on an 800-point evaluation grid. No bootstrap resampling or uncertainty interval is used for this diagnostic.

Hierarchy-level mis-ranking. Figure 3b further examines how candidate mis-ranking varies across the taxonomic hierarchy using the same RFT and PhyRC checkpoints, evaluation data, and seed as Figure 3a. For each taxonomic level ℓ, from Phylum to Species, we define

$$\mathrm{MR}_\ell = \frac{100}{|\mathcal{Q}_\ell|} \sum_{q \in \mathcal{Q}_\ell} \mathbb{I}[\Delta_q < 0], \tag{25}$$

where $\mathcal{Q}_\ell$ contains the evaluation queries whose target lies at level ℓ. Each level contains $9{,}659$ queries, and RFT and PhyRC are evaluated on identical queries and candidate sets using the same length-normalized teacher-forced candidate scoring procedure. Lower values indicate fewer cases in which a competing category receives a higher score than the ground-truth candidate. No error bars or bootstrap resampling are used in this analysis.

This diagnostic focuses on the local ranking relation between the target and its strongest competitor, rather than directly measuring the final hierarchical accuracy. Specifically, MR_ℓ counts cases where the ground-truth candidate is not ranked first among the $K = 4$ candidates. Although both this diagnostic and POR are derived from candidate-ranked predictions, they characterize different aspects of performance: mis-ranking measures target–hardest-competitor failures, whereas POR aggregates prediction correctness according to hierarchical matching.

Relation-conditioned decision separation. To isolate the effect of taxonomy-calibrated separation from simply imposing a positive margin, Figure 3c compares the taxonomy-calibrated margin variant with the corresponding uniform-margin control. Both use Qwen3-VL-2B-Instruct checkpoints trained with seed 42 under the same full-candidate competition protocol.

For each matched target–competitor pair, we measure the change in decision gap as

$$\Delta_{\mathrm{rel}} = (s_y - s_j)_{\mathrm{taxonomy}} - (s_y - s_j)_{\mathrm{uniform}}, \tag{26}$$

where positive values indicate that taxonomy-calibrated margins increase the target–competitor separation relative to uniform margins.

Competitors are grouped according to the normalized taxonomic relation r_j into four intervals: *Very Close*, $[0, 0.25)$; *Close*, $[0.25, 0.50)$; *Medium*, $[0.50, 0.75)$; and *Far*, $[0.75, 1]$. For iNat-Plant, these bins contain $9{,}527$, $25{,}689$, $38{,}828$, and $2{,}834$ target–competitor pairs, respectively. For iNat-Animal, the corresponding counts are $17{,}237$, $34{,}345$, $34{,}060$, and $11{,}342$.

The error bars in Figure 3c denote 95% percentile confidence intervals obtained from $10{,}000$ bootstrap resamples with random seed 20260916. Resampling is performed at the image-ID cluster level. This diagnostic tests whether taxonomy-calibrated margins induce relation-dependent changes in target–competitor decision separation rather than uniformly enlarging all score gaps. In particular, far competitors exhibit a larger positive separation shift than very-close competitors, while the intermediate relation bins are not strictly monotonic. The constraint is applied independently between the target and each competitor and does not impose a total ordering among negative candidates.

Shared visual-space diagnostics. We additionally examine whether PhyRC is associated with a visual representation geometry that better reflects taxonomic relatedness. The visual diagnostics compare matched RFT and full PhyRC checkpoints of Qwen3-VL-2B-Instruct trained with seed 42. To ensure a direct comparison in a representation space shared by both models, we extract the final LLM-layer hidden states (`hidden_states[-1]`) at image-token positions identified by `processor.image_token_id`. For each image x, these states are mean-pooled and normalized in FP32:

$$\widetilde{\mathbf{h}}_x = \mathrm{Norm}\left(\mathrm{MeanPool}\left(\mathbf{H}_x^{(L)}\right)\right), \tag{27}$$

where $\mathbf{H}_x^{(L)}$ denotes the final-layer hidden states associated with image tokens. Cosine distance is then used to compare representations.

Importantly, these diagnostics operate on the shared backbone representations and do not use the hierarchy-level embedding, the PhyRC-specific projection head, or the FIFO memory queue at evaluation time. These components are used only during PhyRC training.

The analysis uses a fixed manifest containing $13{,}129$ anchor–near–far triplets, corresponding to $26{,}258$ comparison arms, where the near and far examples satisfy a strict ordering in taxonomic distance. For each triplet, we evaluate the ordering-satisfaction probability

$$P_{\mathrm{ord}} = \Pr\left[d(\widetilde{\mathbf{h}}_y, \widetilde{\mathbf{h}}_{\mathrm{near}}) < d(\widetilde{\mathbf{h}}_y, \widetilde{\mathbf{h}}_{\mathrm{far}})\right], \tag{28}$$

and the corresponding local metric gap

$$G_{\mathrm{metric}} = d(\widetilde{\mathbf{h}}_y, \widetilde{\mathbf{h}}_{\mathrm{far}}) - d(\widetilde{\mathbf{h}}_y, \widetilde{\mathbf{h}}_{\mathrm{near}}). \tag{29}$$

A positive G_{metric} indicates that the local taxonomic ordering is satisfied. The visual analyses additionally examine distances across taxonomic-relation groups as a complementary diagnostic. Uncertainty is reported using 95% percentile confidence intervals from $10{,}000$ bootstrap resamples, with resampling performed at the anchor-image cluster level using seed 42.

These diagnostics evaluate local ordinal consistency rather than assuming that the complete visual representation space reconstructs the taxonomy as a globally monotonic geometry.

B Supplementary Experiments and Analyses

B.1 Reference Comparison with Reported HiR^2 Results

HiR^2 (He et al., 2026a) is closely related to our work in studying hierarchy-aware post-training for multimodal hierarchical recognition. For completeness, we report its published Qwen2.5-VL-3B results on iNat21-Plant and iNat21-Animal in Table 6, together with the corresponding results obtained under our protocol.

We emphasize that the HiR^2 numbers are taken directly from the original paper and are *not* reproduced using our training and evaluation pipeline. Although both studies use Qwen2.5-VL-3B and the same iNaturalist benchmarks, the original HiR^2 paper reports results under its own Base/Novel hierarchy partition, whereas our experiments use the Task-Seen/Task-Unseen partition defined by exposure during task-specific post-training. We cannot verify that the two studies use identical training-data construction, candidate hypotheses, prompting, optimization budgets, or evaluation implementations. Accordingly, these numbers are included only as contextual reference and should not be interpreted as a strictly controlled head-to-head comparison.

For consistency with the metrics reported by HiR^2, we show HCA, POR, S-POR, and TOR in this comparison. Leaf accuracy is omitted because it is not reported in the corresponding HiR^2 table.

Table 6: Reference comparison on Qwen2.5-VL-3B. Panel (a) reports results obtained under our Task-Seen/Task-Unseen protocol, whereas Panel (b) reproduces the Base/Novel results reported in the original HiR^2 paper (He et al., 2026a). The two partitioning and training protocols are not assumed to be equivalent; HiR^2 results are therefore provided for contextual reference only rather than as controlled comparisons. All values are percentages.

