

In the appendices, we provide all implementation details to promote reproducibility of our work, more experimental results, and further discussions about this work. Section A is about the hierarchical image and text classification datasets. Section B supplements the main experiments in the paper, including the models, experiment setups, quantitative and qualitative results, and ablation studies. Section C presents various prompting and linear probing results, including those with the Gemma model and larger VLLMs up to 72B parameters. Section D allows one to reproduce our finetuning results and shows comparison results of text-only finetuning and on general VQA benchmarks. Finally, Sections E, F, and G broaden the discussion by including limitations, broader impacts of the work, and more related works.

## A Hierarchical Classification Benchmarks Curation

### A.1 Hierarchical Image Classification Benchmarks

Following prior work on hierarchical image classification, we adopted several commonly used hierarchical classification datasets, including ImageNet [13], iNaturalist-2021 [53], CUB-200-2011 [54] and Food-101 [5]. Due to the inherent unconstrained nature of open-ended predictions by VLLMs, even when provided with detailed instructions, their performance in open-ended hierarchical classification remains extremely limited, with an  $\text{Acc}_{\text{leaf}}$  as low as 3.88% by Qwen2.5-VL-7B. To more effectively evaluate model performance, we construct approximately one million multiple-choice questions in a four-choice VQA format. We provide the data construction process of hierarchy VQA benchmarks shown in Figure 7. To better illustrate the data format, we also provide several examples from different datasets as shown in Figure 8.

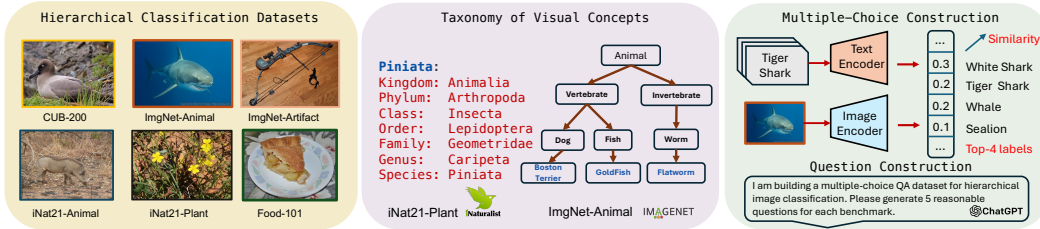


Figure 7: **Overview of hierarchical image classification benchmarks construction process.** Our hierarchical VQA benchmark is built on four datasets and covers six taxonomies. We first obtain the hierarchical structure for each taxonomy (biology standard and WordNet [67] semantics). Then, we use SigLIP [66] to generate four choices for each image based on the image-text similarities. Finally, we leverage GPT to generate the corresponding questions.

### A.2 Hierarchical Text Classification Benchmarks

For each curated hierarchical image classification benchmark, we derive a text-only variant. Concretely, we replace the image token in each prompt with the leaf node label of the corresponding hierarchy, while preserving the original answer choices, which were deliberately selected as *confusing labels*. An example of the resulting prompt template is illustrated in Figure 9.

## B Detailed Experiment Results

### B.1 Models

An overview of the models used in the evaluation experiments is provided in Table 6.

System Prompt: "You are an expert in hierarchical image classification. Given an image, classify it at its current hierarchy level by selecting the most appropriate option from the provided choices (labeled with letters). Respond with only the corresponding letter."

	<p>Based on the image, what is the taxonomic classification at the order level?</p> <p>A. Anseriforme B. Pelecaniforme C. Procellariiformes D. Podicipediformes</p> <p>Answer with the option's letter from the given choices directly.</p>
	<p>How can the bird in this image be categorized taxonomically?</p> <p>A. Pomarine jaeger B. Black-footed albatross C. Laysan albatross D. Sooty albatross</p> <p>Answer with the option's letter from the given choices directly.</p>
	<p>Given the plant in the image, what is its taxonomic classification at the order level?</p> <p>A. Gentianales B. Apiales C. Cornales D. Dipsacales</p> <p>Answer with the option's letter from the given choices directly.</p>
	<p>What is the systematic position of the plant in the image in the biological classification hierarchy?</p> <p>A. Bailey B. Draba C. Erysimum D. Barbarea</p> <p>Answer with the option's letter from the given choices directly.</p>

Figure 8: Examples of the prompt formats used in our four-choice hierarchical VQA tasks.

