BEYOND TEXTUAL SHORTCUTS: A TRAINING FRAMEWORK FOR HIERARCHICAL VISUAL RECOGNITION WITH VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical visual recognition requires vision-language models (VLMs) to make accurate and taxonomically consistent predictions across multiple semantic levels. However, existing hierarchical visual instruction tuning methods expose preceding ground-truth answers during training, potentially encouraging textual shortcuts rather than visually grounded recognition. Through a reference-taxonomy analysis and controlled interventions, we reveal that historical textual context can determine many training answers without visual evidence and that trained models exhibit substantial reliance on preceding taxonomic cues. To address this issue, we propose a decoupled hierarchical training framework that separates hierarchical supervision from historical textual dependencies. Specifically, it combines Exchange-Isolated Causal Attention (EICA), which isolates preceding question–answer exchanges while preserving shared visual information, with a probabilistic hierarchy loss that explicitly encourages taxonomic consistency across prediction levels. Experiments on iNat21-Plant and iNat21-Animal with two generative VLM backbones demonstrate consistent improvements in hierarchical consistency and fine-grained recognition across seen, unseen, and cross-domain settings.

1 INTRODUCTION

Hierarchical visual recognition aims to identify visual concepts at multiple levels of granularity, from coarse categories to fine-grained classes, and plays an important role in understanding the structured relationships among visual concepts (Silla Jr & Freitas, 2011; Kosmopoulos et al., 2015; Wu et al., 2024). Unlike conventional flat classification, hierarchical recognition requires models to distinguish categories at different semantic levels while maintaining consistency across a predefined taxonomy (Novack et al., 2023; Wang et al., 2023). This task is particularly challenging in fine-grained domains, where closely related categories often exhibit subtle visual differences (Van Horn et al., 2021; Yu et al., 2026). Effectively capturing hierarchical relationships while maintaining accurate visual recognition across different levels therefore remains an important research challenge.

Recent advances in vision–language models (VLMs) have opened new possibilities for hierarchical visual recognition by leveraging their visual understanding and language generation capabilities (Radford et al., 2021; Liu et al., 2023). In particular, recent studies have explored incorporating taxonomic knowledge into generative VLMs through hierarchical supervision, where models are trained to predict categories across multiple taxonomic levels (He et al., 2026b;a; Tan et al., 2026). A common approach organizes hierarchical predictions into sequential question–answer exchanges (Tan et al., 2026) or coarse-to-fine textual sequences (Li et al., 2026), allowing fine-grained predictions to condition on preceding taxonomic information. While such formulations provide a convenient way to incorporate hierarchical structure into visual instruction tuning, they also raise an important question: *to what extent do models learn to recognize visual evidence at each taxonomic level, rather than relying on previously provided hierarchical textual context?*

To answer this question, we investigate how historical textual context influences hierarchical predictions during visual instruction tuning. As shown in Fig. 1(a), preceding ground-truth answers are accessible to subsequent predictions during training. Such a teacher-forcing scheme may encourage models to infer fine-grained categories from taxonomic cue rather than visual evidence. At inference

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.

time, however, hierarchy levels are queried independently without access to historical context. As illustrated in Fig. 1(b), this can lead to hierarchically inconsistent predictions, where the model may correctly identify a fine-grained category but misclassify its ancestor. Through a systematic analysis in Sec. 3.2, we find that historical textual context can uniquely determine the correct answer for a substantial proportion of training examples, and that trained models are more sensitive to perturbations of preceding taxonomic information than to the tested image replacements. These findings reveal a potential textual shortcut in existing hierarchical visual instruction tuning methods and motivate the need to disentangle hierarchical supervision from historical textual dependencies.

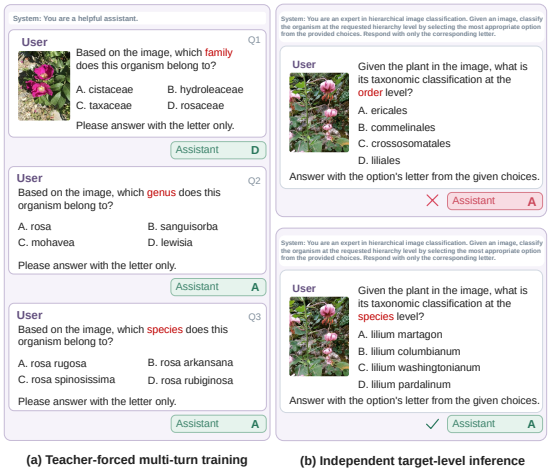


Figure 1: Training (a) and inference (b) in sequential hierarchical visual instruction tuning.

Motivated by these findings, we propose a training framework that decouples hierarchical supervision from historical textual dependencies while preserving taxonomic consistency. Our framework consists of two complementary components. First, we introduce Exchange-Isolated Causal Attention (EICA), which prevents information flow across hierarchical question–answer exchanges while preserving access to shared visual information. By isolating historical textual context, EICA encourages visually grounded predictions at each taxonomic level without introducing additional trainable parameters. Second, to preserve hierarchical relationships without relying on preceding textual answers, we introduce a probabilistic hierarchy loss that explicitly encourages taxonomic consistency across adjacent levels through their predicted probability distributions. Together, these components separate hierarchical supervision from textual history, enabling models to learn taxonomically consistent predictions grounded in visual evidence.

We evaluate our framework on the iNat21-Plant and iNat21-Animal hierarchical recognition benchmarks (Van Horn et al., 2021; Tan et al., 2026) using Qwen2.5-VL-3B-Instruct (Bai et al., 2025b) and Qwen3-VL-8B-Instruct (Bai et al., 2025a). Extensive experiments demonstrate consistent improvements in hierarchical consistency and fine-grained recognition across seen categories, unseen categories, and a previously unseen taxonomic domain. Furthermore, experiments under a one-shot training regime demonstrate the effectiveness of our framework with limited training examples.

Our main contributions are summarized as follows: 1) We identify and systematically investigate hierarchical textual shortcuts in visual instruction tuning. Through a reference-taxonomy analysis and controlled interventions, we reveal that historical textual context can often determine target categories without visual evidence and that trained models exhibit substantial reliance on preceding taxonomic cues. 2) We propose a decoupled hierarchical training framework that combines Exchange-Isolated Causal Attention (EICA) with a probabilistic hierarchy loss. By isolating historical textual dependencies while explicitly preserving taxonomic relationships, our framework encourages visually grounded and hierarchically consistent predictions. 3) We conduct extensive experiments across two VLM backbones and three evaluation splits, demonstrating consistent improvements in hierarchical consistency and fine-grained recognition over existing methods.

2 Related Work

Hierarchical Visual Recognition. Hierarchical visual recognition (Silla Jr & Freitas, 2011; Kosmopoulos et al., 2015; Wu et al., 2024) is important for understanding the visual world (Yi et al., 2022; Chen et al., 2022) and language concepts (Zhou et al., 2020; He et al., 2024; Zhou et al., 2025). Early deep-learning approaches developed hierarchical classifiers with convolutional networks (Yan et al., 2015; Chang et al., 2021; Chen et al., 2022; Wang et al., 2026), while subsequent studies explored vision transformer-based architectures (Wang et al., 2023; Park et al., 2025). With the emergence of vision–language models such as CLIP (Radford et al., 2021), hierarchical visual

recognition has increasingly incorporated language supervision through image–text alignment and prompt-based prediction (Geng et al., 2023; Novack et al., 2023; Wu et al., 2024). More recently, hierarchical visual recognition has been explored with generative vision–language models (He et al., 2026b;a). Tan et al. (2026) systematically evaluated their hierarchical recognition capabilities, revealing substantial limitations in hierarchical consistency and leaf-level accuracy.

