This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Taxonomy-Aware Representation Alignment for Hierarchical Visual Recognition with Large Multimodal Models

Hulingxiao He, Zhi Tan, Yuxin Peng*
Wangxuan Institute of Computer Technology, Peking University
hehulingxiao@stu.pku.edu.cn, tanzhi@stu.pku.edu.cn, pengyuxin@pku.edu.cn

Abstract

*A high-performing, general-purpose visual understanding model should map visual inputs to a taxonomic tree of labels, identify novel categories beyond the training set for which few or no publicly available images exist. Large Multimodal Models (LMMs) have achieved remarkable progress in fine-grained visual recognition (FGVR) for known categories. However, they remain limited in hierarchical visual recognition (HVR) that aims at predicting consistent label paths from coarse to fine categories, especially for novel categories. To tackle these challenges, we propose **T**axonomy-**A**ware **R**epresentation **A**lignment (**TARA**), a simple yet effective strategy to inject taxonomic knowledge into LMMs. TARA leverages representations from biology foundation models (BFMs) that encode rich biological relationships through hierarchical contrastive learning. By aligning the intermediate representations of visual features with those of BFMs, LMMs are encouraged to extract discriminative visual cues well structured in the taxonomy tree. Additionally, we align the representations of the first answer token with the ground-truth label, flexibly bridging the gap between contextualized visual features and categories of varying granularity according to user intent. Experiments demonstrate that TARA consistently enhances LMMs' hierarchical consistency and leaf node accuracy, enabling reliable recognition of both known and novel categories within complex biological taxonomies. Code is available at https://github.com/PKU-ICST-MIPL/TARA_CVPR2026.*

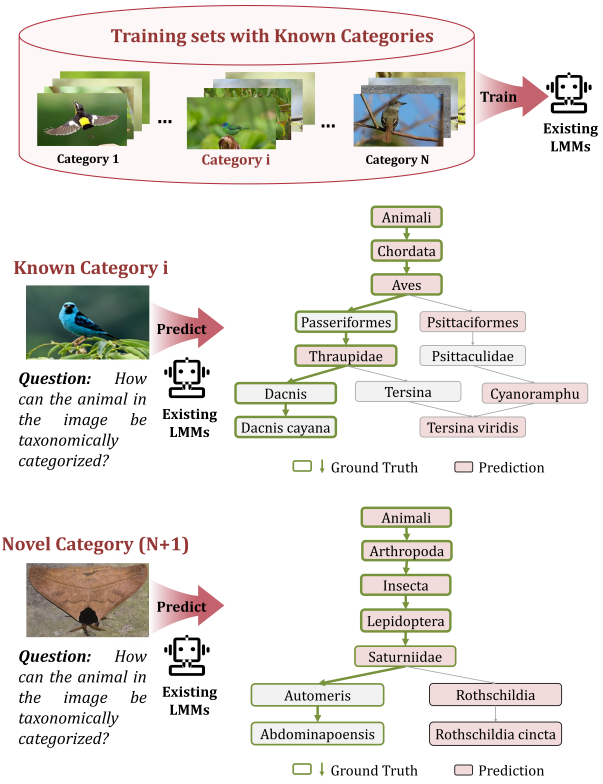


Figure 1. LMMs struggle with hierarchical visual recognition (HVR), failing to obey the hierarchical consistency on both known and novel categories.

1. Introduction

Hierarchical visual recognition (HVR) [3, 6, 20, 34] aims to predict a semantic tree of labels, capturing visual categories from coarse to fine granularity. This structured output enables flexible usage: an expert user may seek a specific label such as *Acadian Flycatcher*, while a general user may only require the broader category *Bird*. Moreover, predicting the entire hierarchy enhances robustness and scalability, as models learn to generalize across different abstraction levels and can be easily extended to incorporate new parent or child categories.

A high-performing, general-purpose visual understanding system should not only recognize fine-grained leaf classes but also robustly map inputs to coarser, higher-level categories within a taxonomy. While existing large multimodal models (LMMs) [2, 15, 27] have demon-

*Corresponding author.

strated strong performance in fine-grained visual recognition (FGVR), they remain weak in hierarchical visual recognition (HVR) and often fail to maintain hierarchical consistency [43]. As illustrated in Figure 1, their predictions can violate the taxonomy path—for example, breaking the sequence Animalia → Chordata → Aves → Passeriformes → Thraupidae → Dacnis → Dacnis cayana. Furthermore, as new categories continually emerge, LMMs must be capable of identifying novel classes that are absent from the training set and potentially lack sufficient public imagery. Because annotating data across taxonomic hierarchies requires substantial domain expertise, constructing large-scale datasets that comprehensively cover all semantic levels is infeasible [48]. As a result, in realistic settings, LMMs often struggle to recognize novel categories.

Taxonomy is a natural and fundamental component of visual understanding. Biological taxonomies, in particular, encompass a vast range of entities that constitute our visual world [46]. Recently, biological foundation models (BFMs) [13, 42, 59] have emerged as large-scale repositories of biological taxonomic knowledge. For example, BioCLIP2 [13] employs hierarchical contrastive learning with taxonomic supervision, leading to embedding spaces in which species representations reflect their ecological and functional relationships. In this way, visual parts [1, 10, 24], attributes [9, 23, 31], and inter-object relationships [22] naturally form hierarchical groupings based on shared characteristics, providing strong priors for inferring unseen or newly discovered categories within a semantic tree.

To absorb this taxonomic knowledge within existing LMMs, we propose **T**axonomy-**A**ware **R**epresentation **A**lignment (**TARA**), a simple yet effective strategy for boosting the performance of HVR. TARA explicitly supervises the intermediate representations of LMMs, encouraging them to learn inter-species ecological alignment and intra-species variation from BFMs. Concretely, we first align the internal visual representations of LMMs with those of BFMs using a cosine-similarity-based alignment loss [51]. Furthermore, recognizing that a single image may correspond to multiple taxonomic levels, we additionally align the first hidden states of the output answers with the BFM-encoded category representation at a specific level. Since BFMs are trained with hierarchical contrastive objectives, they provide rich multimodal embeddings that encode strong taxonomy priors. Aligning LMM representations with those from BFMs thus enables the model to preserve fine-grained visual cues while flexibly mapping to categories of varying granularity according to user intent. Trained in an alternating manner with No-Thinking RL [25], we demonstrate that TARA consistently achieves substantial performance gains across various base models, on both known and novel categories.