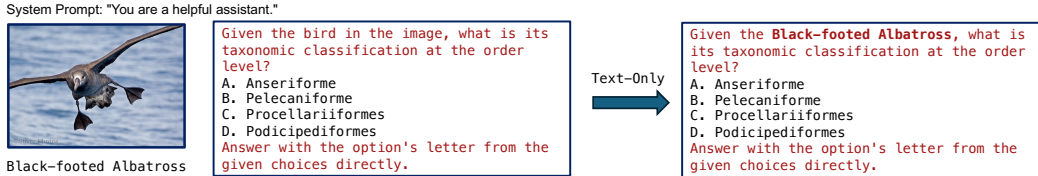


Figure 9: An example of the text QA construction from the hierarchical VQA.

Table 6: Models used in evaluation experiments and their sources.

Model	Source
LLaVA-OV-7B [29]	<a href="https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov">https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov</a>
InternVL2.5-8B [9]	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-8B">https://huggingface.co/OpenGVLab/InternVL2_5-8B</a>
InternVL3-8B [72]	<a href="https://huggingface.co/OpenGVLab/InternVL3-8B">https://huggingface.co/OpenGVLab/InternVL3-8B</a>
Qwen2.5-VL-7B [4]	<a href="https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct">https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct</a>
Qwen2.5-VL-32B [4]	<a href="https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct">https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct</a>
Qwen2.5-VL-72B [4]	<a href="https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct">https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct</a>
GPT-4o <sup>1</sup> [11]	<a href="https://openai.com/api/">https://openai.com/api/</a>
OpenCLIP [10]	<a href="https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K">https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K</a>
SigLIP [64]	<a href="https://huggingface.co/google/siglip-so400m-patch14-384">https://huggingface.co/google/siglip-so400m-patch14-384</a>

<sup>1</sup>GPT-4o results reported in this paper use the gpt-4o-2024-04-01-preview model for image-based tasks and the gpt-4o-2024-11-20 model for text-only evaluations.

## B.2 Hierarchical Evaluation Metrics

In addition to the metrics introduced in Section 2.2, we report results on three complementary metrics that probe different aspects of hierarchical classification ability.

**Point-Overlap Ratio (POR) [61].** To provide a more comprehensive evaluation of model performance across the full hierarchy, Yi et al. [61] proposed the point-overlap ratio, defined as:

$$\text{POR} = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{j=1}^{L_i} \mathbb{1}[f_{\theta}(x_i; \mathcal{Y}_j) = y_j^i]}{L_i}. \quad (3)$$

Unlike HCA, which requires an exact match along the entire path, POR allows for partial correctness by computing the average proportion of correctly predicted nodes. This metric offers a more fine-grained view of model performance over the taxonomy and captures the extent to which predictions align with the target hierarchy.

**Strict Point-Overlap Ratio (S-POR).** S-POR sharpens the original POR criterion by rewarding only *contiguous* stretches of correct predictions. For the  $i$ -th sample, we locate the longest run of consecutive correctly labelled layers and normalise by the hierarchy depth  $L_i$ :

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

This stricter definition penalizes sporadic correctness and encourages full-path consistency.

**Top Overlap Ratio (TOR).** Following Wu et al. [58], TOR measures *local* consistency by treating each pair of adjacent layers as an evaluation unit:

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

A TOR value of 1 indicates that every neighbouring pair is correctly predicted, whereas lower scores reflect violations of pairwise hierarchical coherence.

## B.3 Comprehensive Evaluation Results with All Metrics

Table 7: Evaluation results across all VLMs on CUB-200, ImgNet-Animal, ImgNet-Artifact and iNat21-Plant with POR, S-POR and TOR reported.