Dataset	Method	Result Source	Split 1				Split 2			
			HCA	POR	S-POR	TOR	HCA	POR	S-POR	TOR
(a) Our protocol: Split 1 = Task-Seen, Split 2 = Task-Unseen										
iNat-Plant	SFT	This work	18.32	73.02	57.78	57.99	17.29	72.25	55.04	55.82
	SFT + PhyRC	This work	**24.62**	**77.97**	**63.96**	**65.02**	**23.57**	**76.89**	**60.67**	**62.43**
	DFT	This work	16.35	70.94	56.15	55.59	16.32	71.27	54.32	54.87
	DFT + PhyRC	This work	**23.83**	**77.00**	**61.81**	**63.17**	**23.01**	**76.03**	**60.67**	**61.73**
iNat-Animal	SFT	This work	23.39	77.00	68.61	65.88	21.20	75.74	65.88	63.11
	SFT + PhyRC	This work	**30.03**	**80.09**	**72.17**	**69.99**	**28.03**	**79.16**	**69.03**	**67.32**
	DFT	This work	21.24	75.85	67.20	64.38	18.11	74.42	64.56	61.78
	DFT + PhyRC	This work	**28.98**	**79.44**	**71.29**	**69.12**	**24.57**	**77.74**	**67.62**	**65.62**
(b) HiR^2 paper: Split 1 = Base, Split 2 = Novel (reported results)										
iNat-Plant	SFT	HiR^2 paper	9.98	63.63	46.01	45.69	10.86	63.74	46.15	45.16
	SFT + HiR^2	HiR^2 paper	**11.90**	**66.65**	**50.87**	**49.76**	**13.06**	**67.16**	**50.69**	**49.69**
	DFT	HiR^2 paper	8.15	62.45	40.47	41.98	9.32	63.26	41.35	42.64
	DFT + HiR^2	HiR^2 paper	**11.66**	**66.07**	**49.83**	**48.81**	**12.87**	**66.72**	**50.12**	**49.10**
iNat-Animal	SFT	HiR^2 paper	14.06	71.41	60.40	58.03	15.74	72.12	61.42	58.86
	SFT + HiR^2	HiR^2 paper	**18.07**	**73.35**	**64.27**	**61.17**	**17.60**	**73.77**	**64.63**	**61.49**
	DFT	HiR^2 paper	14.69	71.90	60.27	58.81	16.52	72.65	61.64	60.04
	DFT + HiR^2	HiR^2 paper	**16.33**	**72.54**	**62.44**	**59.58**	**17.04**	**73.16**	**63.01**	**60.49**

B.2 Additional Ablation Study

Component contribution. We first investigate the contribution of individual supervision components in PhyRC. Starting from the RFT baseline, we construct independent variants equipped with target-only supervision (TOS), full-candidate competition (FCC), phylogeny-consistent visual regularization (PCVR), or taxonomy-calibrated separation (TCS). All variants are trained separately under the same protocol, allowing each component's contribution to be evaluated without cumulative effects.

As shown in Table 7, each component provides complementary benefits over the RFT baseline, although the gains vary across datasets and evaluation metrics. TOS and FCC improve candidate-level recognition by introducing explicit supervision over target responses and candidate competition, respectively. TCS further improves several recognition metrics by incorporating taxonomy relations into the separation objective. Although PCVR alone does not consistently improve all candidate-level metrics, combining PCVR with TCS yields the complete PhyRC objective, which achieves the strongest overall performance across the two datasets. These results suggest that taxonomy-calibrated decision constraints and phylogeny-consistent visual regularization provide complementary signals when jointly optimized.

Effect of taxonomy-calibrated separation. We further isolate the design of the separation objective under the same full-candidate competition setting. Instead of assigning identical margins to all competitors, PhyRC introduces taxonomy-calibrated margins according to inter-category relationships.

Table 7 shows that TCS improves several recognition metrics over uniform candidate separation (UCS) and randomized relation assignment (RRA), particularly on HCA, Leaf, and hierarchy-aware metrics. However, the relative advantage varies across individual metrics, reflecting the different sensitivities of recognition and hierarchy-based evaluations to the separation objective.

To examine the robustness of this observation, we repeat the comparison with three independent training seeds (42–44). As reported in Table 8, TCS achieves the best average performance on most metrics across both datasets while maintaining comparable variance to other relation calibration strategies. These results indicate that taxonomy-aware separation provides a more stable training signal than relation-agnostic alternatives, while its effect may vary across individual evaluation criteria.

Table 7: Ablation of PhyRC on Qwen3-VL-2B using the Task-Seen iNat-Plant and iNat-Animal evaluation sets. Panel (a) decomposes the major supervision components, where PhyRC combines TCS and PCVR. Panel (b) isolates the separation design under the same full-candidate competition protocol. All variants are independently trained with seed 42 under the canonical ablation protocol. Results are percentages. Boldface and underlining denote the highest and second-highest results, respectively, for each metric across both panels.

	iNat-Plant					*iNat-Animal*				
Variant	HCA ↑	Leaf ↑	POR ↑	S-POR ↑	TOR ↑	HCA ↑	Leaf ↑	POR ↑	S-POR ↑	TOR ↑
(a) Component contribution										
RFT	9.88	31.88	65.67	52.28	49.43	9.47	29.35	67.73	58.91	55.39
+TOS	15.76	35.79	72.52	60.44	59.02	13.31	35.79	71.38	<u>61.64</u>	59.48
+FCC	16.85	37.65	72.74	<u>60.83</u>	58.90	14.15	37.98	71.06	59.09	59.33
+PCVR	11.87	31.99	68.20	55.07	52.86	9.84	29.70	67.93	58.81	55.65
+TCS	17.08	<u>38.29</u>	**73.83**	59.88	<u>60.38</u>	<u>14.60</u>	<u>39.22</u>	<u>71.73</u>	60.90	<u>59.63</u>
+PhyRC (TCS + PCVR)	**17.90**	**39.08**	<u>73.30</u>	**61.38**	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**
(b) Relation calibration under full-candidate competition										
+FCC	16.85	37.65	72.74	<u>60.83</u>	58.90	14.15	37.98	71.06	59.09	59.33
+UCS	16.91	37.06	72.91	60.47	59.33	14.15	38.19	71.12	59.09	58.49
+RRA	16.91	37.23	72.98	60.65	59.58	12.87	38.63	71.15	57.61	58.80
+TCS	<u>17.08</u>	<u>38.29</u>	**73.83**	59.88	<u>60.38</u>	<u>14.60</u>	<u>39.22</u>	<u>71.73</u>	60.90	<u>59.63</u>

Table 8: Multi-seed robustness of relation calibration on Qwen3-VL-2B using the Task-Seen iNat-Plant and iNat-Animal evaluation sets. All variants follow the same full-candidate competition protocol. Results are reported as mean±sample standard deviation over three independent training seeds (42–44). Boldface denotes the best mean result for each metric.