Our contributions are summarized as follows:

- We highlight a key limitation of LMMs in performing HVR, particularly for novel categories without training images, hindering general-purpose visual understanding.
- We propose TARA, a simple but effective framework that explicitly aligns intermediate representations of LMMs with visual and text features from pretrained BFMs, thereby injecting taxonomic knowledge and enabling richer, hierarchy-aware visual recognition.
- Through comprehensive experiments on both known and novel categories, we demonstrate consistent and significant improvements over base models, and we conduct detailed ablation studies and analyses to validate the effectiveness of each design choice.

2. Related Work

LMMs for FGVR. Large Multimodal Models (LMMs) demonstrate strong general visual understanding but still fall short in fine-grained visual recognition (FGVR) [11, 15, 35, 58], which demands distinguishing between visually similar subcategories. Several studies fine-tuned LMMs with classification-oriented data either by incorporating it during pretraining as captions or during instruction tuning as QA pairs, showing that explicit object mentions in training data are critical for accurate recognition [11]. Other works suggested that the performance gap between LMMs and vision-language models (VLMs) primarily arises from data-related factors: classification-relevant cues are already encoded in the latent space but require sufficient supervision to be effectively decoded [58]. Integrating adequate classification data allows LMMs to approach or match specialized models and to develop stronger object-centric reasoning. From another perspective, Finedefics [15] attributed this gap to misalignment between visual objects and their category names, employing attribute-based descriptions to bridge the two. Although such methods improved recognition accuracy, they lacked domain-specific objectives and explicit explanatory reasoning. To mitigate this, [38] introduced a visual rejection-sampling framework that iteratively synthesizes interpretable, feature-based explanations to enhance explainability. Moreover, Fine-R1 [16] proposed a two-stage framework to learn the reasoning process with only few-shot samples per category, surpassing various strong CLIP-like models. In contrast, training-free approaches such as Sparse Attention Vectors (SAV) [28] adapted LMMs for FGVR by extracting discriminative features from sparse attention heads, eliminating the need for additional training.

Hierarchical Visual Recognition. Hierarchical visual recognition [21, 39] plays a crucial role in achieving a comprehensive understanding of both the visual world [5, 32, 33, 40, 52, 54] and language concepts [17, 47, 61, 62]. Recent research has revisited this long-standing problem, revealing that CLIP-style models [36] often fail to maintain

consistency across taxonomic levels [12, 49]. [49] evaluated CLIP under multiple levels of semantic granularity and proposed a hierarchy-consistent prompt tuning method. [30] improved CLIP's hierarchical representations by embedding them into a hyperbolic space, while [50] extended this direction with graph-based representation learning. Similarly, [29] leveraged hierarchical information to enhance zero-shot classification performance. Beyond CLIP-style models, [41] proposed evaluating LMMs on open-set predictions [56] using taxonomic similarity rather than exact string matching. [43] first investigated LMMs from the perspective of HVR, and positioned the limitations of hierarchical consistency and leaf node accuracy.

3. Preliminaries

3.1. Hierarchical Visual Recognition with LMMs

In conventional visual recognition tasks, the label space is typically flat: each image $x \in \mathcal{X}$ is assigned a single class label $y \in \mathcal{Y}$, where $\mathcal{Y}$ denotes a predefined set of mutually exclusive categories. However, many real-world visual concepts exhibit inherent hierarchical structures, where labels are naturally organized within a taxonomy $\mathcal{T} = (\mathcal{Y}, \mathcal{E})$ [33, 49, 50, 52], such as a tree or a directed acyclic graph (DAG). Here, $\mathcal{E} \subseteq \mathcal{Y} \times \mathcal{Y}$ represents the set of directed edges encoding parent–child relationships, where $(y_i, y_j) \in \mathcal{E}$ indicates that y_i is the parent of y_j. In hierarchical visual recognition (HVR), the objective extends beyond predicting a single leaf-node label $y \in \mathcal{Y}_{\text{leaf}} \subseteq \mathcal{Y}$; instead, the model must also recover the complete ancestral path $(y_0, y_1, \ldots, y_L)$ in $\mathcal{T}$, where y_0 denotes the root node and L is the hierarchy depth.

LMMs are treated as image classifiers f_θ, and language prompts are leveraged to steer their outputs toward specific taxonomy levels. Concretely, we follow [43] to define a VQA-style task for each image and target taxonomy level, denoted as $(x^i, \mathcal{Y}_j)$, where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, L^i$. To enable evaluation in a closed-set setting, we construct four-choice VQA questions, which alleviate the challenges of open-set generation—namely, the vast output space [57] and the ambiguity of prediction granularity. Generally, they follow this format [43]:

<image> Given the plant in the image, what is its taxonomic classification at the <hierarchy> (e.g., kingdom) level?
A.<similar class> B.<ground truth>
C.<similar class> D.<similar class>
Answer with the option letter only.
(Choices are shuffled in the experiments)

Although four-choice VQA tasks are arguably easier than conventional hierarchical classification, we compensate for this simplicity by designing confusing choices. Specifically, for each taxonomy level, we use SigLIP [55] to compute cosine similarity scores between the image and all incorrect text labels, and select the top three most similar labels as distractors. This ensures that all four options belong to the same taxonomy level and present meaningful semantic confusion.

3.2. No-Thinking Reinforcement Fine-tuning

Recently, rule-based reinforcement fine-tuning (RFT) has achieved remarkable progress in enhancing reasoning capabilities of LLMs (e.g., DeepSeek-R1 [14] and Pangu Embedded [4]), often surpassing traditional supervised fine-tuning (SFT) in performance [14, 19, 44]. RFT leverages verifiable rewards to guide training, encouraging models to engage in an explicit thinking process before producing answers for better solution exploration [14]. This explicit reasoning is widely regarded as a key factor behind RFT's success, and many studies on multi-modal RFT [18, 60] have sought to replicate the length-increasing and "aha moment" phenomena observed in DeepSeek-R1 [14].