Model	CUB-200			ImgNet-Animal			ImgNet-Artifact			iNat21-Plant		
	POR	S-POR	TOR	POR	S-POR	TOR	POR	S-POR	TOR	POR	S-POR	TOR
<b>Open-Source VLLMs</b>												
LLaVA-OV-7B [29]	58.46	42.06	35.01	83.56	70.56	72.36	63.36	26.33	44.74	55.50	43.08	37.09
InternVL2.5-8B [9]	66.58	55.10	47.34	84.59	76.08	74.71	65.65	35.35	45.96	57.82	43.20	39.48
InternVL3-8B [72]	69.80	53.62	51.75	86.34	79.33	77.72	62.96	31.35	42.48	62.54	48.83	44.72
Qwen2.5-VL-7B [4]	80.85	70.52	67.97	90.52	84.52	83.84	64.12	26.47	44.53	71.95	59.37	57.45
Qwen2.5-VL-32B [4]	86.86	81.62	78.71	92.14	87.89	87.00	69.25	38.82	50.55	76.14	65.37	63.90
Qwen2.5-VL-72B [4]	89.79	86.20	83.63	92.43	88.74	87.77	67.21	30.86	48.10	80.23	70.92	69.86
<b>CLIP Models</b>												
OpenCLIP [10]	47.22	19.04	17.28	71.68	37.71	47.88	53.80	20.60	28.35	34.40	14.17	11.38
SigLIP [66]	66.56	45.84	41.79	78.95	48.46	59.24	50.90	16.15	24.77	34.67	16.97	15.00
<b>Proprietary VLLMs</b>												
GPT-4o [11]	94.46	92.29	91.00	93.33	89.16	88.83	70.45	40.42	51.47	79.92	69.37	68.62

We report more comprehensive evaluation results in Table 2 and Table 3. From these tables, we observe that VLLMs achieve relatively high POR scores, indicating strong classification performance across different levels of granularity. However, both S-POR and TOR scores remain relatively low, reflecting inconsistency in the model predictions.

Table 8: Evaluation results across all VLMs on iNat21-Animal with all metrics reported.

Model	Acc <sub>leaf</sub>	HCA	POR	S-POR	TOR
<b>Open-source VLLMs</b>					
LLaVA-OV-7B [29]	26.47	4.53	60.31	45.96	45.53
InternVL2.5-8B [9]	27.65	8.52	66.26	57.07	53.50
InternVL3-8B [72]	35.40	11.93	69.00	59.13	55.55
Qwen2.5-VL-7B [4]	41.66	19.73	74.80	66.92	63.71
Qwen2.5-VL-32B [4]	26.90	46.98	78.38	72.09	68.93
Qwen2.5-VL-72B [4]	35.73	54.20	81.76	76.05	73.55
<b>CLIP Models</b>					
OpenCLIP [10]	23.53	1.04	41.11	19.02	21.12
SigLIP [66]	12.71	2.15	38.24	38.24	33.95
<b>Proprietary VLLM</b>					
GPT-4o [11]	63.79	42.95	84.25	77.74	76.15

As the capacity of the VLLM increases (e.g., from Qwen2.5-VL 7B to 32B and 72B), the gap between POR and S-POR narrows, suggesting improved consistency in preserving the hierarchical structure during prediction. For GPT-4o, the gap between POR and S-POR on CUB-200 is only 2.17%, indicating that the correctly predicted nodes are mostly concentrated in the upper levels of the hierarchy. Additionally, the gap between TOR and POR also shrinks as model capacity increases, suggesting that better local hierarchical consistency is achieved.

While many individual nodes along the taxonomy path are predicted correctly, as evidenced by high POR scores, the probability of correctly predicting the entire path from root to leaf remains low. Although prior work [12] has noted that models often succeed in predicting coarse-grained categories but fail at fine-grained distinctions, our evaluation reveals that models sometimes predict the correct fine-grained label while misclassifying the corresponding coarse category. Therefore, beyond assessing fine-grained classification accuracy, it is equally important to evaluate the hierarchical consistency of VLLMs across different levels of granularity.

Compared with results in Table 2, models with higher POR, S POR, and TOR scores tend to exhibit better hierarchical consistency.

#### B.4 Visualization of Error Predictions

We visualize some hierarchical prediction errors made by open-source VLLMs in Figure 10.