	iNat-Plant					*iNat-Animal*				
Variant	HCA ↑	Leaf ↑	POR ↑	S-POR ↑	TOR ↑	HCA ↑	Leaf ↑	POR ↑	S-POR ↑	TOR ↑
+FCC	16.79±0.41	37.65±0.51	72.74±0.44	60.09±0.74	59.10±0.53	13.33±0.83	37.91±0.34	71.06±0.97	59.07±0.59	58.85±0.83
+UCS	16.73±0.18	37.06±0.71	72.83±0.08	**60.47±0.58**	59.33±0.20	13.63±0.52	38.20±0.55	71.12±0.76	58.79±0.98	59.33±0.84
+RRA	16.68±0.36	36.99±0.48	72.83±0.28	60.23±0.37	59.44±0.38	13.33±1.09	38.64±0.15	71.09±1.09	57.93±3.94	58.75±1.41
+TCS	**17.08±0.14**	**38.20±0.48**	**73.32±0.51**	60.42±0.87	**59.75±0.63**	**14.28±0.51**	**38.87±0.55**	**71.73±0.82**	**59.92±0.98**	**59.63±0.93**

Table 9: Ablation of PhyRC on Qwen3-VL-2B using the Task-Unseen iNat-Plant and iNat-Animal evaluation sets. Panel (a) reports the contribution of individual supervision components under the Task-Unseen evaluation protocol. Panel (b) compares different candidate separation strategies under the same full-candidate competition setting. All variants are trained with seed 42 under the strict Task-Unseen protocol. Results are percentages. Boldface and underlining denote the highest and second-highest results, respectively, for each metric across both panels.

	iNat-Plant					*iNat-Animal*				
Variant	HCA ↑	Leaf ↑	POR ↑	S-POR ↑	TOR ↑	HCA ↑	Leaf ↑	POR ↑	S-POR ↑	TOR ↑
(a) Component contribution										
RFT	9.90	34.45	64.67	50.19	48.34	7.10	27.21	65.85	56.48	52.87
+TOS	15.06	38.08	71.62	58.09	57.68	8.92	31.76	68.55	59.34	55.61
+FCC	15.06	40.45	71.87	58.11	57.94	<u>12.01</u>	37.49	69.97	59.42	<u>57.16</u>
+PCVR	11.58	34.87	66.76	53.42	51.32	7.92	29.12	65.79	56.45	52.14
+TCS	**15.90**	40.86	72.01	58.23	57.74	11.19	<u>38.31</u>	<u>70.35</u>	<u>59.93</u>	<u>57.16</u>
+PhyRC(TCS+PCVR)	**15.90**	**41.98**	**73.08**	**59.25**	**59.08**	**14.01**	**41.31**	**71.99**	**60.93**	**59.20**
(b) Relation calibration under full-candidate competition										
+FCC	15.06	40.45	71.87	58.11	57.94	<u>12.01</u>	37.49	69.97	59.42	<u>57.16</u>
+UCS	15.62	40.31	71.41	58.79	57.13	11.19	38.13	70.14	59.30	56.85
+RRA	<u>15.76</u>	<u>41.84</u>	<u>72.36</u>	<u>58.90</u>	<u>58.13</u>	10.37	37.49	69.81	59.34	56.62
+TCS	**15.90**	40.86	72.01	58.23	57.74	11.19	<u>38.31</u>	<u>70.35</u>	<u>59.93</u>	<u>57.16</u>

B.3 Qualitative Results

Figures 5–10 provide six qualitative examples from iNat-Plant and iNat-Animal, covering family-, genus-, and species-level recognition. For each query, we show the input image, the local taxonomy of the candidate set, and the mean answer-token log-likelihood assigned by RFT, TARA, and PhyRC(Ours). The examples are selected to highlight differences in how the methods separate the ground-truth category from competing candidates.

Across these examples, the baselines often assign competitive scores to taxonomically related distractors, resulting in small or incorrect target–competitor margins. PhyRC(Ours) more reliably shifts the ground-truth category toward the high-score region while maintaining relation-dependent separation from the competing candidates. This behavior is especially evident when the strongest distractor is taxonomically close to the target, where a small change in relative score can determine the final prediction. Together, these examples complement the quantitative results by illustrating how phylogeny-constrained supervision changes candidate competition in the generative decision space. Controlled relation-calibration experiments and mechanism analyses in the ablations and Figure 3c further examine whether these changes arise from taxonomy-aware relative separation rather than target-score optimization alone.

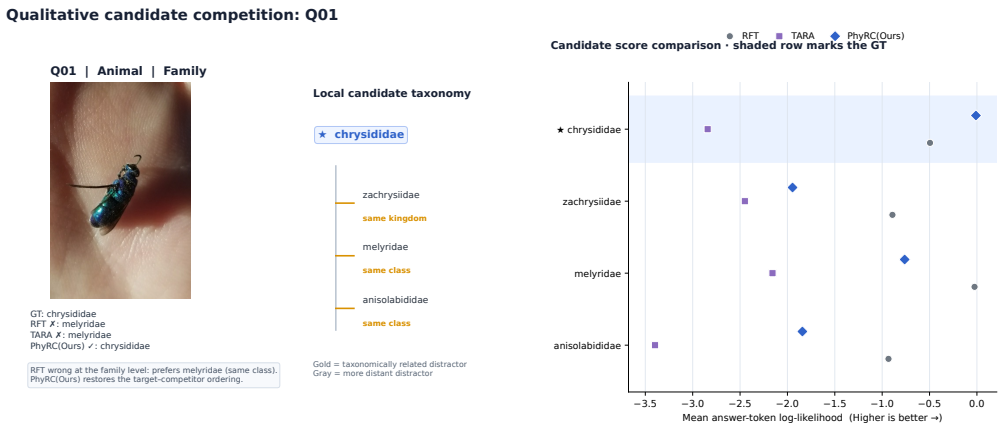


Figure 5: **Animal family-level example.** Candidate taxonomy and generative scores under RFT, TARA, and PhyRC(Ours).

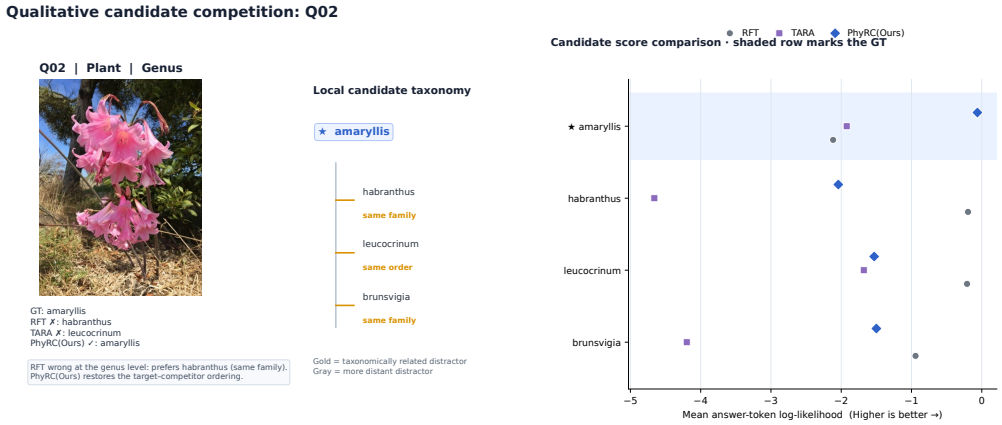


Figure 6: **Plantae genus-level example.** Candidate taxonomy and generative scores under RFT, TARA, and PhyRC(Ours).