Hierarchical Supervision with Vision-Language Models. Hierarchical supervision has been increasingly incorporated into vision–language models to improve hierarchical visual recognition (Stevens et al., 2024; Zheng et al., 2025; Gu et al., 2026). For CLIP-style models, prior work has explored hyperbolic representations (Pal et al., 2025), graph-based representation learning (Xia et al., 2025), and hierarchical label sets for zero-shot classification (Novack et al., 2023). For generative VLMs, TARA (He et al., 2026b) and HiR^2 (He et al., 2026a) introduce taxonomic structure through representation alignment and geometric regularization, respectively. Another line of work formulates hierarchical supervision as sequential textual prediction, organizing hierarchical labels into successive prediction steps. Tan et al. (2026) employ multi-turn question–answer exchanges across taxonomic levels, while VL-Taxon (Li et al., 2026) generates coarse-to-fine taxonomic sequences within a single response during supervised fine-tuning. Unlike prior work focused on incorporating hierarchical knowledge, we investigate textual shortcuts in sequential supervision and propose a framework that decouples historical textual dependencies from hierarchical consistency constraints.

3 Diagnosing Hierarchical Textual Shortcuts

In this section, we investigate potential textual shortcuts in VLM-based hierarchical recognition, where visual instruction tuning may allow models to exploit historical context to infer taxonomic labels without sufficient visual evidence. To this end, we first formalize the hierarchical recognition setting and then examine the presence of such shortcuts and their influence on model predictions.

3.1 Preliminaries

Hierarchical Visual Recognition. This task identifies images at multiple taxonomic levels, from coarse categories to fine-grained species. We represent the taxonomy as a rooted tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the category set and each edge $(y_i, y_j) \in \mathcal{E}$ links a parent y_i to its child y_j. For an image x, its ground-truth labels form a hierarchical path $\pi(x) = (y_1, \dots, y_L)$ across L taxonomic levels, with $(y_\ell, y_{\ell+1}) \in \mathcal{E}$. Following the four-choice VQA protocol of Tan et al. (2026), we formulate each level as a multiple-choice question with four candidate categories, including the ground-truth label. Given the image and the level-specific question, the model predicts the target category by generating the corresponding option letter. At inference, each level is queried independently, without ancestor labels or previous predictions. A level-specific query is shown in Fig. 1(b).

Hierarchical visual instruction tuning. Tan et al. (2026) formulate hierarchical visual recognition as multi-turn instruction tuning following LLaVA (Liu et al., 2023). For brevity, we refer to this approach as *HierSFT*. Given an image x with a ground-truth path $\pi(x) = (y_1, \dots, y_L)$, the training sequence is $S = [v(x), q_1, a_1, \dots, q_L, a_L]$, where $v(x)$ denotes the shared system and visual prefix, and q_ℓ and a_ℓ denote the level-specific question and ground-truth answer at level ℓ, respectively.

During training, the complete sequence is processed in parallel, with preceding questions and ground-truth answers accessible to subsequent turns through standard causal attention:

$$\mathcal{A}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V, \tag{1}$$

where Q, K, and V are the query, key, and value matrices, respectively. The causal mask M allows each query position i to attend only to key positions $j \le i$, with $M_{ij} = 0$ for $j \le i$ and $M_{ij} = -\infty$ otherwise. Consequently, each prediction can attend to all the preceding question–answer exchanges. The causal context for the answer at level ℓ is $C_\ell = (v(x), q_1, a_1, \dots, q_{\ell-1}, a_{\ell-1}, q_\ell)$, and the teacher-forced training objective is

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{1}{L}\sum_{\ell=1}^{L} \log p_\theta(a_\ell \mid C_\ell; M), \tag{2}$$

where p_θ denotes the probability assigned to the ground-truth answer at each level.

Table 1: Taxonomic constraint induced by ground-truth ancestors on the 37,710-image training set. Values denote the proportion of questions with exactly one taxonomically compatible candidate.

Transition	# Questions	Compatible Candidate
Class → Order	37,710	12.13%
Order → Family	37,710	76.09%
Family → Genus	37,710	56.32%
Genus → Species	37,710	63.86%
All	**150,840**	**52.10%**

Under this formulation, the model has access to preceding questions and ground-truth ancestor answers, whereas each taxonomic level is queried independently without historical context during evaluation. This train–test mismatch creates opportunities for textual shortcut learning, potentially encouraging reliance on preceding taxonomic cues rather than visual evidence. Similar dependencies arise in the single-response formulation of Li et al. (2026), where finer-level predictions can access preceding ancestor labels in a coarse-to-fine taxonomic sequence. We therefore investigate the informativeness of historical textual context and the extent to which models exploit it in practice.

3.2 Textual Shortcuts Analysis

We investigate two questions: whether historical textual context can reveal the correct answer without visual evidence, and whether trained models rely on this information when making predictions.

Finding 1. Hierarchical context can uniquely determine the answer for many exchanges. We first examine how much information the hierarchical textual context alone provides about the answer by analyzing the reference taxonomy associated with the training set of Tan et al. (2026). As shown in Table 1, among the 150,840 training questions below the Class level, 52.10% contain only one candidate that is taxonomically compatible with the ground-truth ancestor path. Thus, for more than half of these questions, the answer can in principle be uniquely determined from the hierarchical context and the reference taxonomy without examining the image. This proportion further increases to 76.09% for Order→Family questions. For example, consider a Family-level question with four candidates: `Sapindaceae`, `Betulaceae`, `Pinaceae`, and `Salicaceae`. Given the ground-truth ancestor path `Pinopsida` → `Pinales`, only `Pinaceae` is taxonomically compatible, while the other three candidates can be ruled out without visual evidence. This finding reveals a potential hierarchical shortcut in the training data, but does not establish whether models have learned the underlying taxonomic relations or actually exploit them when making predictions.

Finding 2. Model predictions are highly sensitive to historical textual context. We next examine whether the trained model relies on historical context through controlled input interventions. Using a fixed HierSFT checkpoint, we provide the model with the target question and ground-truth question–answer exchanges from preceding levels, and compare three conditions. *Correct Context* retains the original image and textual history. *Text Swap* preserves the image and current question but replaces the immediate parent's correct answer with a taxonomically incompatible option from its original candidate set. Conversely, *Image Swap* preserves the textual context but replaces the image with one from a taxonomically incompatible Class. Detailed intervention settings are provided in Appendix G.2. All predictions are evaluated against the original image's target labels. As shown in Fig. 2, Text Swap causes substantially larger accuracy drops than Image Swap across all three evaluation splits. On Plantae Seen, for example, accuracy falls from 82.28% under Correct Context to 21.88% under Text Swap, compared with 70.68% under Image Swap. These results show that changing historical textual information alone can substantially alter the model's predictions, providing evidence that the model relies on preceding taxonomic cues under teacher-forced diagnostic conditions. These results indicate that the model is substantially more sensitive to perturbations of historical textual context than to the tested image replacements.