#### B.5 Results on CUB-200 and iNat21-Plant with Random Choices

As shown in Table 2, using random choices significantly improves the model’s fine-grained accuracy—reaching up to 90% for Qwen2.5-VL. However, even with random choices, the gap between Acc<sub>leaf</sub> and HCA still exceeds 20%. For models like LLaVA-OV-7B and InternVLs, this gap is even more pronounced, reaching up to 40% on the iNat21-Plant benchmark, despite their relatively high Acc<sub>leaf</sub>. Therefore, our conclusion and analysis are still valid regardless of how the choices are constructed. However, the random choice construction does not reflect real-world scenarios, as it drastically reduces the task difficulty: three out of the four choices are likely to be completely unrelated to the query concept. For VLLMs, constructing similar choices based on image-text similarity better reflects practical scenarios, as end users are more likely to compare closely related concepts rather than unrelated ones.

#### B.6 Open-set Evaluation Results

We also evaluate the open-set scenario on Qwen2.5-VL-7B (Table 10), where no answer choices are provided. In this setting, model performance drops significantly, particularly on the iNat21-Plant

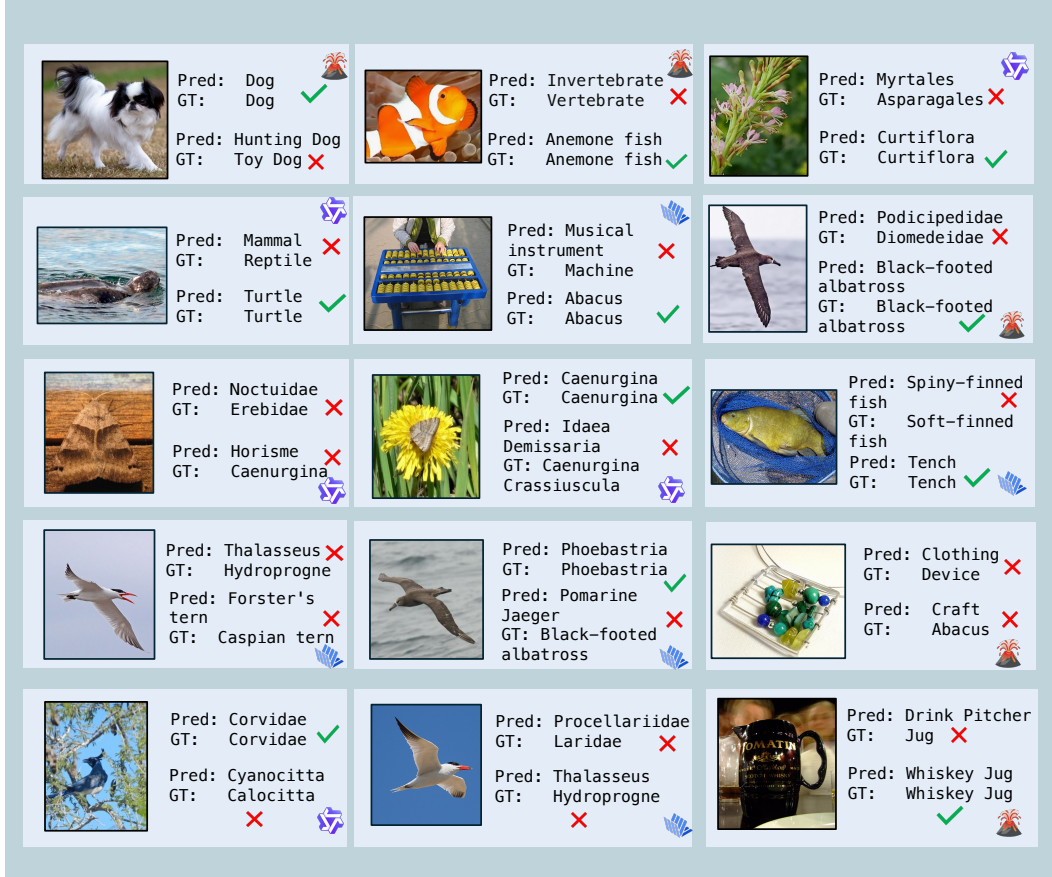


Figure 10: **Error Examples of the hierarchical predictions of VLLMs.** Examples are drawn from different VLLMs (Qwen2.5-VL-7B, InternVL2.5-8B and LLaVA-OV-7B) to reflect the diverse error modes observed across taxonomic levels.

941 benchmark, where the model struggles to generate correct answers. This results in very low fine-  
 942 grained accuracy and HCA.