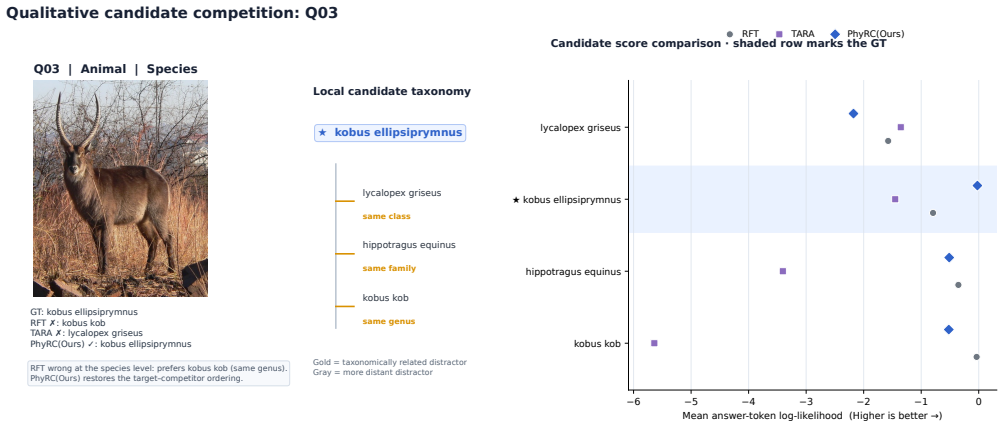


Figure 7: **Animal species-level example.** Candidate taxonomy and generative scores under RFT, TARA, and PhyRC(Ours).

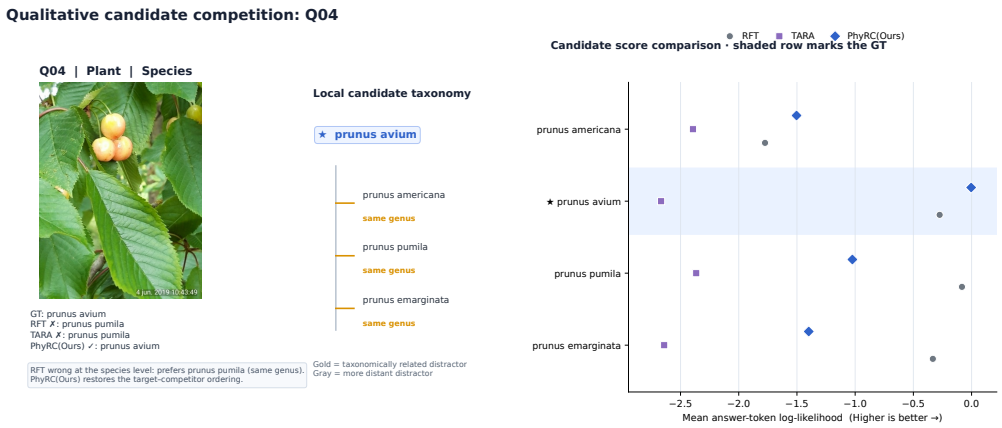


Figure 8: **Plantae species-level example.** Candidate taxonomy and generative scores under RFT, TARA, and PhyRC(Ours).

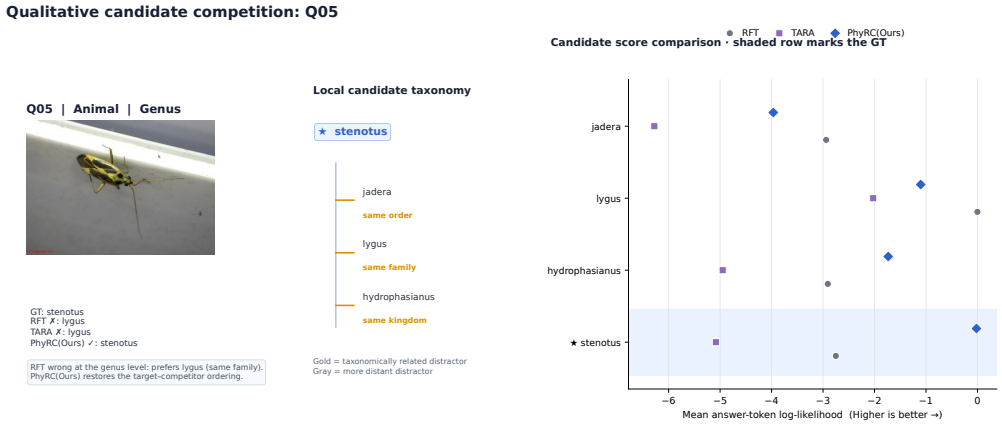


Figure 9: **Animal genus-level example.** Candidate taxonomy and generative scores under RFT, TARA, and PhyRC(Ours).

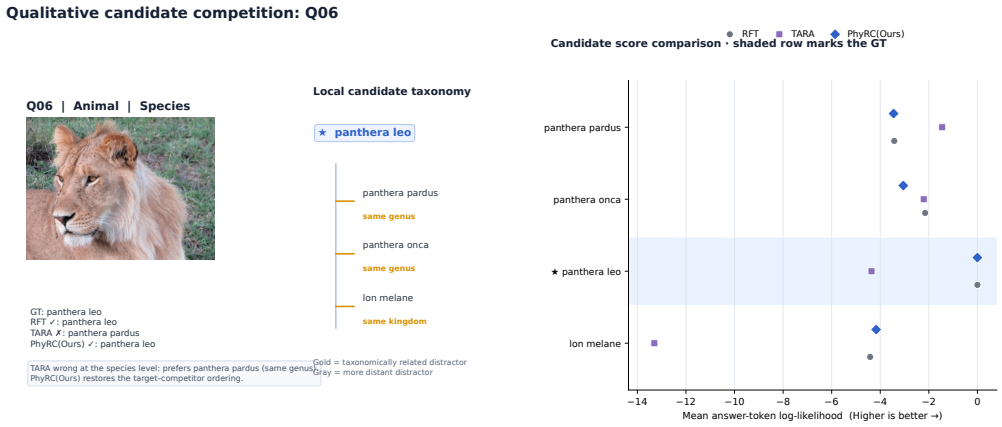


Figure 10: **Animal species-level example.** Candidate taxonomy and generative scores under RFT, TARA, and PhyRC(Ours).

B.4 Representation and Hyperparameter Sensitivity

Visual representation layer. We first examine whether PCVR is sensitive to the transformer layer from which visual representations are extracted. For Qwen3-VL-2B, we compare the middle, late, and final language-transformer layers, corresponding to layers 14, 21, and 28, respectively. All variants use seed 42 and otherwise identical training and evaluation settings.

Table 10: Effect of the visual representation layer used by PCVR on Qwen3-VL-2B under the Task-Seen evaluation setting. Middle, Late, and Final correspond to language-transformer layers 14, 21, and 28, respectively. All variants use seed 42 and otherwise identical training settings. Results are percentages; bold indicates the best result in each column.