However, recent evidence suggests that RFT without explicit thinking can outperform its thinking-based counterpart on classification tasks, indicating that reasoning traces are not always necessary or beneficial—especially for smaller models [25]. Therefore, we train the base model with No-Thinking RFT as default, where instruction prompt and reward functions are as follows:

Instruction prompt. Unlike the Thinking-RFT prompt, which encourages step-by-step reasoning, the No-Thinking-RFT prompt explicitly prohibits the model from engaging in any thought process. It is formulated as: `{Question} Please directly output the answer.`

Reward Function. No-Thinking-RFT removes the format reward and employs only an accuracy reward. The accuracy reward R_{accuracy} assigns a value of 1 if the model's output exactly matches the ground truth and 0 otherwise. This strict equality-based reward discourages unnecessary reasoning and enforces concise answers, significantly shorter than the reasoning-heavy outputs of Thinking-RFT.

4. Method

Biology foundation models (BFMs), trained with hierarchical supervision and contrastive objectives, have demonstrated the ability to learn biologically meaningful embedding spaces [13, 42, 59]. In this section, we introduce our Taxonomy-aware Representation Alignment (TARA) for LMMs. TARA aligns representations at two levels: Taxonomic Visual Representation Alignment and Free-grained Label Representation Alignment, corresponding to the specific targets of alignment. When alternately trained with No-Thinking-RFT, LMMs leverage the guidance from BFMs to enhance HVR performance larger and faster. The overall framework is illustrated in Figure 2.

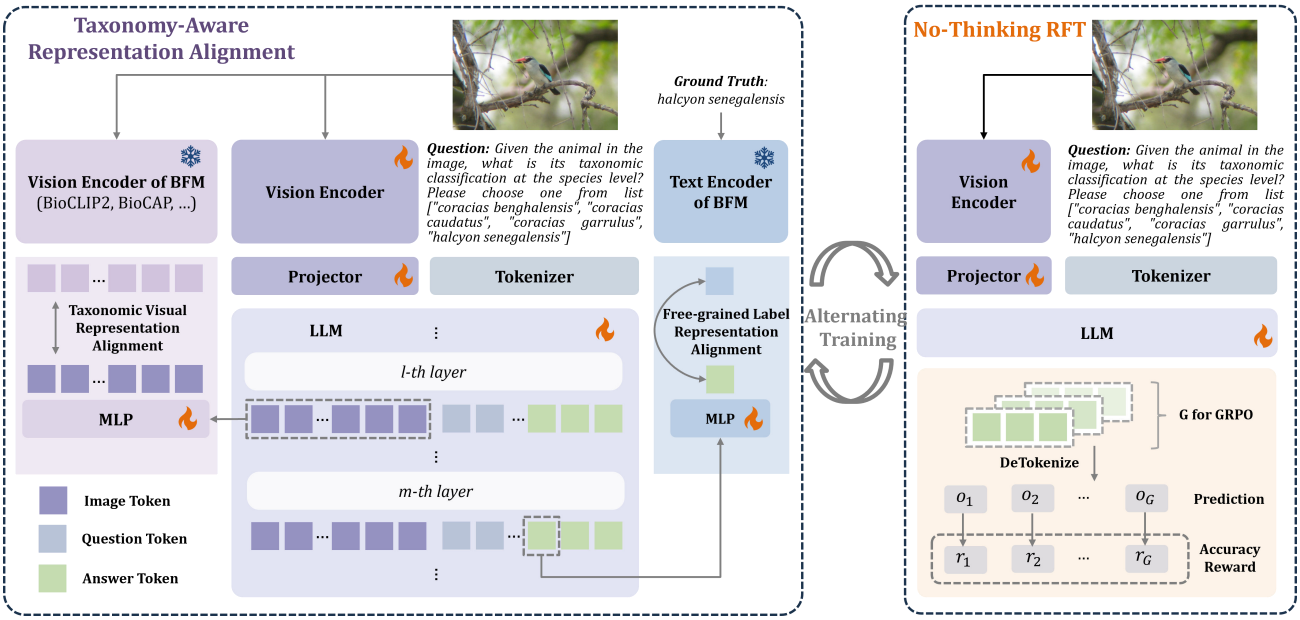


Figure 2. Illustration of the training framework. Taxonomy-Aware Representation Alignment (TARA) is conducted alternately with No-Thinking RFT to improve the hierarchical recognition performance of LMMs with taxonomic knowledge absorbed from BFMs.

4.1. Taxonomic Visual Representation Alignment

Inspired by [53], we employ taxonomy-aware BFMs (e.g., BiocLIP [42], BioCLIP2 [13], and BioCAP [59]) as teacher models to guide the internal visual representations of LMMs. These BFMs provide rich, taxonomically informed supervision that enhances the extraction of discriminative visual cues. Concretely, we align intermediate visual features from the LMM with those produced by a pretrained BFM, thereby transferring domain-specific visual knowledge. Let $\mathcal{E}_{\mathrm{V}}(\cdot)$ denote the pretrained BFM visual encoder. Given an input image I and the question q for identifying the category at a specific level (e.g., *Given the animal in the image, what is its taxonomic classification at the species level? Please choose one from list [similar class, ground truth, similar class, similar class].)*, the encoder outputs target features $\mathbf{y}^{\mathrm{img}} = \mathcal{E}_{\mathrm{V}}(I) \in \mathbb{R}^{N\times d}$, where d is the BFM feature dimension. Let $\mathbf{e}_{\ell}^{\mathrm{img}} \in \mathbb{R}^{N\times D}$ represent the LMM's visual representations at layer ℓ, and let $P_{\mathrm{V}}(\cdot)$ be a learnable projection mapping $\mathbf{e}_{\ell}^{\mathrm{img}}$ into the BFM feature space. The visual alignment loss is defined as

$$\mathcal{L}_{\mathrm{V}} = -\frac{1}{N}\sum_{i=1}^{N} \mathrm{sim}\Big(P_{\mathrm{V}}(\mathbf{e}_{\ell,i}^{\mathrm{img}}),\, \mathbf{y}_{i}^{\mathrm{img}}\Big), \tag{1}$$

where $\mathrm{sim}(\cdot)$ denotes cosine similarity and gradients do not flow into $\mathbf{y}^{\mathrm{img}}$. Minimizing $\mathcal{L}_{\mathrm{V}}$ regularizes the LMM's internal representations, encouraging them to align with the biologically grounded visual space of the BFM.