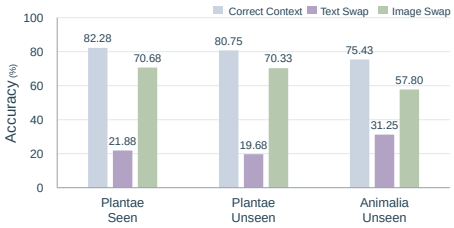


Figure 2: Effects of textual and visual interventions on HierSFT.

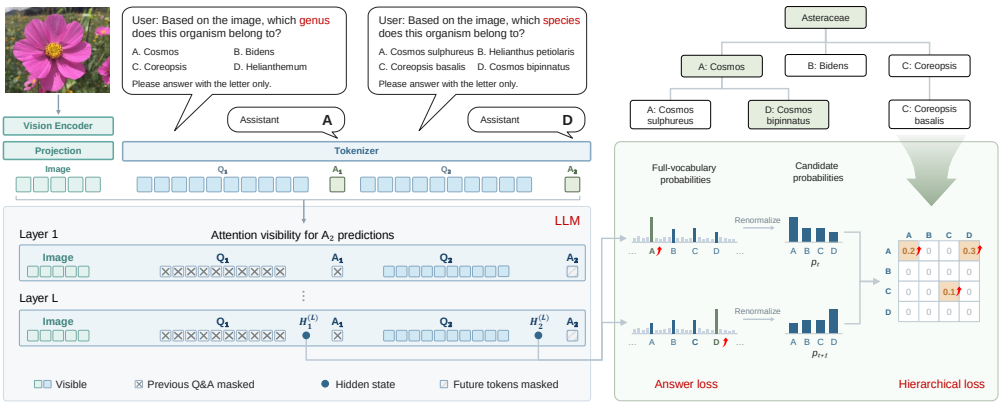


Figure 3: Overview of our framework. During training, EICA processes all level-wise question–answer exchanges in a single decoder forward pass, preserving access to the shared visual prefix while isolating each exchange from preceding ones. The resulting candidate distributions are then coupled by a taxonomy-derived compatibility matrix to enforce consistency across adjacent levels.

Takeaway. Together, these findings reveal a potential shortcut in hierarchical supervision: historical question–answer exchanges can provide sufficient taxonomic cues to identify the correct answer, and trained models exhibit substantial reliance on this context. This motivates our approach of preserving multi-level supervision while isolating historical exchanges to encourage visually grounded predictions at each taxonomic level.

4 METHODOLOGY

4.1 EXCHANGE-ISOLATED CAUSAL ATTENTION

Our analysis in Sec. 3.2 reveals that preceding question-answer exchanges can provide sufficient taxonomic cues to identify the correct answer without visual evidence, and that trained models exhibit substantial sensitivity to such context. To mitigate this textual shortcut, we introduce Exchange-Isolated Causal Attention (EICA), which augments the standard causal mask to restrict cross-level textual dependencies while preserving access to the visual input and current exchange (see Fig. 3).

Following the notation in Sec. 3.1, let $S = [v(x), q_1, a_1, \ldots, q_L, a_L]$ denote the hierarchical training sequence, and let $E_\ell = [q_\ell, a_\ell]$ represent the question–answer exchange at level ℓ. EICA preserves access to the shared visual prefix and the current exchange while masking all preceding exchanges. Specifically, we introduce an exchange-level masking bias into standard causal attention:

$$\widetilde{\mathcal{A}}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M + B\right)V, \tag{3}$$

where M is the causal mask, and $B_{ij} = -\infty$ if $i \in E_\ell$ and $j \in E_k$ for $k < \ell$, and 0 otherwise.

Under EICA, each taxonomic level retains access to the shared visual input and the tokens within its own exchange, while textual information from earlier levels is inaccessible. In particular, preceding questions, candidate sets, and ground-truth answers cannot serve as conditioning signals for subsequent predictions. The resulting effective context for level ℓ is therefore restricted to the image and the current level-specific interaction, rather than the complete hierarchical conversation. We perform supervised fine-tuning under the modified attention mask using the standard answer-level objective:

$$\mathcal{L}_{\text{ans}} = -\frac{1}{L}\sum_{\ell=1}^{L} \log p_\theta\left(a_\ell \mid C_\ell; M, B\right), \tag{4}$$

where C_ℓ is the causal context defined in Sec. 3.1. We complement the answer-level objective with an auxiliary format loss $\mathcal{L}_{\text{fmt}}$, which averages the cross-entropy for predicting ground-truth end-of-turn markers across exchanges to ensure proper response termination.

This exchange-isolated training scheme offers three key advantages:

- **Architectural compatibility.** It modifies only the attention mask, preserving the original autoregressive architecture without introducing additional trainable parameters. It can therefore be integrated into existing VLMs without modifying their attention operators.
- **Visual grounding.** By isolating historical question–answer exchanges, it aligns the training context with independent level-wise inference. Each prediction retains access to the image and its current question, encouraging visual grounding rather than historical textual cues.
- **Preserved hierarchical supervision.** It decouples cross-level textual dependencies without discarding hierarchical relationships, which are instead explicitly modeled by the hierarchy loss introduced in the next section.

4.2 Probabilistic Hierarchy Loss

While EICA removes direct cross-level textual dependencies, it does not explicitly enforce consistency among predictions at different taxonomic levels. To address this limitation, we introduce a probabilistic hierarchy loss that leverages parent–child relationships in the reference taxonomy to encourage consistent prediction distributions across adjacent levels.

We first obtain a probability distribution over the candidate categories at each taxonomic level. Let $\mathbf{z}_t$ denote the full-vocabulary logits at prediction position d_t, and let b_k denote the token ID of the k-th answer choice in $(\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D})$. The choice-level probability distribution is computed as

$$\mathbf{p}_t = \mathrm{softmax}\left((\mathbf{z}_t[b_1], \mathbf{z}_t[b_2], \mathbf{z}_t[b_3], \mathbf{z}_t[b_4])^\top\right). \tag{5}$$

This yields a normalized distribution over the candidate categories for hierarchical regularization, while the answer loss $\mathcal{L}_{\mathrm{ans}}$ and formatting loss $\mathcal{L}_{\mathrm{fmt}}$ retain full-vocabulary normalization.