Layer	*iNat-Plant*					*iNat-Animal*				
	HCA	Leaf	POR	S-POR	TOR	HCA	Leaf	POR	S-POR	TOR
Middle	16.94	38.07	73.18	60.35	59.46	14.20	38.61	71.84	61.19	59.78
Late	16.60	38.21	**73.30**	60.72	59.73	13.92	37.70	71.54	60.23	59.45
Final	**17.90**	**39.08**	**73.30**	**61.38**	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**

Performance remains broadly comparable across the three representation layers, suggesting that the method is not highly sensitive to the precise layer used for visual feature extraction. The final-layer representation nevertheless provides the most consistent results, achieving the best or tied-best performance across all reported metrics on both datasets. This empirical pattern motivates our use of the final language-transformer layer in the canonical configuration.

Hyperparameter sensitivity. We further examine sensitivity to the weight of the visual regularizer, λ_{vis}, and the taxonomy-margin scaling coefficient, γ, by varying one hyperparameter at a time. The canonical configuration uses $\lambda_{\text{vis}} = 0.05$ and $\gamma = 0.10$. The canonical rows in both sensitivity panels reuse the same verified final-layer checkpoint from the preceding layer ablation.

Table 11: Sensitivity to λ_{vis} and γ on Qwen3-VL-2B under the Task-Seen evaluation setting. Only one hyperparameter is varied at a time, with all remaining settings fixed to the canonical configuration. The canonical values are $\lambda_{\text{vis}} = 0.05$ and $\gamma = 0.10$. All variants use seed 42. Results are percentages; bold indicates the best result within each panel and column.

Parameter	Value	*iNat-Plant*					*iNat-Animal*				
		HCA	Leaf	POR	S-POR	TOR	HCA	Leaf	POR	S-POR	TOR
(a) Visual regularization weight											
λ_{vis}	0.025	16.94	37.73	**73.41**	61.10	59.99	14.83	38.82	71.89	60.36	59.91
	0.05	**17.90**	**39.08**	73.30	**61.38**	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**
	0.10	16.80	37.37	73.27	61.20	59.83	14.22	38.75	71.60	59.63	59.57
(b) Taxonomy-margin scaling											
γ	0.05	17.16	38.97	73.55	60.83	60.10	14.69	38.52	72.04	60.89	60.01
	0.10	**17.90**	**39.08**	73.30	61.38	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**
	0.20	16.88	37.51	**73.57**	**61.50**	60.23	14.25	38.33	72.03	61.54	60.06

Sensitivity analysis. Across the tested ranges, the canonical configuration provides the strongest overall performance, while neighboring hyperparameter values remain competitive on most metrics.

For λ_{vis}, the canonical value of 0.05 achieves the best results on most metrics across both datasets. Smaller or larger visual-regularization weights remain competitive and occasionally perform better on individual Plant metrics, indicating that the method is not highly sensitive to a narrowly tuned visual-loss weight. At the same time, the consistent advantage of the canonical setting on most metrics suggests that the visual regularization strength should not be treated as entirely interchangeable.

A similar pattern is observed for the taxonomy-margin scaling coefficient γ. The canonical value of 0.10 provides the strongest overall results, whereas smaller and larger values remain competitive on selected metrics. In particular, increasing γ does not lead to a general degradation, although the relative advantages vary across metrics. This suggests that PhyRC is reasonably stable to moderate

changes in the strength of taxonomy-conditioned separation without implying complete insensitivity to the margin scale.

Overall, the sensitivity results support the canonical choices $\lambda_{\text{vis}} = 0.05$ and $\gamma = 0.10$ as empirically stable settings rather than uniquely optimal values. Together with the representation-layer ablation, these results suggest that PhyRC remains effective under moderate changes to its visual representation and relation-calibration settings. We therefore use the final language-transformer layer, $\lambda_{\text{vis}} = 0.05$, and $\gamma = 0.10$ throughout the main experiments.

B.5 Candidate-Competition Ablations

Candidate coverage. Table 12 examines the effect of progressively increasing the number of competing candidates used by the structured objective. Performance generally improves as supervision expands from a single hard negative to the complete candidate set. Full-candidate supervision achieves the strongest overall results on both taxonomies, with only a minor exception on Plantae S-POR where Top-2 performs slightly better. This trend suggests that exposing the model to the complete set of competing hypotheses provides a more informative learning signal than optimizing against only a subset of negatives, supporting the use of full-candidate competition in PhyRC.

Table 12: Candidate-coverage ablation on Qwen3-VL-2B with seed 42 under the Task-Seen evaluation setting. Top-1 and Top-2 use one and two negative candidates, respectively, while Full uses all three negatives in the four-way candidate set. All other settings are held fixed. Bold indicates the best result in each column.

Coverage	*iNat-Plant*					*iNat-Animal*				
	HCA	Leaf	POR	S-POR	TOR	HCA	Leaf	POR	S-POR	TOR
Top-1	16.01	35.68	72.33	60.43	59.08	11.80	35.14	70.57	59.88	58.43
Top-2	17.25	37.11	73.21	**61.87**	60.39	14.13	37.79	71.85	61.70	59.86
Full	**17.90**	**39.08**	**73.30**	61.38	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**

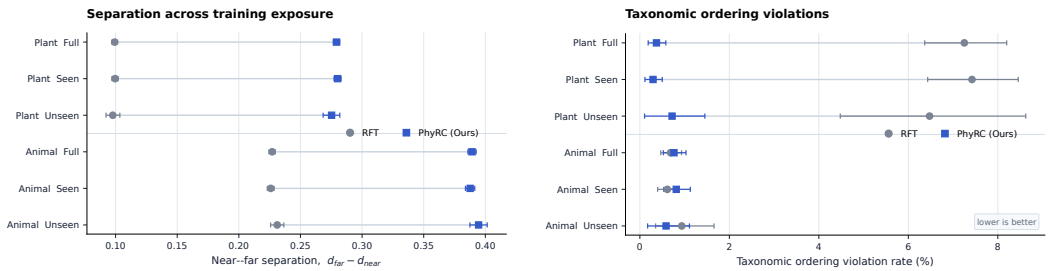
Listwise versus pairwise competition. Table 13 compares alternative formulations for learning from competing candidates. Pairwise Mean improves eight of the ten reported metrics over hardest-only supervision, while showing slightly lower S-POR on both taxonomies. The listwise formulation achieves the best performance across all reported metrics on both taxonomies by jointly modeling competition among the complete candidate set. This result suggests that preserving the relative structure among all candidate hypotheses is beneficial compared with decomposing competition into independent target–negative comparisons. Therefore, PhyRC adopts the listwise formulation to model multi-candidate competition.

Table 13: Comparison of candidate-ranking formulations on Qwen3-VL-2B under the Task-Seen evaluation setting. Hardest-only uses one negative, Pairwise Mean averages independent positive–negative losses, and Listwise jointly normalizes over the complete set. Bold indicates the best result in each column.

Ranking	*iNat-Plant*					*iNat-Animal*				
	HCA	Leaf	POR	S-POR	TOR	HCA	Leaf	POR	S-POR	TOR
Hardest-only	16.01	35.68	72.33	60.43	59.08	11.80	35.14	70.57	59.88	58.43
Pairwise Mean	16.80	37.99	72.97	60.11	59.11	13.85	38.12	71.43	59.43	59.29
Listwise	**17.90**	**39.08**	**73.30**	**61.38**	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**

B.6 Additional Representation-Space Diagnostics

To further examine whether the visual structure induced by PhyRC extends beyond the categories observed during task-specific post-training, we analyze the local taxonomic geometry separately on Task-Seen and Task-Unseen categories. We use the same mean-pooled and normalized final-layer image-token features for RFT and PhyRC, and evaluate taxonomy-defined anchor–near–far relations without introducing an additional learned projection.