4.2. Free-grained Label Representation Alignment

Unlike one-hot labels, taxonomic labels naturally encode hierarchical biological structures across multiple levels [13]. A single image may correspond to categories of varying granularity. For example, an expert might aim to identify an instance as an *Acadian Flycatcher* at the species level, whereas a general user may only need the broader label *Bird*. To accommodate this flexibility, we introduce free-grained label representation alignment, which aligns intermediate answer representations with the embeddings of their corresponding labels at the specified granularity level, obtained from pretrained BFMs.

Formally, the BFM encoder $\mathcal{E}_{\mathrm{T}}(\cdot)$ generates a target feature $\mathbf{y}^{\mathrm{label}} = \mathcal{E}_{\mathrm{T}}(C) \in \mathbb{R}^{d}$, where C denotes the ground-truth label at the desired granularity. Let $\mathbf{e}_{m}^{\mathrm{answer}} \in \mathbb{R}^{N'\times D}$ represent the answer embeddings from the LMM at layer m. We then employ a projector $P_{\mathrm{T}}(\cdot)$ to map the first-token embeddings of $\mathbf{e}_{m}^{\mathrm{answer}}$ into the BFM textual feature space. The alignment loss is defined as:

$$\mathcal{L}_{\mathrm{C}} = -\mathrm{sim}\left(P_{\mathrm{T}}(\mathbf{e}_{m,0}^{\mathrm{answer}}), \mathbf{y}^{\mathrm{label}}\right), \tag{2}$$

where $\mathbf{e}_{m,0}^{\mathrm{answer}}$ denotes the embedding of the first token in the generated answer. Minimizing $\mathcal{L}_{\mathrm{C}}$ encourages the answer representations to align with the ground-truth labels at the appropriate semantic level, thereby enabling the model to learn structured, hierarchy-aware representations beneficial for downstream HVR tasks. The overall alignment objective of TARA is given by:

$$\mathcal{L}_{\mathrm{alignment}} = (\mathcal{L}_{\mathrm{V}} + \mathcal{L}_{\mathrm{C}})/2. \tag{3}$$

Table 1. Effects of TARA on iNat-Plant and iNat-Animal datasets.

Base Model	RL	TARA	iNat21-Plant					iNat21-Animal				
			HCA	Acc_{leaf}	POR	S-POR	TOR	HCA	Acc_{leaf}	POR	S-POR	TOR
Qwen3-VL-2B	✗	✗	6.46	30.16	60.15	44.74	41.36	7.18	27.86	65.23	55.57	51.77
	✓	✗	9.23	31.96	64.58	50.81	47.86	8.57	29.32	66.98	57.84	54.10
	✓	✓	**12.78**	**32.66**	**68.57**	**55.98**	**53.72**	**10.26**	**30.77**	**68.42**	**58.85**	**55.86**
			+3.55	+0.70	+3.99	+5.17	+5.86	+1.69	+1.45	+1.44	+1.01	+1.76
Qwen2.5-VL-3B	✗	✗	10.89	39.73	65.77	48.90	48.24	16.70	40.26	73.03	64.05	60.80
	✓	✗	17.91	44.35	72.15	57.29	57.08	21.99	46.25	76.15	67.66	64.83
	✓	✓	**19.53**	**45.66**	**73.87**	**58.92**	**59.38**	**24.02**	**49.16**	**77.26**	**68.62**	**66.11**
			+1.62	+1.31	+1.72	+1.63	+2.30	+2.03	+2.91	+1.11	+0.96	+1.28

Algorithm 1 No-Thinking RFT + TARA

Require: input image I, question q, label C; LMM π_θ; BFM encoders $\mathcal{E}_{\text{V}}, \mathcal{E}_{\text{T}}$; projectors $P_{\text{V}}, P_{\text{T}}$
Ensure: Updated LMM π_θ
1: **for** each step **do**
2: Generate G answers; extract e_ℓ^{img}, $e_{m,0}^{\text{answer}}$.
3: Compute $\mathcal{L}_{\text{RFT}}$.
4: Encode GT: $\mathbf{y}^{\text{img}} = \mathcal{E}_{\text{V}}(I)$, $\mathbf{y}^{\text{label}} = \mathcal{E}_{\text{T}}(C)$.
5: Project: $\hat{\mathbf{y}}^{\text{img}} = P_{\text{V}}(e_\ell^{\text{img}})$, $\hat{\mathbf{y}}^{\text{label}} = P_{\text{T}}(e_{m,0}^{\text{answer}})$.
6: Compute $\mathcal{L}_{\text{alignment}}$.
7: Update $\pi_\theta, P_{\text{V}}, P_{\text{T}}$ with $\mathcal{L}_{\text{align}}$.
8: Update π_θ with $\mathcal{L}_{\text{RFT}}$ (with $P_{\text{V}}, P_{\text{T}}$ frozen).
9: **end for**

4.3. Training and Inference

Algorithm 1 shows the training process. In addition to applying No-Thinking RFT to optimize LMMs, we integrate TARA to jointly update the LMMs and two lightweight MLP-style projectors, $P_{\text{V}}(\cdot)$ and $P_{\text{T}}(\cdot)$, which map intermediate representations of the LMM. This alternating optimization scheme enables the model to effectively absorb taxonomic knowledge, leading to improved performance and generalization in HVR. During inference, both the BFMs and projectors are discarded, and LMMs are directly prompted to perform HVR.