## 943 B.7 Food-101 Results

944 A comprehensive evaluation on Food-101 with all metrics is shown in Table [10](#). On the Food-101  
 945 dataset, all models achieve relatively high fine-grained classification accuracy. Unlike other datasets,  
 946 LLaVA-OV-7B attains the highest HCA on this benchmark among the 7B/8B open-source VLLMs,  
 947 even though its leaf-level accuracy is not the highest.

## 948 B.8 Could the Poor Hierarchical Consistency Originate from the Four-choice VQA Format?

949 In general, VQA benchmarks [\[30, 36\]](#) adopt a multiple-choice question format, with four-choice  
 950 questions comprising the majority. Current open source VLLMs [\[4, 72, 9, 29\]](#) have already demon-  
 951 strated strong performance on these general VQA benchmarks. Therefore, the poor performance  
 952 observed in our setting is unlikely to be caused by the question format or prompt design, but rather  
 953 by the limitations of the VLLMs themselves. A more comprehensive analysis of the effects of  
 954 prompt design and question formats on the hierarchical understanding of VLLMs is provided in  
 955 Appendix [C](#).

Table 9: Hierarchical evaluation results (image) on CUB-200 and iNat21-Plant benchmarks with random choices.

Model	HCA	Acc <sub>leaf</sub>	POR	S-POR	TOR
<b>CUB-200</b>					
LLaVA-OV-7B [29]	40.25	86.14	78.12	59.30	58.94
InternVL2.5-8B [9]	64.20	91.06	88.36	77.13	76.91
InternVL3-8B [72]	67.50	93.80	90.55	75.96	82.23
Qwen2.5-VL-7B [4]	82.34	97.15	95.05	87.78	90.23
<b>iNat21-Plant</b>					
LLaVA-OV-7B [29]	28.41	69.04	75.36	58.53	60.03
InternVL2.5-8B [9]	36.45	75.97	80.25	60.81	67.22
InternVL3-8B [72]	51.94	89.70	87.19	70.96	76.28
Qwen2.5-VL-7B [4]	70.09	93.76	92.75	82.88	86.15

Table 10: HCA and leaf-level accuracy Acc<sub>leaf</sub> of Qwen2.5-VL-7B on open-set VQA tasks across five benchmarks.

Model	iNat21-Animal		iNat21-Plant		ImgNet-Artifact		ImgNet-Animal		CUB-200	
	HCA	Acc <sub>leaf</sub>	HCA	Acc <sub>leaf</sub>	HCA	Acc <sub>leaf</sub>	HCA	Acc <sub>leaf</sub>	HCA	Acc <sub>leaf</sub>
Qwen2.5-VL7B [4]	0.01	8.33	0.11	3.88	N/A	7.43	N/A	15.46	9.39	22.02

Table 11: Evaluation results across all VLMs on Food-101 with all metrics reported.

Model	Acc <sub>leaf</sub>	HCA	POR	S-POR	TOR
<b>Open-source VLLMs</b>					
LLaVA-OV-7B [29]	88.80	46.45	77.30	57.70	57.06
InternVL2.5-8B [9]	84.76	41.18	72.91	52.77	51.85
InternVL3-8B [72]	84.88	37.95	71.26	49.11	48.96
Qwen2.5-VL-7B [4]	90.51	43.11	75.15	53.48	53.13
Qwen2.5-VL-32B [4]	89.13	47.32	76.99	56.83	57.93
Qwen2.5-VL-72B [4]	92.02	52.00	80.46	60.95	62.33
<b>CLIP Models</b>					
OpenCLIP [10]	93.89	37.53	72.73	49.76	44.83
SigLIP [56]	97.17	42.49	73.68	50.03	51.69
<b>Proprietary VLLM</b>					
GPT-4o [11]	95.67	55.60	82.97	63.03	66.79

## C Supplementary for Section 3

### C.1 Prompt Engineering

To comprehensively assess how prompt design affects hierarchical classification performance, we evaluate a diverse set of prompt engineering strategies.