(a) Near–far separation across category exposure.

(b) Taxonomic ordering violations.

Figure 11: **Additional diagnostics of phylogeny-consistent visual geometry.** PhyRC produces stronger near–far separation on both Task-Seen and Task-Unseen categories, showing that the learned local taxonomic structure is not restricted to category identities encountered during task-specific post-training. Consistently, PhyRC substantially reduces taxonomic ordering violations on iNat21-Plant and preserves the already low violation rates on iNat21-Animal. Together, these results support local taxonomic ordering in the visual representation space rather than a uniform expansion of feature distances.

Figure 11a shows that the increased separation between taxonomically near and far references persists across both Task-Seen and Task-Unseen categories. This suggests that the relational structure encouraged by PhyRC extends beyond the specific category identities encountered during task-specific post-training. Figure 11b provides a complementary ordinal view of the same geometry. PhyRC markedly reduces violations of the expected near-before-far ordering on iNat21-Plant, while maintaining the already low violation rates on iNat21-Animal. These trends are consistent with the intended *phylogeny-consistent hierarchical metric space*, in which local visual distances reflect taxonomic proximity. Importantly, the analysis concerns local relational ordering and does not assume that the complete taxonomy is embedded with globally monotonic distances. Detailed feature extraction and diagnostic definitions are provided in Appendix A.5.

B.7 Robustness Analysis

To evaluate the stability and reliability of PhyRC under diverse experimental variations, we conduct a comprehensive robustness analysis from multiple perspectives. Specifically, we examine the sensitivity of our method to training randomness, input candidate ordering, task-unseen exposure ratio, and hierarchical structural scale. These evaluations investigate whether the observed improvements arise from specific experimental configurations or reflect the intrinsic effectiveness of phylogeny-constrained relative competition.

B.7.1 Robustness to Training Seeds and Candidate Ordering

We further examine the robustness of PhyRC to training randomness and candidate presentation order on Qwen3-VL-2B. For training robustness, we independently repeat RFT and PhyRC with two additional random seeds. For ordering robustness, we evaluate each trained model using a fixed set of eight candidate-order permutations while keeping the underlying recognition instance unchanged. These two analyses separately characterize variation arising from independent training runs and from the presentation order of candidate categories.

Robustness across training seeds. Table 14 reports results over three independently trained seeds on both Task-Seen and Task-Unseen categories. Across iNat-Plant and iNat-Animal, PhyRC consistently achieves higher average performance than RFT on all reported metrics. Importantly, the same trend is preserved on Task-Unseen categories, indicating that the improvements are not restricted to categories encountered during post-training or to a particular training seed. Although the magnitude of cross-seed variation differs across metrics, the overall performance advantage of PhyRC remains stable across the evaluated runs.

Candidate-ordering protocol. To evaluate sensitivity to candidate presentation order, we use eight fixed, balanced permutations of each four-candidate set. Let $[A, B, C, D]$ denote the origi-

Table 14: Robustness across three independent training seeds (42–44) on Task-Seen and Task-Unseen categories with Qwen3-VL-2B. Values are reported as mean±sample standard deviation. All values are percentages.

Split	Dataset	Method	HCA ↑	Leaf ↑	POR ↑	S-POR ↑	TOR ↑
Task-Seen	*iNat-Plant*	RFT	9.94 ± 0.09	31.97 ± 0.66	65.82 ± 0.28	52.62 ± 0.65	49.63 ± 0.47
		PhyRC	$\mathbf{17.78 \pm 0.38}$	$\mathbf{39.04 \pm 0.16}$	$\mathbf{73.20 \pm 0.12}$	$\mathbf{60.97 \pm 0.59}$	$\mathbf{61.64 \pm 0.41}$
	iNat-Animal	RFT	9.56 ± 0.20	29.27 ± 0.13	67.75 ± 0.10	59.06 ± 0.15	55.32 ± 0.11
		PhyRC	$\mathbf{14.84 \pm 0.25}$	$\mathbf{39.19 \pm 0.84}$	$\mathbf{72.20 \pm 0.85}$	$\mathbf{61.87 \pm 0.47}$	$\mathbf{61.03 \pm 0.69}$
Task-Unseen	*iNat-Plant*	RFT	9.86 ± 1.60	33.10 ± 1.27	64.59 ± 0.89	50.36 ± 1.53	48.44 ± 1.38
		PhyRC	$\mathbf{15.90 \pm 0.42}$	$\mathbf{41.98 \pm 0.56}$	$\mathbf{73.11 \pm 0.38}$	$\mathbf{58.72 \pm 0.60}$	$\mathbf{58.89 \pm 0.25}$
	iNat-Animal	RFT	7.46 ± 0.48	26.72 ± 0.46	65.57 ± 0.54	56.66 ± 1.05	52.72 ± 0.61
		PhyRC	$\mathbf{14.01 \pm 0.27}$	$\mathbf{41.31 \pm 0.46}$	$\mathbf{72.13 \pm 0.32}$	$\mathbf{60.51 \pm 0.42}$	$\mathbf{58.94 \pm 0.26}$

nal candidate order. The exact permutations are listed in Table 15. The same permutation set is used across methods, datasets, and training seeds. The permutations are balanced such that each original candidate appears equally often at each of the four display positions.

For each permutation, the image, question wording, candidate identities, candidate set, and ground-truth category remain unchanged. Only the presentation order of the candidates is modified. The ground-truth option position is recomputed after reordering so that the same ground-truth category is evaluated under every permutation.

Table 15: Fixed candidate-order permutations used in the ordering-sensitivity evaluation. $[A, B, C, D]$ denotes the original candidate order.

Permutation	Candidate Order
P0	$[A, B, C, D]$
P1	$[B, C, D, A]$
P2	$[C, D, A, B]$
P3	$[D, A, B, C]$
P4	$[D, C, B, A]$
P5	$[A, D, C, B]$
P6	$[B, A, D, C]$
P7	$[C, B, A, D]$

Robustness to candidate ordering. We evaluate the eight candidate orders for each method, dataset, and training seed. For each seed and metric, we compute the sample standard deviation across the eight permutations and then average this quantity across seeds. We additionally report the largest within-seed worst-to-best range across candidate orders as a complementary measure of ordering sensitivity.

Table 16 summarizes the resulting variation. Both RFT and PhyRC exhibit some dependence on candidate presentation order, but the overall variation remains limited across most metrics and datasets. PhyRC does not exhibit a consistent increase in ordering sensitivity relative to RFT. The results therefore suggest that the observed performance trend is not primarily associated with a particular candidate arrangement, while small ordering effects remain and the models should not be interpreted as strictly permutation-invariant.

Takeaway. The two analyses provide complementary robustness evidence. Across independent training runs, the main performance trend of PhyRC is largely preserved, although stability varies across metrics. The candidate-ordering analysis further shows that both RFT and PhyRC retain some sensitivity to presentation order, without a consistent pattern of greater sensitivity for PhyRC. Together, these results support the overall robustness of the observed trend while indicating that training randomness and candidate ordering remain non-negligible sources of variation.