Let $c_{t,k}$ denote the category label of the k-th candidate at level t. To relate these candidate-level distributions across adjacent hierarchy levels, we use the reference taxonomy to define the sample-specific compatibility matrix $C_t \in \{0,1\}^{4\times 4}$:

$$C_t(k,j) = \mathbb{1}[(c_{t,k}, c_{t+1,j}) \in \mathcal{E}], \tag{6}$$

where k and j index the candidates at levels t and $t+1$, respectively. The matrix C_t identifies the valid parent–child edges between candidate labels at the two adjacent levels according to the reference taxonomy. Accordingly, the probability mass assigned to compatible candidate pairs is

$$m_t = \mathbf{p}_t^\top C_t \mathbf{p}_{t+1}. \tag{7}$$

This quantity measures the total probability mass assigned to taxonomically compatible candidate pairs across the two adjacent levels. To encourage large compatible probability mass, we define the hierarchical loss as

$$\mathcal{L}_{\mathrm{hier}} = -\frac{1}{L-1}\sum_{t=1}^{L-1}\log(\max\{m_t, \epsilon\}), \tag{8}$$

where $\epsilon > 0$ ensures numerical stability. Importantly, the compatibility matrix accounts for all taxonomically valid candidate pairs rather than only the ground-truth pair. As reported in Appendix E, an average of 2.61 out of the 16 possible candidate pairs is taxonomically valid, corresponding to a density of 16.29%. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ans}} + \lambda_{\mathrm{fmt}}\mathcal{L}_{\mathrm{fmt}} + \lambda_{\mathrm{hier}}\mathcal{L}_{\mathrm{hier}}. \tag{9}$$

5 Experiments

5.1 Experimental Setup

Datasets. We follow the data construction and multi-turn visual instruction format of Tan et al. (2026), based on LLaVA (Liu et al., 2023). The iNat21-Plant (Van Horn et al., 2021) dataset contains 4,271 species, of which 3,771 are used for fine-tuning (with ten images per species), and the remaining 500 species are held out. For evaluation on iNat21-Plant, we use two splits from Tan et al. (2026): *Plantae Seen* contains 37,710 images from the 3,771 species used during fine-tuning, but the evaluation images are disjoint from the fine-tuning images; *Plantae Unseen* contains 5,000 images from the 500 held-out species. For cross-domain evaluation, we additionally use *Animalia Unseen*,

Table 2: Performance using Qwen2.5-VL-3B-Instruct and Qwen3-VL-8B-Instruct backbones.

Method	**Plantae Seen** ($N = 37{,}710$)					**Plantae Unseen** ($N = 5{,}000$)					**Animalia Unseen** ($N = 53{,}880$)				
	HCA	Acc_{leaf}	POR	S-POR	TOR	HCA	Acc_{leaf}	POR	S-POR	TOR	HCA	Acc_{leaf}	POR	S-POR	TOR
Qwen2.5-VL-3B	10.51	38.33	65.35	53.22	47.92	10.18	37.26	63.57	52.43	46.32	15.93	38.86	72.22	64.80	59.90
HierSFT [CVPR26]	23.56	44.94	75.70	68.04	64.15	19.09	43.13	71.73	63.46	58.63	21.04	45.76	74.71	67.27	63.16
TARA [CVPR26]	18.06	44.05	73.29	62.80	58.68	16.65	44.02	70.81	60.75	55.94	20.44	45.37	75.05	67.60	63.38
Ours	**28.29**	**48.73**	**79.48**	**72.33**	**69.08**	**23.05**	**46.86**	**74.73**	**66.88**	**62.58**	**22.96**	**47.72**	**75.79**	**68.52**	**64.58**
Gain	+4.73	+3.79	+3.78	+4.29	+4.93	+3.96	+3.73	+3.00	+3.42	+3.95	+1.92	+1.96	+1.08	+1.25	+1.42
Qwen3-VL-8B	18.85	40.85	73.07	64.94	60.40	16.16	40.66	70.66	62.29	57.12	19.61	39.81	75.22	69.50	64.89
HierSFT [CVPR26]	29.45	46.23	79.47	74.24	70.51	23.83	44.05	75.36	69.06	64.76	24.08	43.90	77.02	71.51	67.29
TARA [CVPR26]	24.33	45.93	77.29	70.20	66.23	21.47	46.29	74.41	66.83	62.37	23.20	45.98	77.26	71.25	67.03
Ours	**32.40**	**50.36**	**81.67**	**76.20**	**72.99**	**25.07**	**48.06**	**76.72**	**69.89**	**65.87**	**25.25**	**46.24**	**77.75**	**72.06**	**68.02**
Gain	+2.95	+4.13	+2.20	+1.96	+2.48	+1.24	+4.01	+1.36	+0.83	+1.11	+1.17	+2.34	+0.73	+0.55	+0.73

which contains 53,880 images from the iNat21-Animal taxonomy. At each taxonomic level, the candidate set includes the ground-truth label and three most similar distractors selected by SigLIP image–text similarity (Zhai et al., 2023), and is fixed across methods. At test time, each taxonomic level is queried independently, without supplying either ground-truth or predicted ancestor answers.

Compared Methods. We experiment with two VLM backbones, Qwen2.5-VL-3B-Instruct and Qwen3-VL-8B-Instruct, and compare our method with two recent methods: *HierSFT* (Tan et al., 2026), and *TARA* (He et al., 2026b). *HierSFT* follows the sequential VQA protocol where subsequent training turns can attend to earlier question–answer pairs, and only the language decoder is fine-tuned with LoRA (Hu et al., 2021). *TARA* alternates reinforcement-learning updates with representation alignment to a frozen BioCLIP2 (Gu et al., 2026). Unless otherwise specified, we adapt its language decoder with LoRA while fully training its two auxiliary alignment projectors. Thus, our TARA results reflect this adaptation rather than the original full-parameter training.

Implementation Details. For both VLM backbones, all methods involving fine-tuning use the same training images. We freeze the vision encoder and vision–language projector, and train for one epoch with an effective global batch size of 128. For our method, λ_{hier} is linearly warmed up from zero to its target value during the first 20% of training, and we set $\lambda_{\text{fmt}} = 0.05$. Appendices A and B provide more implementation details. All results are averaged over three random seeds. We report the mean values in the main text, with the corresponding standard deviations provided in Appendix H.

Evaluation Metrics. Following Tan et al. (2026), we report five complementary metrics. Hierarchical Consistent Accuracy (HCA) measures the proportion of images correctly classified at all evaluated levels. Species-level accuracy $(\text{Acc}_{\text{leaf}})$ measures leaf-level prediction accuracy. Point-Overlap Ratio (POR) averages the proportion of correct levels, equivalent to mean level-wise accuracy for equal-length paths. Strict Point-Overlap Ratio (S-POR) measures the longest consecutive sequence of correct predictions normalized by path length, while Top Overlap Ratio (TOR) averages correctness over adjacent-level pairs. Formal definitions are provided in Appendix D. All metrics are reported as percentages, with higher values indicating better performance.

5.2 Main Results

We evaluate our method across seen categories, unseen categories, and a previously unseen taxonomic domain. Table 2 reports the comparison with the untuned backbones, HierSFT, and TARA. Our method achieves the best performance on all five metrics across both backbones.

Results on Seen Categories. On Plantae Seen, our method outperforms HierSFT by 4.73% and 2.95% in HCA under the Qwen2.5-VL-3B and Qwen3-VL-8B backbones, respectively. These improvements are accompanied by consistent gains in leaf-level accuracy and all hierarchical overlap metrics. Notably, gains over HierSFT extend beyond leaf recognition to accurate prediction of complete hierarchical paths, suggesting that our method improves hierarchical consistency beyond merely enhancing final-category recognition. Our method also consistently outperforms TARA, demonstrating the effectiveness of the proposed hierarchical optimization strategy.