#### C.1.1 Prompt Variation

Across all benchmarks we employ five distinct prompt templates (Table 12), comprising both hierarchy-aware (Hierarchical) and general formulations (General). For CUB200, iNat21-Animal, and iNat21-Plant, we use two hierarchy-specific prompts and three general prompts. For ImgNet-Animal and ImgNet-Artifact, all five prompts are general because the corresponding taxonomy trees are highly unbalanced, making level-specific queries ill-posed. We report the results in Table 13,



966 averaging performance separately over general (General Prompts) and hierarchy-aware prompts (Hi-  
967 erarchy Prompts). Overall, hierarchy-aware prompts outperform general prompts on CUB-200 and  
968 iNat21-Plant.

Table 12: Prompt templates used across datasets. Placeholders: (i) **CUB-200**:  $\text{level} \in \{\text{order, family, genus, species}\}$ ; (ii) **iNat21**:  $\text{object} \in \{\text{animal, plant}\}$ ,  $\text{level} \in \{\text{kingdom, phylum, class, order, family, genus, species}\}$ ; (iii) **ImgNet**:  $\text{class} \in \{\text{animal, artifact}\}$ .

Dataset	Format	Prompt Template
CUB-200	Hierarchical	Based on taxonomy, what is the $\{\text{level}\}$ of the bird in this image? Based on the image, what is the taxonomic classification at the $\{\text{level}\}$ level?
	General	What is the taxonomic classification of the bird in this image? How can the bird in this image be categorized taxonomically? What is the systematic position of the bird shown in the image?
iNat21	Hierarchical	Based on taxonomy, where does the $\{\text{object}\}$ in the image fall in terms of $\{\text{level}\}$ ? Given the $\{\text{object}\}$ in the image, what is its taxonomic classification at the $\{\text{level}\}$ level?
	General	What could the $\{\text{object}\}$ in the image be classified as? How can the $\{\text{object}\}$ in the image be taxonomically categorized? What is the systematic position of the $\{\text{object}\}$ in the image within the biological hierarchy?
ImgNet	General	What is the taxonomic category of the $\{\text{class}\}$ in this image? How can the $\{\text{class}\}$ in this image be categorized in taxonomy? Based on classification, what type of $\{\text{class}\}$ is this? What is the hierarchical class of the $\{\text{class}\}$ shown here? Where does this $\{\text{class}\}$ belong in the taxonomic hierarchy?

Table 13: Evaluation of open-source VLLMs on hierarchical image classification benchmarks using different prompt engineering methods.

Model	Prompt	CUB-200		ImgNet-Animal		iNat21-Plant	
		HCA	Acc <sub>leaf</sub>	HCA	Acc <sub>leaf</sub>	HCA	Acc <sub>leaf</sub>
LLaVA-OV-7B [29]	General Prompts	11.44	43.44	35.58	65.45	3.88	27.24
	Hierarchy Prompts	11.24	43.94	N/A	N/A	4.54	27.45
	+ CoT [57]	10.99	42.98	35.72	65.82	4.46	27.34
	+ Taxonomy	14.45	40.25	N/A	N/A	N/A	N/A
InternVL2.5-8B [9]	General Prompts	19.81	45.58	37.20	65.12	4.98	28.11
	Hierarchy Prompts	21.25	45.30	N/A	N/A	5.61	28.28
	+ CoT [57]	21.17	45.21	36.99	65.38	5.49	28.26
	+ Taxonomy	16.19	37.88	N/A	N/A	N/A	N/A
Qwen2.5-VL-7B [4]	General Prompts	40.78	65.38	55.33	79.93	16.15	40.51
	Hierarchy Prompts	43.66	65.54	N/A	N/A	17.21	41.49
	+ CoT [57]	43.21	64.91	56.17	80.43	18.06	42.53
	+ Taxonomy	32.52	52.66	N/A	N/A	N/A	N/A

### 969 C.1.2 Chain of Thought (CoT)

970 To examine whether Chain-of-Thought reasoning improves hierarchical inference, we follow [24,  
971 57]. Concretely, we append the phrase "Let's think step by step." to the end of each ques-  
972 tion prompt followed the work [68]. The results are presented in Table 13, where no significant  
973 improvement is observed when building CoT upon the *Hierarchical Prompt*. Apart from the sim-  
974 ple "Let's think step by step." prompt, we also evaluated a biologically grounded chain-of-  
975 thought prompting strategy on the iNat21-Plant and iNat21-Animal datasets, which feature more  
976 comprehensive and standardized taxonomies. Specifically, we incorporated the biological reasoning  
977 process directly into the system prompt on iNat21-Animal as follows:

You are an expert in hierarchical image classification.\n  
 Given an image and a multiple-choice question about a specific taxonomy level (e.g., genus, family), first infer the most likely species from the image, then reason step-by-step through the taxonomy hierarchy to identify the correct label.\n  
 Respond with only the letter corresponding to the correct answer.\n  
 Example: if the image depicts *Caenurgina crassiuscula*, then the correct genus is *Caenurgina*, family is *Erebidae*, order is *Lepidoptera*, class is *Insecta*, phylum is *Arthropoda*, and kingdom is *Animalia*.\n  
 For instance, if the question is:\n Given the animal in the image, what is its taxonomic classification at the phylum level?\n  
 A. Annelida\n  
 B. Arthropoda\n  
 C. Mollusca\n  
 D. Chordata\n  
 You should select option B, labeled with Arthropoda.

978

979 The hierarchical reasoning example in the system prompt for iNat21-Plant is adapted accordingly  
 980 using a representative example from the iNat21-Plant taxonomy.

981 We report the evaluation results of Qwen2.5-VL-7B in Table 14. Notably, incorporating the biolog-  
 982 ical chain-of-thought does not yield performance improvements and even underperforms compared  
 983 to the simple chain-of-thought prompting strategy, as shown in Table 13.

Table 14: Biological chain-of-thought results on iNat21-Plant and iNat21-Animal using Qwen2.5-VL-7B.

Model	iNat21-Plant		iNat21-Animal	
	HCA	Acc <sub>leaf</sub>	HCA	Acc <sub>leaf</sub>
Qwen2.5-VL-7B [4]	15.42	40.96	15.39	40.47

### 984 C.1.3 Taxonomy as Context

985 The taxonomy is encoded as a JSON dictionary that maps each leaf node to the ordered list of  
 986 its ancestors up to the root. We provide this structure verbatim at the beginning of the prompt by  
 987 concatenating "Here's a taxonomy: " + {Taxonomy JSON} + {original prompt}. This  
 988 supplies the model with the full taxonomic context. We report results on representative open-source  
 989 VLLMs using the CUB-200 dataset in Table 13. Surprisingly, explicitly providing the taxonomy as  
 990 context to VLLMs does not improve performance; instead, it leads to a degradation in HCA. This  
 991 may be attributed to the additional taxonomy consuming a portion of the model's attention capacity,  
 992 thereby reducing the attention available for visual tokens. In addition, we include a text-only eval-  
 993 uation where each prompt is contextualized with the full taxonomy. The results are summarized in  
 994 Table 15. Notably, even when the explicit textual taxonomy is provided, the text-only HCA reaches  
 995 only 74.82%, which remains substantially below our expectations for LLMs.

Table 15: (Text) HCA of Qwen2.5-VL-7B on the CUB-200 dataset with taxonomy as context.

LLM of	HCA	POR	S-POR	TOR
Qwen2.5-VL-7B	66.26	89.94	77.08	77.44
Qwen2.5-VL-7B w/ Taxonomy	74.82	93.14	83.72	83.37

### 996 C.1.4 Question with Binary Answer

997 We also evaluate a binary question-answering format with Yes or No responses. For each original  
 998 four-choice question, we convert the four candidate answers into four separate statements. We then  
 999 perform four separate forward passes on the same image to obtain the final predictions using majority  
 1000 voting with the results from the standard prompt. The binary-format questions are formulated as  
 1001 follows:



**Statement 1:** <image> The bird in the image belongs to the <hierarchy> (e.g., Order) of <ground truth> (e.g., Passeriformes).  
Is this statement correct? Please answer Yes or No.

**Statement 2:** <image> The bird in the image belongs to the <hierarchy> (e.g., Order) of <similar class>.  
Is this statement correct? Please answer Yes or No.

**Statement 3:** <image> The bird in the image belongs to the <hierarchy> (e.g., Order) of <similar class>.  
Is this statement correct? Please answer Yes or No.

**Statement 4:** <image> The bird in the image belongs to the <hierarchy> (e.g., Order) of <similar class>.  
Is this statement correct? Please answer Yes or No.