Table 16: Sensitivity to candidate ordering over the eight fixed permutations P0–P7 defined in Table 15. For each training seed, the metric-wise variability is measured by the sample standard deviation across the eight candidate orders, and the reported metric columns average this quantity across seeds 42–44. **Max Range** denotes the largest within-seed, within-metric worst-to-best range across the eight permutations for each method–dataset pair. All values are percentage points.

Method	Dataset	HCA	Leaf	POR	S-POR	TOR	Max Range
RFT	*iNat-Plant*	0.269	0.413	0.123	0.228	0.180	1.264
RFT	*iNat-Animal*	0.340	0.340	0.106	0.154	0.168	0.965
PhyRC	*iNat-Plant*	0.316	0.374	0.130	0.222	0.192	1.218
PhyRC	*iNat-Animal*	0.313	0.444	0.110	0.205	0.138	1.559

B.7.2 Robustness to Task-Unseen Category Exposure

Robustness under varying Task-Unseen category exposure. We further evaluate the robustness of PhyRC under different proportions of Task-Unseen categories. Specifically, we construct multiple category splits with approximately 20%, 40%, 60%, and 80% Task-Unseen species, which progressively reduce the availability of directly supervised categories during post-training.

Following the definition in the main paper, a species is considered Task-Unseen when its task label is strictly withheld from post-training in all taxonomic levels. Accordingly, images belonging to Task-Unseen species are excluded from post-training examples at all taxonomic levels. No images from Task-Unseen species are included in post-training, regardless of their associated taxonomic-level annotations. For each split, RFT and PhyRC are trained and evaluated using the same category partition, candidate construction protocol, and evaluation procedure.

As shown in Table 17, PhyRC consistently improves over the matched RFT baseline under different Task-Unseen category ratios across both biological domains. These results demonstrate that the effectiveness of PhyRC remains stable when the amount of directly observed category supervision varies substantially.

Table 17: Robustness to different Task-Unseen category ratios on Qwen3-VL-2B. We report HCA and Leaf accuracy on the corresponding Task-Unseen species. Δ denotes the absolute gain of PhyRC over the matched RFT baseline in percentage points. The realized Task-Seen/Task-Unseen ratios may differ slightly from the nominal settings because species are partitioned discretely.

Setting	Dataset	Task-Seen / Task-Unseen Ratio (%)		Task-Unseen Performance		Δ PhyRC–RFT	
		Supervised	Withheld	RFT HCA / Leaf	PhyRC HCA / Leaf	HCA	Leaf
Seen80	*iNat-Plant*	80.00	20.00	10.66 / 34.78	19.67 / 42.74	+9.01	+7.96
	iNat-Animal	79.60	20.40	7.19 / 25.84	12.92 / 41.13	+5.73	+15.29
Seen60	*iNat-Plant*	60.01	39.99	10.66 / 33.90	18.03 / 38.99	+7.37	+5.09
	iNat-Animal	60.00	40.00	8.54 / 26.77	13.69 / 36.47	+5.15	+9.70
Seen40	*iNat-Plant*	39.99	60.01	10.42 / 32.15	18.22 / 37.26	+7.80	+5.11
	iNat-Animal	40.00	60.00	8.41 / 27.65	13.27 / 35.08	+4.86	+7.42
Seen20	*iNat-Plant*	20.00	80.00	10.07 / 30.93	16.01 / 34.27	+5.94	+3.34
	iNat-Animal	20.01	79.99	9.72 / 27.68	13.50 / 34.45	+3.78	+6.77

Across all exposure ratios and both taxonomies, PhyRC consistently improves HCA and Leaf accuracy on Task-Unseen over the matched RFT baseline. The improvements remain effective under substantially different proportions of supervised and withheld categories, indicating that PhyRC is robust to substantial variations in the proportion of withheld categories during post-training.

Figure 12 further summarizes the gain trends under different Task-Unseen ratios. Although the improvement magnitude varies across datasets and metrics, the results do not exhibit a strictly monotonic relationship with the proportion of withheld categories. These observations suggest that the relational supervision introduced by PhyRC remains effective across diverse category-exposure conditions.

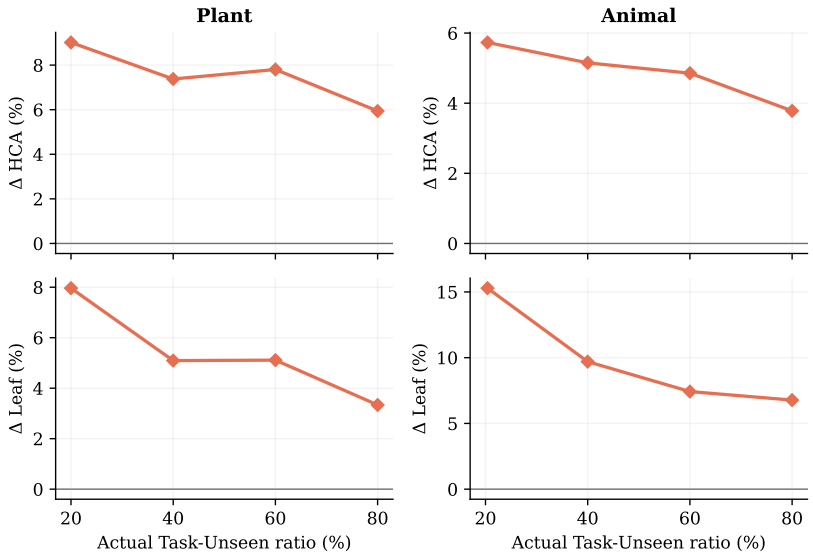


Figure 12: **PhyRC gains under different Task-Unseen category ratios.** We visualize the absolute HCA and Leaf improvements of PhyRC over the matched RFT baseline on Task-Unseen species under different category exposure settings. PhyRC maintains positive gains across all evaluated exposure ratios on both iNat-Plant and iNat-Animal, while the improvement magnitude is not strictly monotonic as the Task-Unseen proportion increases.

B.7.3 Robustness to Structural Scale

Candidate-set size. Table 18 studies the effect of different candidate-set sizes. For each setting, the number of candidates K is consistently applied to both training and evaluation, while keeping the training protocol otherwise unchanged. PhyRC maintains improvements over the corresponding baseline across different candidate-set sizes and both taxonomies, indicating that its effectiveness is not limited to the canonical four-way candidate formulation.

Meanwhile, increasing K does not lead to uniformly increasing performance. Larger candidate sets introduce more challenging competition among categories while also distributing supervision across more alternatives, resulting in a trade-off between competition difficulty and optimization complexity. These results demonstrate that PhyRC remains effective under varying scales of candidate competition, rather than relying on a fixed candidate-set size.

Table 18: Effect of candidate-set size under the Task-Seen evaluation setting. K is the number of options. The canonical four-way setting reuses the original training data; the other settings change candidate-set size under the reported matched training budget. Bold indicates the best result within each K.