Results on Unseen Categories and Domains. The improvements generalize beyond observed categories. On Plantae Unseen and Animalia Unseen, our method consistently outperforms HierSFT, achieving HCA gains of 1.17–3.96 percentage points and leaf-level accuracy gains of 1.96–4.01 percentage points across different backbone scales. It also achieves consistent improvements in

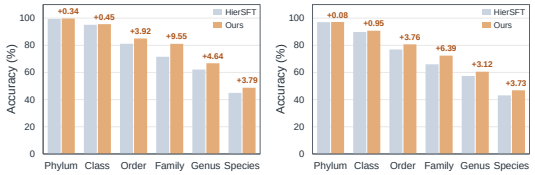
Table 3: Performance using Qwen2.5-VL-3B-Instruct backbone under the one-shot setting.

Method	Plantae Seen ($N = 37{,}710$)					Plantae Unseen ($N = 5{,}000$)					Animalia Unseen ($N = 53{,}880$)				
	HCA	Acc_{leaf}	POR	S-POR	TOR	HCA	Acc_{leaf}	POR	S-POR	TOR	HCA	Acc_{leaf}	POR	S-POR	TOR
Qwen2.5-VL-3B	10.51	38.33	65.35	53.22	47.92	10.18	37.26	63.57	52.43	46.32	15.93	38.86	72.22	64.80	59.90
HierSFT [CVPR26]	18.39	41.34	72.75	63.26	59.18	15.78	39.20	69.77	60.62	55.68	18.86	41.55	73.80	66.65	62.15
TARA [CVPR26]	16.72	42.59	71.88	60.92	56.73	15.31	**41.31**	69.17	58.89	53.89	18.79	41.83	73.97	66.76	62.28
Ours	**21.64**	**42.80**	**75.78**	**67.19**	**63.53**	**19.57**	41.29	**72.61**	**64.43**	**59.87**	**19.73**	**41.99**	**74.51**	**67.61**	**63.19**
Gain	+3.25	+1.46	+3.03	+3.93	+4.35	+3.79	+2.09	+2.84	+3.81	+4.19	+0.87	+0.44	+0.71	+0.96	+1.04

POR, S-POR, and TOR. Compared with TARA, our method achieves higher performance across both unseen evaluation splits and both backbone scales. These results demonstrate that our method maintains effective hierarchical recognition under category and domain shifts.

5.3 Discussion and Ablation Study

Level-wise Recognition Accuracy. Fig. 4 shows that our method consistently improves recognition accuracy across all taxonomic levels on both Plantae Seen and Plantae Unseen. The improvements are particularly evident at intermediate levels, such as Family and Order, where accurate predictions are essential for maintaining a coherent hierarchical path. Meanwhile, Species-level recognition also benefits from our method, indicating that hierarchical improvements do not come at the cost of fine-grained recognition. These results suggest that HCA improvements arise not only from enhanced leaf-level classification but also from more accurate intermediate-level predictions, leading to more reliable hierarchical inference.



(a) Plantae Seen (b) Plantae Unseen

Figure 4: Level-wise accuracy on the Plantae test sets using Qwen2.5-VL-3B-Instruct.

One-shot Hierarchical Recognition. To evaluate the data efficiency of our method, we conduct experiments under the one-shot protocol of TARA He et al. (2026b), with only one training image provided per species. As shown in Table 3, our method consistently outperforms both TARA and HierSFT across all evaluation splits and hierarchical metrics, while maintaining competitive leaf-level accuracy. For example, on Plantae Seen, our method achieves an HCA of 21.64%, compared with 16.72% for TARA and 18.39% for HierSFT. These results demonstrate its effectiveness in data-scarce settings with limited supervision. See Appendix F for detailed experimental settings.

Evaluation on General VQA Benchmarks. Following HierSFT Tan et al. (2026),we further evaluate our method on general VQA benchmarks to assess whether hierarchical fine-tuning preserves general-purpose visual reasoning. We use VLMEvalKit (Duan et al., 2024) with identical prompting and decoding settings across models. As shown in Table 4, our method achieves comparable or better performance than HierSFT on both MME (Fu et al., 2026) and MMBench (Liu et al., 2024b), with consistent improvements across the two benchmarks. These results indicate that our framework enhances hierarchical recognition while maintaining the general VQA capability of the pretrained VLM.

Table 4: General VQA performance. MME reports total scores, while MMBench (English dev) reports accuracy (%).

Method	MME↑	MMBench↑
Qwen2.5-VL-3B	2177.77	79.12
HierSFT [CVPR26]	2191.22	79.21
Ours	**2210.65**	**79.84**
Δ (vs. HierSFT)	+19.43	+0.63

Qualitative Results. Fig. 5 presents an example of hierarchical recognition under independent level-wise queries. Given the plant image, HierSFT correctly identifies the genus and species but makes inconsistent predictions at intermediate levels, violating the underlying taxonomy. In contrast, our method predicts a coherent taxonomic path from class to species, correctly identifying both coarse- and fine-grained categories. This example highlights the importance of modeling hierarchical dependencies rather than optimizing each level independently.

Ablation Study. Table 5 presents a progressive analysis of our framework. We first investigate the effect of EICA by comparing different context masking strategies, and then analyze the additional contribution of the hierarchy loss.

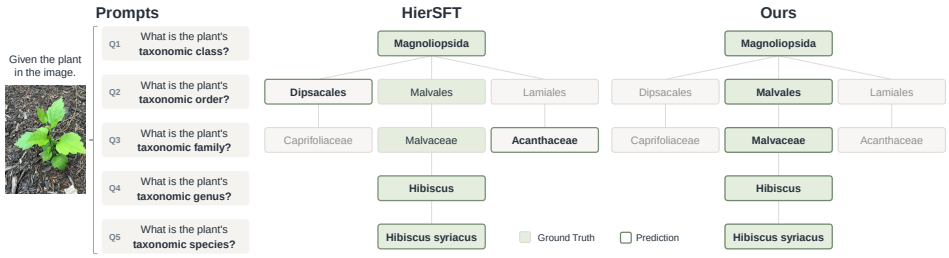


Figure 5: Qualitative comparison of hierarchical predictions between HierSFT and our method. Independent level-wise queries reveal taxonomy-inconsistent predictions from HierSFT and coherent hierarchical paths produced by our method.

Table 5: Ablation of training components and masking scope.

Method	Plantae Seen ($N = 37{,}710$)			Plantae Unseen ($N = 5{,}000$)			Animalia Unseen ($N = 53{,}880$)		
	HCA↑	$\text{HCA}_L \uparrow$	$\text{Acc}_{\text{leaf}} \uparrow$	HCA↑	$\text{HCA}_L \uparrow$	$\text{Acc}_{\text{leaf}} \uparrow$	HCA↑	$\text{HCA}_L \uparrow$	$\text{Acc}_{\text{leaf}} \uparrow$
Qwen2.5-VL-3B	10.51	27.42	38.33	10.18	27.32	37.26	15.93	40.99	38.86
+SFT (HierSFT)	23.56	52.42	44.94	19.09	44.26	43.13	21.04	45.99	45.76
Ablation on EICA									
EICA w/ Answer Masking	23.39	50.95	45.91	18.73	43.03	43.54	20.92	45.43	46.04
EICA w/ Exchange Masking	26.65	54.18	**49.19**	21.32	45.48	**46.87**	22.05	45.62	**48.32**
EICA + Hierarchy Loss	**28.29**	**58.04**	48.73	**23.05**	**49.19**	46.86	**22.96**	**48.10**	47.72

(1) *Effect of Exchange- Isolated Causal Attention.* HierSFT serves as a supervised fine-tuning baseline without context masking. Compared with the zero-shot Qwen2.5-VL-3B-Instruct backbone, HierSFT substantially improves HCA on Plantae Seen, demonstrating the benefit of supervised fine-tuning for hierarchical recognition. However, its reliance on the full question–answer context may allow the model to exploit cross-level textual dependencies. Applying EICA with answer-only masking (EICA w/ Answer Masking) achieves similar performance to HierSFT, indicating that masking only previous answers is insufficient to remove such shortcuts. In contrast, masking complete question–answer exchanges (EICA w/ Exchange Masking) substantially improves HCA to 26.65%, 21.32%, and 22.05% on Plantae Seen, Plantae Unseen, and Animalia Unseen, respectively. These gains demonstrate that exchange-level isolation effectively reduces textual propagation and encourages predictions grounded in visual evidence.