5. Experiments

5.1. Experimental Setup

Datasets. We employ the iNaturalist-2021 (iNat21) dataset [46], a large-scale collection featuring species-level annotations across diverse biological taxa. We separate it into two taxonomies, Plant and Animal, comprising 4,271 and 5,388 leaf nodes, respectively, across six hierarchical levels. To assess the model's ability to recognize unseen species, we further incorporate the TerraIncognita dataset [8]. This dataset includes a mixture of expertly annotated images of insect species that are likely familiar to frontier AI models, as well as images of rare or poorly documented species with few or no publicly available samples. These novel-category images were collected during field expeditions in biodiversity hotspots across Central and South America. For many of these samples, only higher-level taxonomic labels are reliable, and numerous taxa are believed to be entirely new to science.

Implementation Details. We adopt the Qwen family of models as base models due to their strong zero-shot performance. Specifically, we employ Qwen3-VL-2B-Instruct [45] and Qwen2.5-VL-3B-Instruct [2] as our base models, and fine-tune all parameters following the training configurations of [7, 60]. Our No-Thinking RFT dataset is constructed from the iNat21-Animal and iNat21-Plant training splits, containing 1-shot VQA samples across 9,659 leaf categories. For hierarchical recognition, we formulate questions corresponding to the *order*, *family*, *genus*, and *species* levels, sampled at a ratio of 1:2:4:8. Evaluation is conducted using 1-shot VQA samples derived from the corresponding iNat-Animal and iNat-Plant validation sets. The vision and text projectors, $P_{\text{V}}(\cdot)$ and $P_{\text{T}}(\cdot)$, are implemented as lightweight three-layer MLPs with SiLU activations, while $\mathcal{E}_{\text{V}}(\cdot)$ and $\mathcal{E}_{\text{T}}(\cdot)$ are the vision and text encoders of BioCLIP2 [13], respectively. All experiments are performed on 8×A6000 GPUs with a batch size of 1 per GPU and a two-step gradient accumulation [7, 37]. Each model is trained for one epoch. We adopt the few-shot classification hyper-parameters from [25]. All input images are resized to 328×328, and no data augmentation is applied.

5.2. Evaluation Metrics

To comprehensively evaluate model performance, we focus on the hierarchical consistency of predictions [32, 49], complemented by the leaf-level classification accuracy [15, 26, 57], which serves as the upper bound of hierarchical consistency. The evaluation metrics are detailed below.

Table 2. Effects of TARA on TerraIncognita dataset [8].

Species	RL	TARA	Order F1	Family F1
Known	✗	✗	17.16	10.83
	✓	✗	23.30	11.47
	✓	✓	41.56	25.47
			+18.26	+14.00
Novel	✗	✗	17.16	10.83
	✓	✗	23.30	11.47
	✓	✓	33.45	12.67
			+10.15	+1.20

Hierarchical Consistent Accuracy (HCA) [32, 49]. This metric is defined as

$$\mathrm{HCA} = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{L^i}\mathbb{1}\left[f_\theta\left(x^i;\mathcal{Y}_j\right)=y_j^i\right], \tag{4}$$

Here, N is the number of test samples, L^i is the depth of the hierarchy for the i-th input x^i, and $\mathcal{Y}j$ denotes the label set at level j. HCA computes the proportion of samples whose predicted paths exactly match the ground truth from root to leaf. It is therefore a stricter criterion than flat accuracy and serves as our primary evaluation metric for hierarchical classification.

Leaf-Level Accuracy ($\mathrm{Acc}_{\mathrm{leaf}}$) [15, 26, 57]. It reflects the model's discriminative ability at the most fine-grained level:

$$\mathrm{Acc}_{\mathrm{leaf}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left[f_\theta\left(x^i;\mathcal{Y}_L\right)=y_L^i\right]. \tag{5}$$

Since correctly predicting a leaf node implies correctness at that level, $\mathrm{Acc}_{\mathrm{leaf}}$ naturally upper-bounds HCA. However, a sample contributes to HCA only if all nodes along its path $(y_0, y_1, \ldots, y_L)$ are predicted correctly, making HCA a more stringent measure of hierarchical consistency.

Point-Overlap Ratio (POR) [52]. It measures hierarchical performance beyond strict correctness, defined as:

$$\mathrm{POR} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{L_i}\mathbb{1}\left[f_\theta\left(x_i;\mathcal{Y}_j\right)=y_j^i\right]}{L_i}. \tag{6}$$

Unlike HCA, which requires an exact match along the entire path, POR allows partial correctness by averaging the proportion of correctly predicted nodes. This provides a fine-grained assessment of how well model outputs align with the target hierarchy.

Strict Point-Overlap Ratio (S-POR). S-POR refines POR by rewarding only contiguous segments of correct predictions. For the i-th sample, we locate the longest run of consecutive correctly predicted nodes and normalize by the hierarchy depth L_i:

$$\text{S-POR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_i}\max_{1\le a\le b\le L_i}\Big[(b-a+1) \times\prod_{j=a}^{b}\mathbb{1}\big[f_\theta(x_i;\mathcal{Y}_j)=y_j^i\big]\Big]. \tag{7}$$

This stricter definition penalizes isolated correct predictions and emphasizes full-path consistency within the hierarchy.

Top Overlap Ratio (TOR). Following [49], TOR evaluates local hierarchical consistency by considering adjacent layer pairs as independent evaluation units:

$$\mathrm{TOR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_i-1}\sum_{j=1}^{L_i-1}\mathbb{1}\big[f_\theta(x_i;\mathcal{Y}_j)=y_j^i\big] \times\mathbb{1}\big[f_\theta(x_i;\mathcal{Y}_{j+1})=y_{j+1}^i\big]. \tag{8}$$

A TOR value of 1 indicates perfect pairwise consistency between consecutive layers, while lower scores reveal local violations of the hierarchical structure.

F1 Score. For evaluations on TerraIncognita, we follow the setup of [8] and use a fixed system prompt instructing the model to classify an insect specimen across the taxonomic hierarchy, returning "Unknown" when it is uncertain. We report F1 score at the Order and Family level since full taxonomic labels are missing for novel species. It captures the harmonic mean of precision and recall and thus provides a balanced assessment of the model's classification performance.