K	Method	*iNat-Plant*					*iNat-Animal*				
		HCA	Leaf	POR	S-POR	TOR	HCA	Leaf	POR	S-POR	TOR
2	RFT	9.79	30.87	65.92	52.96	50.06	9.12	29.96	67.31	**57.84**	54.43
	RFT + PhyRC	**14.15**	**35.23**	**71.88**	**58.75**	**57.27**	**12.10**	**36.40**	**69.87**	56.26	**57.18**
4	RFT	9.88	31.88	65.67	52.28	49.43	9.47	29.35	67.73	58.91	55.39
	RFT + PhyRC	**17.90**	**39.08**	**73.30**	**61.38**	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**
8	RFT	8.02	29.91	62.40	48.72	45.04	8.70	29.05	66.99	**58.34**	54.12
	RFT + PhyRC	**15.98**	**36.69**	**72.99**	**59.69**	**59.13**	**12.75**	**36.42**	**70.71**	57.71	**58.61**
16	RFT	7.96	30.53	62.66	48.19	45.18	8.18	28.26	66.43	57.34	53.54
	RFT + PhyRC	**15.39**	**35.99**	**71.39**	**58.47**	**57.15**	**11.49**	**33.74**	**69.67**	**58.73**	**57.34**

Hierarchical supervision scale. Table 19 investigates the robustness of PhyRC under different hierarchical supervision scales. Specifically, we vary the number of taxonomy levels involved in constructing training supervision, from Species-only supervision (H1) to the full Phylum–Species hierarchy (H6). Since changing the hierarchy scale modifies the available structured supervision, each setting is treated as an independent training configuration. Both RFT and PhyRC are retrained under the same hierarchy setting with matched sample and optimization budgets.

The hierarchy depth controls the amount of taxonomic information available during training rather than introducing additional visual samples or new category instances. For higher-depth settings (H5/H6), additional taxonomic relations are derived from existing annotations associated with the same images.

PhyRC consistently improves over the corresponding RFT baseline across all six hierarchy depths, indicating that its benefit is not tied to a particular amount of taxonomic structure. Hierarchical recognition performance generally benefits from additional taxonomic supervision up to H4, while further increasing the hierarchy depth does not yield uniform gains across all metrics. Further increasing the hierarchy depth beyond H4 does not yield uniform gains across all metrics. H4 provides the strongest overall trade-off across the two datasets, while deeper settings remain competitive and occasionally perform better on individual hierarchy-oriented metrics.

These results suggest that exposing more taxonomic levels is not by itself sufficient to improve recognition under a fixed training budget. One possible explanation is that increasing hierarchy depth expands structural coverage while distributing a fixed amount of supervision across more levels, which may change the balance between fine-grained discrimination and broader hierarchical consistency. Overall, PhyRC remains effective across a wide range of supervision depths, and the H4 setting used in the main experiments provides a practical balance between these objectives.

Table 19: Effect of hierarchical supervision scale under the Task-Seen evaluation setting. We evaluate six hierarchy configurations (H1–H6), ranging from Species-only supervision (H1) to the full Phylum–Species hierarchy (H6), under four-way candidate questions and matched sample and optimization budgets. H4 (Order–Species) corresponds to the canonical hierarchy used in the main experiments. Bold indicates the better result between RFT and RFT + PhyRC within each depth.

Depth	Training Hierarchy	Method	*iNat-Plant*					*iNat-Animal*				
			HCA	Leaf	POR	S-POR	TOR	HCA	Leaf	POR	S-POR	TOR
H1	Species	RFT	6.98	31.88	60.44	45.19	41.65	7.67	29.63	65.37	55.39	51.45
		RFT + PhyRC	**13.62**	**41.84**	**68.47**	**54.15**	**52.12**	**11.05**	**43.72**	**68.83**	**57.44**	**53.92**
H2	Genus–Species	RFT	7.88	31.01	62.48	47.64	44.90	8.18	29.66	66.30	56.54	52.98
		RFT + PhyRC	**14.69**	**38.41**	**70.56**	**56.96**	**55.36**	**12.94**	**39.99**	**70.72**	**59.46**	**57.08**
H3	Family–Species	RFT	9.26	31.46	64.44	50.36	47.58	9.21	29.68	67.37	58.55	54.54
		RFT + PhyRC	**16.40**	**38.13**	**72.52**	**59.28**	**57.95**	**13.99**	**38.66**	**72.00**	**61.02**	**59.47**
H4	Order–Species	RFT	9.88	31.88	65.67	52.28	49.43	9.47	29.35	67.73	58.91	55.39
		RFT + PhyRC	**17.90**	**39.08**	**73.30**	**61.38**	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**
H5	Class–Species	RFT	9.68	31.37	65.73	52.28	49.26	9.42	29.98	67.46	58.67	54.70
		RFT + PhyRC	**17.19**	**36.49**	**73.73**	**63.54**	**61.14**	**13.17**	**37.68**	**71.47**	**60.15**	**59.38**
H6	Phylum–Species	RFT	9.23	31.29	65.19	51.93	48.70	9.19	29.45	67.22	58.32	54.45
		RFT + PhyRC	**15.76**	**36.61**	**72.86**	**61.20**	**59.05**	**13.97**	**37.70**	**72.51**	**64.35**	**60.45**

B.8 Controlled Analysis of Longer RFT Optimization

PhyRC introduces an additional structured update alongside each RFT policy update, resulting in a longer optimization schedule than vanilla RFT. We therefore examine whether its improvement can be explained simply by extending optimization of the vanilla RFT objective.

To this end, we introduce **Long-RFT**, which doubles the number of RFT policy updates from 604 to 1,208 while leaving the training objective unchanged. PhyRC instead uses 604 RFT policy updates together with 604 structured updates. This comparison tests whether longer optimization of the original RFT objective is sufficient to recover the gains achieved by PhyRC.

Effect of longer RFT optimization. Long-RFT improves over vanilla RFT on several metrics, indicating that extending optimization of the RFT objective can provide some benefit. However,

Table 20: Optimization schedules used to test whether the gains of PhyRC can be explained by longer RFT training.

Setting	RFT Policy Updates	Structured Updates	Total Updates
Vanilla RFT	604	0	604
Long-RFT	1,208	0	1,208
PhyRC	604	604	1,208

Table 21: Controlled comparison with longer RFT optimization on Qwen3-VL-2B under the Task-Seen evaluation setting. Long-RFT doubles the number of RFT policy updates while keeping the base objective unchanged. PhyRC uses the same total number of optimizer updates but replaces half of them with structured updates. Results are percentages, and bold denotes the best result in each column.

	iNat-Plant					*iNat-Animal*				
Method	HCA	Leaf	POR	S-POR	TOR	HCA	Leaf	POR	S-POR	TOR
Vanilla RFT	9.88	31.88	65.67	52.28	49.43	9.47	29.35	67.73	58.91	55.39
Long-RFT	11.82	33.20	67.21	54.59	52.03	10.07	31.08	68.09	58.71	55.37
RFT + PhyRC	**17.90**	**39.08**	**73.30**	**61.38**	**61.54**	**14.97**	**39.36**	**72.20**	**61.96**	**60.34**

longer RFT training remains consistently below PhyRC across the reported evaluation metrics. The difference is especially apparent on the primary recognition measures, while the gains from Long-RFT are generally smaller and less consistent on the hierarchy-aware metrics.

These results suggest that simply extending RFT optimization is insufficient to reproduce the performance achieved by PhyRC. The improvement therefore appears to depend on the additional structured supervision introduced by PhyRC rather than on longer training alone.