(2) *Effect of Hierarchy Loss.* Building upon EICA with exchange masking, adding the hierarchy loss (EICA + Hierarchy Loss) further improves HCA to 28.29%, 23.05%, and 22.96% across the three evaluation splits. To better understand this improvement, we introduce HCA_L, which measures the proportion of samples with a fully correct hierarchical path among those with correct leaf-level predictions. The improvement in HCA_L indicates that the hierarchy loss mainly enhances ancestor–leaf consistency rather than simply improving leaf recognition. Together with the stable leaf-level accuracy, these results reveal complementary roles of the two components: EICA improves recognition by reducing cross-level textual shortcuts, whereas the hierarchy loss promotes coherent predictions between leaves and their corresponding ancestor paths. We further examine how the hierarchy-loss weight affects the trade-off between hierarchical consistency and leaf-level accuracy in Appendix G.1.

6 Conclusion

In this work, we highlight a fundamental challenge in hierarchical visual instruction tuning: sequential supervision can introduce textual shortcuts that encourage reliance on preceding taxonomic information rather than visual evidence. We show that hierarchical supervision can instead be formulated by isolating historical textual context while explicitly enforcing taxonomic consistency across prediction levels. Our experiments demonstrate the effectiveness of this approach across different VLM backbones, category splits, and domains, including data-scarce settings. More broadly, our findings suggest that hierarchical knowledge need not be propagated through textual history to benefit visual recognition, offering a direction for training VLMs that better balance visual grounding and structural consistency.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, code development, and experimental workflow organization. All AI-assisted text, code, analyses, and claims were reviewed and verified by the authors. We take responsibility for the final content of this work, including artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the model formulation, attention mask definition, training objectives, and inference protocol in Section 4, and describe the datasets, baselines, implementation settings, and evaluation metrics in Section 5. Complete preprocessing details, hyperparameters, and additional ablations will be included in the appendix and supplementary material.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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*, 2025b.

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.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

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.

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

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. In *International Conference on Learning Representations*, volume 2025, pp. 89927–89944, 2025.

Hulingxiao He, Zhi Tan, and Yuxin Peng. Learning taxonomic trees with hierarchical representation regularization for large multimodal models. *arXiv preprint arXiv:2607.02909*, 2026a.

Hulingxiao He, Zhi Tan, and Yuxin Peng. Taxonomy-aware representation alignment for hierarchical visual recognition with large multimodal models. *arXiv preprint arXiv:2603.00431*, 2026b.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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*, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

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*, pp. 26342–26362. PMLR, 2023.

Avik Pal, Max Van Spengler, Guido D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 87371–87399, 2025.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

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.

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

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, 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.

Renzhen Wang, De Cai, Kaiwen Xiao, Xixi Jia, Xiao Han, and Deyu Meng. Label hierarchy transition: Delving into class hierarchies to enhance deep classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Wenhao Wang, Yifan Sun, Wei Li, and Yi Yang. Transhp: Image classification with hierarchical prompting. *Advances in Neural Information Processing Systems*, 36:28187–28200, 2023.

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.

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.

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 2740–2748, 2015.

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.

Hongtao Yu, Yuxin Peng, Serge Belongie, and Xiu-Shen Wei. Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 44677–44700, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

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? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024.

Hao Zheng, Shunzhi Yang, Zhuoxin He, Jinfeng Yang, and Zhenhua Huang. Hierarchical cross-modal prompt learning for vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1891–1901. IEEE, 2025.

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*, pp. 1106–1117, 2020.

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*, pp. 5645–5655, 2025.

A Implementation Details

For the main 10-shot experiments, we conduct experiments using Qwen2.5-VL-3B-Instruct and Qwen3-VL-8B-Instruct as VLM backbones. For both models, we freeze the original vision encoder and vision–language projector and optimize the language decoder using LoRA. The LoRA adapters are applied to the eligible linear modules of the language decoder with rank 64, scaling factor 64, and dropout 0.2. All fine-tuned methods use the same training images, one training epoch, and an effective global batch size of 128. We use AdamW with a learning rate of 5×10^{-5}, zero weight decay, a cosine learning-rate schedule, and a warm-up ratio of 0.03. For TARA (He et al., 2026b), we use the same training images and VLM optimization settings described above, while retaining its native update procedure. Each image is assigned one target level from Order, Family, Genus, and Species according to the ratio 1:2:4:8, and the model is trained to output the category name. The BioCLIP2 teacher remains frozen, while TARA's two auxiliary alignment projectors are fully trained using their native optimization settings. For our method, the hierarchy-loss coefficient is set to $\lambda_{\text{hier}} = 3.0$, with a linear warm-up from zero to the respective value over the first 20% of training. The formatting-loss coefficient is set to $\lambda_{\text{fmt}} = 0.05$ for both backbones. We report the mean and standard deviation over three training seeds, while the frozen base model is evaluated once because it involves no training randomness. All experiments are implemented in PyTorch (Paszke et al., 2019) and conducted on 8 NVIDIA GeForce RTX 5090 GPUs.

B Prompt Template

Training. Our method uses five four-choice questions per image, covering Class, Order, Family, Genus, and Species. The system prompt and level-wise question template are shown below.

```
System:
You are a helpful assistant.

User:
Based on the image, which {level} does this organism belong to?
A. {candidate_A}
B. {candidate_B}
C. {candidate_C}
D. {candidate_D}

Please answer with the letter only.

Assistant:
{ground_truth_letter}
```

Here, {`level`} is replaced with `class`, `order`, `family`, `genus`, or `species`, and the target is a single letter from {`A`, `B`, `C`, `D`}. The five question–answer pairs are packed into one sequence, with the image supplied in the first user message.

Evaluation. Each requested level is evaluated independently without ancestor history. The following template illustrates a plant query.

```
System:
You are an expert in hierarchical image classification.
Given an image, classify the organism at the requested
hierarchy level by selecting the most appropriate option
from the provided choices. Respond with only the
corresponding letter.

User:
Given the plant in the image, what is its taxonomic
classification at the {level} level?
A. {candidate_A}
B. {candidate_B}
C. {candidate_C}
D. {candidate_D}
```

```
Answer with the option's letter from the given choices directly.
```

C Training and Inference Cost

To assess the practical efficiency of hierarchical supervision, we compare the training and inference costs of our method with the HierSFT baseline and TARA under the same experimental setting. Table 6 reports the training time, training GPU-hours, and inference time for the main Qwen2.5-VL-3B-Instruct experiments, averaged over three seeds. All methods are trained for one epoch on the same 37,710 Plantae training images. Our method requires 3.90 GPU-hours for training, comparable to HierSFT (3.69 GPU-hours) and substantially lower than TARA (41.21 GPU-hours), while achieving the lowest inference time at 41.70 minutes. Notably, Exchange- Isolated Causal Attention and the hierarchy loss introduce neither additional trainable parameters nor extra inference stages. These results demonstrate that our approach incorporates hierarchical supervision with minimal computational overhead.