Table 3. Ablation on target alignment layers of TARA.

Target Layer		iNat21-Plant				
$\mathcal{L}_{\mathrm{V}}$	$\mathcal{L}_{\mathrm{T}}$	HCA	$\mathrm{Acc}_{\mathrm{leaf}}$	POR	S-POR	TOR
/	/	6.46	30.16	60.15	44.74	41.36
14	/	10.72	33.27	65.35	51.40	49.11
14	14	10.40	**33.43**	65.27	51.14	49.26
14	28	**12.78**	32.66	**68.57**	**55.98**	**53.72**
28	14	11.87	33.20	67.17	54.62	51.88
28	28	10.65	32.73	66.52	53.41	50.48

5.3. Main Results

Known Categories. The results on iNat-Animal and iNat-Plant are summarized in Table 2. With only 1-shot supervision, both base models (Qwen3-VL-2B and Qwen2.5-VL-3B) trained with TARA consistently outperform their baselines, achieving improvements in both hierarchical consistency and leaf-level accuracy. This is obtained through a simple yet effective strategy that aligns intermediate LMM features with BFM targets, enabling the model to absorb

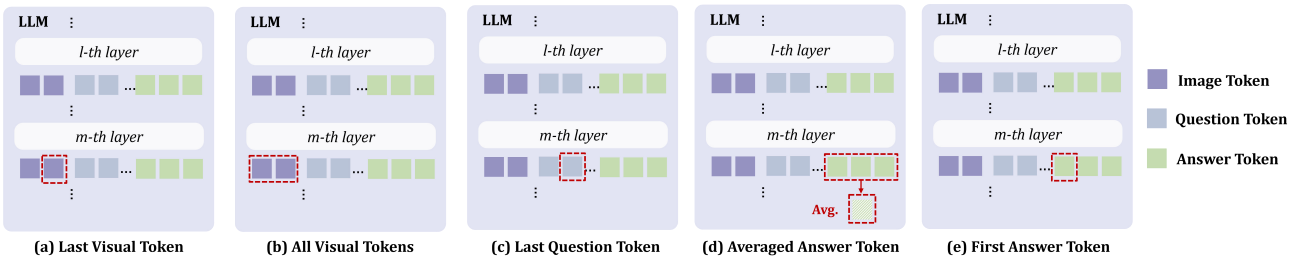


Figure 3. Different designs of target alignment features. (a) and (b) are for $\mathcal{L}_{\mathrm{V}}$, and (c)-(e) are for $\mathcal{L}_{\mathrm{C}}$.

Table 4. Ablation on target alignment features of TARA.

Features		iNat21-Plant				
$\mathcal{L}_{\mathrm{V}}$	$\mathcal{L}_{\mathrm{T}}$	HCA	$\mathrm{Acc}_{\mathrm{leaf}}$	POR	S-POR	TOR
(a)	(e)	10.30	32.22	65.50	52.36	49.59
(b)	(c)	12.01	**33.62**	67.39	55.12	52.25
(b)	(d)	11.59	33.41	67.18	53.50	51.53
(b)	(e)	**12.78**	32.66	**68.57**	**55.98**	**53.72**

Table 5. Left: visual probing results on the last and averaged image hidden states. Right: evaluation results on ImageWikiQA [58].

RL	TARA	Last	Avg.
✗	✗	13.30	85.00
✓	✗	14.40	84.40
✓	✓	18.30	85.90
		+3.90	+1.50

RL	TARA	Acc.
✗	✗	46.60
✓	✗	48.70
✓	✓	51.40
		+2.70

taxonomic structure. On the known split of the TerraIncognita dataset [8], TARA also delivers substantial gains on Order F1 and Family F1. These consistent improvements demonstrate that TARA effectively guides LMMs to recognize known categories more reliably.

Novel Categories. To assess whether these gains stem merely from memorizing known training categories, we further evaluate on the novel split of the TerraIncognita dataset [8], which contains newly captured images of rare or previously unseen species beyond the scope of training data. Even in this challenging setting, TARA provides consistent improvements, indicating that the learned representations generalize beyond observed classes within the tree of life. Overall, these results highlight a broader insight: regularizing intermediate representations is a broadly applicable approach for strengthening LMMs with taxonomic knowledge, even under extremely limited data.

5.4. Component-wise Analysis

In this ablation study, we systematically analyze the key design choices of our framework, examining the contribution of each core component as well as the impact of different target alignment features and the selected layers used in the alignment losses.

Effects of core components. We first analyze the contribution of each alignment loss introduced in TARA, namely $\mathcal{L}_{\mathrm{V}}$ and $\mathcal{L}_{\mathrm{C}}$. As shown in Table 3, incorporating $\mathcal{L}_{\mathrm{V}}$ enhances both hierarchical consistency and leaf-level accuracy, indicating that the alignment signal from the BFM vision encoder carries taxonomy-aware visual knowledge. Introducing $\mathcal{L}_{\mathrm{C}}$ further improves the intermediate LMM representations, helping the model better map an input image to labels across different levels of granularity. This leads to consistent improvements in hierarchical consistency (e.g., a 2.06% gain in HCA.)

Target alignment layers. We then analyze alignment at individual target layers to determine the most effective positions of $\mathcal{L}_{\mathrm{V}}$, $\mathcal{L}_{\mathrm{C}}$. As shown in the target-layers ablation results in Table 3, we report performance at different layers throughout the network. We observe that performance varies depending on the alignment layer, with the 14th, 28th layer of the 28-layer model consistently yielding stronger results. This trend is consistent with the intuition that the visual information is progressively transferred to taxonomic label and finally to the label at a certain level according to the given question.

Target alignment features. For the visual alignment loss $\mathcal{L}_{\mathrm{V}}$, we compare two design choices for constructing target alignment features: using only the embedding of the last visual token (Figure 3 (a)) versus using the embeddings of all visual tokens (Figure 3 (b)). For the text alignment loss $\mathcal{L}_{\mathrm{C}}$, we evaluate three alternatives: using the embedding of the last question token (Figure 3 (c)), the averaged embeddings of all answer tokens (Figure 3 (d)), and our proposed approach, which uses the embedding of the first answer token (Figure 3 (e)). As summarized in Table 4, the "All Visual Tokens" variant yields the best performance for $\mathcal{L}_{\mathrm{V}}$, while the "First Answer Token" strategy is the most effective choice for $\mathcal{L}_{\mathrm{C}}$.

5.5. Probing Analysis

To examine whether TARA enhances visual representations by enabling the extraction of more discriminative cues for HVR, we conduct linear probing experiments on the image token embeddings from the residual stream at the last layer of the LLM. Two pooling strategies are explored: mean pooling across all image tokens and selecting the final image token. A linear classifier is trained on the iNat21-Plant training set with a batch size of 512, a learning rate of 1e-4, and the Adam optimizer for 500 epochs. To ensure balance, we randomly sample 10 images per class from 100 categories (1,000 images total) for training, and use another 1,000 images for testing. Table 5 (left) reports species-level classification accuracy. We observe that TARA outperforms No-Thinking RFT variant, demonstrating its superior ability to extract fine-grained visual cues.

5.6. Classification-based VQA Benchmark

Classification serves as a fundamental building block for developing more advanced visual capabilities. For instance, correctly recognizing an object is often a prerequisite for answering complex questions about it. To demonstrate that TARA extends beyond HVR and benefits more challenging tasks, we further evaluate its performance on ImageWikiQA [58], a dataset containing complex, real-world questions grounded in ImageNet objects. As shown in Table 5 (right), TARA improves accuracy from 48.70% to 51.40%, highlighting that strengthening HVR indeed enhances the advanced reasoning capabilities of LMMs.

5.7. Qualitative Results

We further validate the effectiveness of TARA through qualitative analyses of model outputs. As shown in Figure 4, TARA not only improves fine-grained prediction accuracy but also rectifies errors along the entire hierarchical label path, yielding substantially better hierarchical consistency than merely applying No-Thinking RFT to Qwen3-VL-2B.

5.8. Training Efficiency

To further showcase the additional benefits of TARA, we evaluate iNat21-Plant at every 200 training step from the total of 604 training steps of the No-Thinking RFT stage in Figure 5. Qwen3-VL-2B trained with TARA show that convergence is faster and quickly surpasses the baseline performance in early steps, demonstrating the effectiveness of representation guidance for taxonomy knowledge injection. Since TARA adds minimal overhead to the overall training process, these earlier performance gains are likely to translate into enhanced scalability.

6. Conclusion

In this work, we introduce TARA, a simple yet effective strategy that aligns the internal representations of LMMs with those of pre-trained BFMs. By doing so, our approach enables the extraction of fine-grained visual semantics, facilitating accurate mapping to taxonomic labels and allowing the transfer of visual features across categories at any level of the tree of life. Extensive experiments demonstrate that TARA improves recognition of known categories and generalizes effectively to novel categories in the HVR task.

Limitations. More real-world classification tasks beyond the biological domain involve hierarchical label spaces (e.g., taxonomies or knowledge graphs). Incorporating such structure could further advance LMMs toward general-purpose visual understanding.

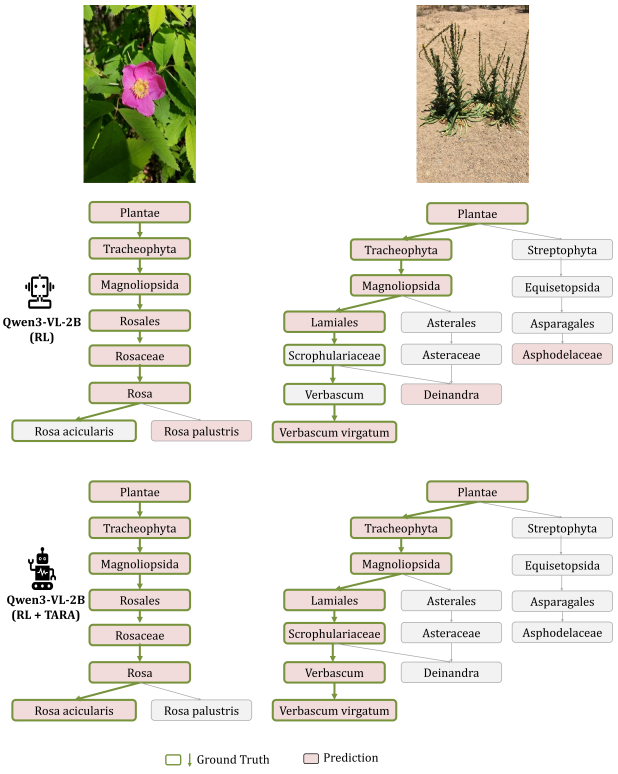


Figure 4. Qualitative comparison of No-Thinking RFT with and without TARA. The two columns show that TARA can achieve better leaf node accuracy and hierarchical consistency.

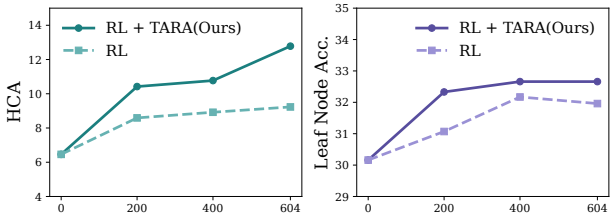


Figure 5. Training efficiency. Models trained with TARA achieve faster convergence.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (62525201, 62132001, 62432001) and Beijing Natural Science Foundation (L247006). This work was partially supported by PKU Kunpeng&Ascend Center of Excellence.