Table 6: Training and inference cost for the main Qwen2.5-VL-3B experiments. Results are reported as mean ± standard deviation over three seeds.

Method	Training (min)	GPU-hours	Inference (min)
HierSFT [CVPR26]	$27.66_{\pm 0.63}$	$3.69_{\pm 0.08}$	$44.35_{\pm 0.10}$
TARA [CVPR26]	$309.06_{\pm 1.40}$	$41.21_{\pm 0.19}$	$62.03_{\pm 0.36}$
Ours	$29.24_{\pm 5.76}$	$3.90_{\pm 0.77}$	$41.70_{\pm 0.22}$

D Evaluation Metrics

For the i-th test sample and hierarchy level t, let $\widehat{y}_t^i$ and y_t^i denote the predicted and ground-truth categories, respectively. We define the correctness indicator as

$$c_t^i = \mathbb{1}\big[\widehat{y}_t^i = y_t^i\big]\,. \tag{10}$$

Hierarchical Consistent Accuracy (HCA). HCA (Park et al., 2025; Wu et al., 2024) measures the proportion of images for which all categories along the complete taxonomic path are predicted correctly:

$$\mathrm{HCA} = \frac{1}{N}\sum_{i=1}^{N}\prod_{t=1}^{L} c_t^i. \tag{11}$$

It is therefore a strict path-level metric: an error at any hierarchy level makes the corresponding image incorrect under HCA.

Leaf-Level Accuracy ($\mathrm{ACC}_{\mathrm{leaf}}$). Leaf-level accuracy (Liu et al., 2024a; Zhang et al., 2024; He et al., 2025) evaluates only the prediction at the finest hierarchy level:

$$\mathrm{ACC}_{\mathrm{leaf}} = \frac{1}{N}\sum_{i=1}^{N} c_L^i. \tag{12}$$

Conditional Hierarchical Consistency HCA_L. Following Li et al. (2026), HCA_L measures hierarchical consistency conditioned on a correct leaf-level prediction:

$$\mathrm{HCA}_L = \frac{\sum_{i=1}^{N}\prod_{t=1}^{L} c_t^i}{\sum_{i=1}^{N} c_L^i}, \tag{13}$$

which is equivalently given by $\mathrm{HCA}_L = \mathrm{HCA}/\mathrm{ACC}_{\mathrm{leaf}}$.

Point-Overlap Ratio (POR). POR (Yi et al., 2022) gives partial credit according to the proportion of correctly predicted nodes along each taxonomic path:

$$\mathrm{POR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\sum_{t=1}^{L} c_t^i. \tag{14}$$

Strict Point-Overlap Ratio (S-POR). S-POR (Tan et al., 2026) rewards only the longest contiguous sequence of correct predictions along a path:

$$\text{S-POR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\max_{1\le a\le b\le L}\left[(b-a+1)\prod_{t=a}^{b}c_t^i\right]. \tag{15}$$

Unlike POR, isolated correct predictions separated by an incorrect level do not jointly contribute to the same overlap segment.

Top Overlap Ratio (TOR). TOR (Wu et al., 2024) evaluates local hierarchical consistency by measuring the proportion of adjacent hierarchy-level pairs that are both predicted correctly:

$$\text{TOR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L-1}\sum_{t=1}^{L-1}c_t^i c_{t+1}^i. \tag{16}$$

All metrics are reported as percentages, and higher values indicate better performance.

E Compatibility Matrix Statistics

Our hierarchy loss couples predictions at adjacent hierarchy levels through taxonomy-derived compatibility matrices. To characterize the structural constraints imposed by these matrices on the candidate space, we analyze their sparsity on the training set. For each adjacent pair of hierarchy levels, we construct a 4×4 compatibility matrix over the corresponding candidate sets, yielding 16 possible parent–child pairs. A pair is marked as valid when the parent and child categories are directly connected in the reference taxonomy. Table 7 summarizes the compatibility statistics across all adjacent hierarchy-level pairs, with an average of 2.61 valid edges out of 16 possible candidate pairs. Thus, only about 16.3% of candidate pairs are taxonomically compatible on average, indicating that the taxonomy rules out a large portion of otherwise possible pairings and provides a strong structural constraint for coupling predictions across adjacent levels.

Table 7: Statistics of candidate-level taxonomic compatibility matrices in the training set.

Transition	# Matrices	Valid Edges	Density
Class → Order	37,710	3.80 / 16	23.77%
Order → Family	37,710	1.84 / 16	11.51%
Family → Genus	37,710	2.02 / 16	12.65%
Genus → Species	37,710	2.76 / 16	17.22%
All transitions	150,840	2.61 / 16	16.29%

F One-Shot Hierarchical Recognition Settings

Datasets. To further evaluate our method under a data-efficient setting, we follow the 1-shot data construction and target-level sampling protocol of TARA (He et al., 2026b). We use the same 3,771 plant images, with one image selected for each training species, and assign each image a target level from Order, Family, Genus, and Species according to the ratio 1:2:4:8. The image selection, target-level assignments, and target-level candidate sets are fixed across methods and training seeds. Each target-level instance contains four candidates, including the ground-truth category. For a fair comparison, we retain the original training formulation of TARA while adapting the hierarchical supervision used by HierSFT and our method to the same 1-shot setting. We evaluate all trained models on the same three benchmark splits: Plantae Seen (N = 37,710), Plantae Unseen (N = 5,000), and Animalia Unseen (N = 53,880). At test time, each image is queried independently at taxonomic levels, without access to ground-truth or predicted ancestor labels.

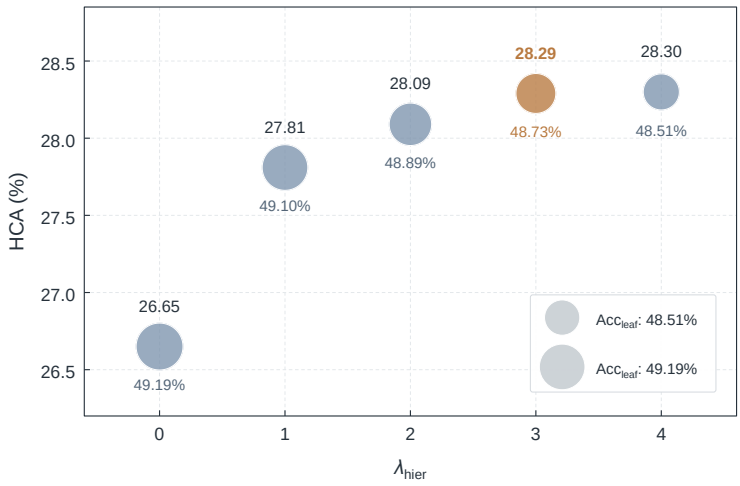


Figure 6: Effect of λ_{hier} on HCA and leaf-level accuracy on Plantae Seen, averaged over three random seeds.