References

[1] Pablo Arbeláez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev, and Jitendra Malik. Semantic segmentation using regions and parts. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3378–3385. IEEE, 2012. 2

[2] 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. *arXiv preprint arXiv:2502.13923*, 2025. 1, 5

[3] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": fine-grained, or not. In *CVPR*, 2021. 1

[4] Hanting Chen, Yasheng Wang, Kai Han, Dong Li, Lin Li, Zhenni Bi, Jinpeng Li, Haoyu Wang, Fei Mi, Mingjian Zhu, et al. Pangu embedded: An efficient dual-system llm reasoner with metacognition. *arXiv preprint arXiv:2505.22375*, 2025. 3

[5] Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4858–4867, 2022. 2

[6] Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *CVPR*, 2022. 1

[7] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02. 5

[8] Shivani Chiranjeevi, Hossein Zaremehrjerdi, Zi K Deng, Talukder Z Jubery, Ari Grele, Arti Singh, Asheesh K Singh, Soumik Sarkar, Nirav Merchant, Harold F Greeney, et al. Terraincognita: A dynamic benchmark for species discovery using frontier models. *arXiv preprint arXiv:2506.03182*, 2025. 5, 6, 7

[9] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009. 2

[10] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[11] Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. *arXiv preprint arXiv:2406.14496*, 2024. 2

[12] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*, 2023. 3

[13] Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E White, James Balhoff, et al. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. *arXiv preprint arXiv:2505.23883*, 2025. 2, 3, 4, 5

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[15] Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. *arXiv preprint arXiv:2501.15140*, 2025. 1, 2, 5, 6

[16] Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-r1: Make multi-modal llms excel in fine-grained visual recognition by chain-of-thought reasoning. *arXiv preprint arXiv:2602.07605*, 2026. 2

[17] Yuan He, Moy Yuan, Jiaoyan Chen, and Ian Horrocks. Language models as hierarchy encoders. *Advances in Neural Information Processing Systems*, 37:14690–14711, 2024. 2

[18] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 3

[19] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 3

[20] Juan Jiang, Jingmin Yang, Wenjie Zhang, and Hongbin Zhang. Hierarchical multi-granularity classification based on bidirectional knowledge transfer. *Multimedia Systems*, 2024. 1

[21] 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*, 29:820–865, 2015. 2

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2

[23] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on com-*

puter vision and pattern recognition, pages 951–958. IEEE, 2009. 2

[24] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401 (6755):788–791, 1999. 2

[25] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*, 2025. 2, 3, 5

[26] Huan Liu, Lingyu Xiao, Jiangjiang Liu, Xiaofan Li, Ze Feng, Sen Yang, and Jingdong Wang. Revisiting mllms: An in-depth analysis of image classification abilities. *arXiv preprint arXiv:2412.16418*, 2024. 5, 6

[27] Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*, 2025. 1

[28] Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers. *arXiv preprint arXiv:2412.00142*, 2024. 2

[29] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 3

[30] 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. *arXiv preprint arXiv:2410.06912*, 2024. 3

[31] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, 22, 2009. 2

[32] Seulki Park, Youren Zhang, Stella X Yu, Sara Beery, and Jonathan Huang. Learning hierarchical semantic classification by grounding on consistent image segmentations. *arXiv preprint arXiv:2406.11608*, 2024. 2, 5, 6

[33] Seulki Park, Youren Zhang, X Yu Stella, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3

[34] Seulki Park, Youren Zhang, Stella X. Yu, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[35] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. *Authorea Preprints*, 2025. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[37] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 5

[38] Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. *arXiv preprint arXiv:2502.14044*, 2025. 2

[39] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 2011. 2

[40] Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. *Advances in Neural Information Processing Systems*, 37:91220–91259, 2024. 2

[41] Vésteinn Snæbjarnarson, Kevin Du, Niklas Stoehr, Serge Belongie, Ryan Cotterell, Nico Lang, and Stella Frank. Taxonomy-aware evaluation of vision-language models. *arXiv preprint arXiv:2504.05457*, 2025. 3

[42] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024. 2, 3, 4

[43] Yuwen Tan, Yuan Qing, and Boqing Gong. Vision llms are bad at hierarchical visual understanding, and llms are the bottleneck. *arXiv preprint arXiv:2505.24840*, 2025. 2, 3

[44] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. 3

[45] Qwen Team. Qwen3 technical report, 2025. 5

[46] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Computer Vision and Pattern Recognition*, 2021. 2, 5

[47] Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. *arXiv preprint arXiv:2203.03825*, 2022. 2

[48] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021. 2

[49] Tz-Ying Wu, Chih-Hui Ho, and Nuno Vasconcelos. Protect: Prompt tuning for taxonomic open set classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16531–16540, 2024. 3, 5, 6

[50] 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. *arXiv preprint arXiv:2311.14064*, 2023. 3

[51] Zhiwen Yang and Yuxin Peng. Gala-2.5 d: Global-local alignment with 2.5 d semantic guidance for camera-based 3d semantic scene completion in autonomous driving. *Chinese Journal of Electronics*, 2025. 2

[52] 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*, pages 116–132. Springer, 2022. 2, 3, 6

[53] Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025. 4

[54] Siqi Zeng, Sixian Du, Makoto Yamada, and Han Zhao. Learning structured representations by embedding class hierarchy with fast optimal transport. *arXiv preprint arXiv:2410.03052*, 2024. 2

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3

[56] Ruru Zhang, E Haihong, Lifei Yuan, Yanhui Wang, Lifei Wang, and Meina Song. Fgm-spcl: Open-set recognition network for medical images based on fine-grained data mixture and spatial position constraint loss. *Chinese Journal of Electronics*, 33(4):1023–1033, 2024. 3

[57] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3, 5, 6

[58] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*, 2024. 2, 7, 8

[59] Ziheng Zhang, Xinyue Ma, Arpita Chowdhury, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Samuel Stevens, Hilmar Lapp, Tanya Berger-Wolf, Yu Su, et al. Biocap: Exploiting synthetic captions beyond labels in biological foundation models. *arXiv preprint arXiv:2510.20095*, 2025. 2, 3, 4

[60] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025. 3, 5

[61] Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1106–1117, 2020. 2

[62] Juncheng Zhou, Lijuan Zhang, Yachen He, Rongli Fan, Lei Zhang, and Jian Wan. A novel negative sample generation method for contrastive learning in hierarchical text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5645–5655, 2025. 2