Table 8: Accuracy drops relative to Correct Context under controlled history and image interventions (percentage points). Text Swap measures sensitivity to changes in historical answers, while Image Swap measures sensitivity to image replacement. Arrows indicate the desired diagnostic directions.

Method	**Plantae Seen** ($N = 1{,}000$)		**Plantae Unseen** ($N = 1{,}000$)		**Animalia Unseen** ($N = 1{,}000$)	
	Text Swap ↓	Image Swap ↑	Text Swap ↓	Image Swap ↑	Text Swap ↓	Image Swap ↑
HierSFT [CVPR26]	60.40	11.60	61.07	10.42	44.18	17.63
EICA w/o hierarchy loss	**15.50**	**19.67**	**20.07**	**17.20**	**17.27**	**26.00**

Implementation Details. All models are initialized from Qwen2.5-VL-3B-Instruct. Following the TARA 1-shot setting, we perform full-parameter fine-tuning of the VLM for all methods. We use AdamW with a learning rate of 1×10^{-6}, zero weight decay, linear learning-rate decay, and no warm-up. Models are trained for one epoch with an effective batch size of 16 image-level records on eight GPUs, using a per-device batch size of one and two gradient-accumulation steps. TARA retains its native auxiliary alignment components and update procedure, with a KL coefficient of 0.04. To align with TARA's target-level setting, HierSFT and Ours emphasize the assigned target by defining the answer loss as the target answer loss plus the mean answer loss over all ancestor levels. For our method, the hierarchy-loss weight is set to $\lambda_{\text{hier}} = 2.0$ and linearly ramped over the first 20% of training. We report HCA, leaf-level accuracy, POR, S-POR, and TOR, averaged over three training seeds with sample standard deviation.

G Additional Experiments

G.1 Effect of the Hierarchy-Loss Weight

Since the hierarchy loss is introduced to improve hierarchical consistency without substantially compromising fine-grained recognition, we examine the effect of λ_{hier} on Qwen2.5-VL-3B on Plantae Seen, averaging results over three random seeds. Figure 6 plots HCA against λ_{hier}, with circle sizes indicating leaf-level accuracy. Increasing λ_{hier} from 0 to 3 improves HCA by 1.64 percentage points, with only a slight decrease in leaf-level accuracy. Raising the weight further to 4 yields only a 0.01% gain in HCA, accompanied by an additional 0.22% decrease in leaf-level accuracy. This suggests that an appropriate hierarchy loss weight can substantially improve HCA while largely preserving leaf-level accuracy.

Table 9: Sample standard deviations of the main results across three random seeds.

Method	Plantae Seen ($N = 37{,}710$)					Plantae Unseen ($N = 5{,}000$)					Animalia Unseen ($N = 53{,}880$)				
	HCA	Acc_{leaf}	POR	S-POR	TOR	HCA	Acc_{leaf}	POR	S-POR	TOR	HCA	Acc_{leaf}	POR	S-POR	TOR
Qwen2.5-VL-3B *10-shot*															
HierSFT [CVPR26]	0.21	0.06	0.12	0.20	0.21	0.18	0.05	0.11	0.08	0.11	0.05	0.09	0.04	0.04	0.05
TARA [CVPR26]	0.29	0.29	0.43	0.45	0.58	0.25	0.54	0.27	0.33	0.42	0.28	0.39	0.20	0.24	0.31
Ours	0.05	0.13	0.07	0.09	0.12	0.34	0.26	0.19	0.22	0.29	0.14	0.18	0.07	0.07	0.07
Qwen3-VL-8B *10-shot*															
HierSFT [CVPR26]	0.13	0.14	0.07	0.12	0.12	0.20	0.13	0.09	0.16	0.15	0.08	0.22	0.02	0.03	0.02
TARA [CVPR26]	0.52	0.45	0.28	0.31	0.38	0.29	0.42	0.29	0.42	0.48	0.25	0.31	0.08	0.16	0.16
Ours	0.30	0.40	0.13	0.10	0.13	0.54	0.38	0.11	0.13	0.13	0.20	0.21	0.11	0.12	0.15
Qwen2.5-VL-3B *1-shot*															
HierSFT [CVPR26]	0.15	0.24	0.08	0.10	0.12	0.12	0.30	0.15	0.11	0.14	0.12	0.16	0.07	0.06	0.08
TARA [CVPR26]	0.42	0.53	0.33	0.46	0.51	0.33	0.65	0.44	0.51	0.59	0.10	0.53	0.15	0.12	0.14
Ours	0.21	0.08	0.19	0.29	0.31	0.09	0.13	0.22	0.33	0.32	0.10	0.07	0.05	0.08	0.08

G.2 Controlled History and Image Interventions

Experimental details. We use the Seed-1 Qwen2.5-VL-3B checkpoints of HierSFT and Our model without the hierarchy loss. For each of Plantae Seen, Plantae Unseen, and Animalia Unseen, we sample 1,000 images uniformly without replacement using seed 2027. Each image is evaluated at Order, Family, Genus, and Species, giving 4,000 decisions per split and condition. No correctness-based filtering is applied. We construct Class-to-Species conversations using the training prompt template, the benchmark candidate orderings, and annotated answer letters. Both checkpoints are evaluated with fixed parameters and standard causal attention. Predictions are obtained by taking the argmax over the logits of $\{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}\}$ at the position immediately preceding each target answer. The target answer itself is causally masked. All conditions are scored against the original image's labels. Reported drops are computed relative to each model's correct-context accuracy and expressed in percentage points. This diagnostic is separate from the ancestor-free benchmark evaluation.

Intervention construction. Correct Context uses the original image and ground-truth history. Text Swap changes only the immediate parent's answer letter, preserving the image, questions, candidate lists, and all other answers. The replacement is selected from the parent's incorrect candidates and must be incompatible with the correct target category under the reference taxonomy. Among eligible candidates, we select the one compatible with the largest number of current target candidates, breaking ties by the original candidate order. Each target-level intervention is constructed separately, without accumulating replacements across turns. Image Swap preserves the textual context and replaces the image with another image from the same sampled split. For each image, we scan forward cyclically through the seeded sample order and select the first image from a different Class. Replacement images may be reused; no class balancing or visual-similarity matching is performed. The same image pairing is used for all four target levels. Samples, textual replacements, and image pairings are identical across checkpoints. The reference taxonomy is used only for intervention construction and is not included in the model input.

Results. We apply the interventions introduced in Section 3.2 to assess whether EICA reduces the dependence on ancestor history observed in HierSFT. As shown in Table 8, Exchange- Isolated Causal Attention (EICA w/o hierarchy loss) reduces the accuracy drops under Text Swap from 44.18–61.07 to 15.50–20.07 percentage points across the three splits. Meanwhile, Image Swap produces larger drops for EICA, increasing from 10.42–17.63 to 17.20–26.00 percentage points across the three splits. These results show reduced sensitivity to ancestor-history replacement and greater responsiveness to image replacement after training with EICA.

H Standard Deviations of Results

We report the sample standard deviations in Table 9. These values correspond to the fine-tuned 10-shot and 1-shot results presented in Tables 2 and 3, respectively. The low standard deviations indicate that our method yields stable and consistent performance across different settings.