1002 In this scenario, if none or multiple “Yes” responses appear among the four statements, we consider  
1003 the model uncertain at the current hierarchy level and mark the prediction as incorrect. A prediction  
1004 is counted as valid only when exactly one “Yes” answer is returned out of the four questions. Based  
1005 on this criterion, we assess the answers and report the results on all metrics on CUB-200 using  
1006 Qwen2.5-VL-7B in Table 16. Compared with the original four choice question answering setting,  
1007 this scenario exhibits a significant performance drop, with approximately 27% degradation in HCA  
1008 and 13% in leaf level accuracy. This result is expected, as the model no longer has access to con-  
1009 trasting choices within a single forward pass. In uncertain cases, the absence of explicit alternatives  
1010 makes it more prone to errors, whereas the four choice setting can implicitly guide the model toward  
1011 a correct selection by constraining the label space.

Table 16: Hierarchical evaluation results using binary QA format on CUB-200.

Model	HCA	Acc <sub>leaf</sub>	POR	S-POR	TOR
Qwen2.5-VL-7B [4]	16.22	51.71	63.23	41.37	42.60

## 1012 C.2 Linear Probing of Visual Features

1013 For linear probing experiments on image features, we use Qwen-2.5-VL-7B and retrieve image token  
1014 embeddings from three checkpoints in the pipeline: (i) vision encoder output, (ii) projector output,  
1015 and (iii) residual stream of the final layer of LLM. We evaluate two pooling heuristics: mean pooling  
1016 across all image tokens versus selecting the final image token, and observe that mean pooling con-  
1017 sistently outperforms the final-token alternative. Accordingly, all results in Section 3.2 are reported  
1018 with mean-pooled representations, echoing the empirical findings of Zhang et al. [68].

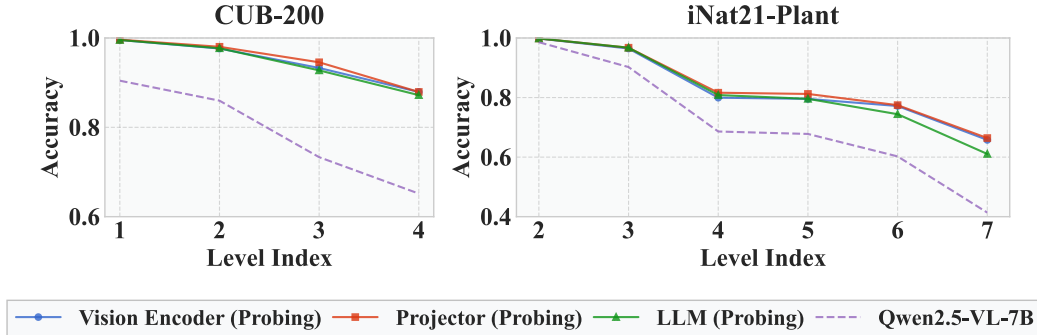


Figure 11: Level-by-level linear probing accuracy on CUB-200 [54] and iNat21-Plant [53] using Qwen2.5-VL-7B [4]. High performance obtained from features taken at the vision encoder, vision projector, and LLM shows that discriminative visual information is preserved end-to-end throughout the VLLM.

1019 We train a linear classifier on the training sets of CUB-200 and iNat21-Plant using a batch size of  
1020 512, a learning rate of 1e-4, and the Adam optimizer for 500 epochs. For CUB-200, we use the

entire training set (5994 images), while for iNat21-Plant, we randomly sample 10 images per class to ensure a balanced subset (42710 images). For testing, we use 5794 testing images from CUB-200 and 42710 images from iNat21-Plant. Furthermore, for each level in the taxonomy, we train a separate linear classifier. After training, we report the best test performance achieved during the training process. We present the level-by-level accuracy of the probing results, as shown in Figure [11](#). On the iNat21-Plant dataset, we observe that the performance gap between the VLLM and the probed components increases with taxonomy depth, indicating that VLLM struggle more at finer-grained levels. In contrast, on the CUB-200 dataset, the probed components significantly outperform the VLLM across all levels. These results demonstrate that the visual embeddings are highly effective for both hierarchical consistency and fine-grained recognition. However, performance still drops when the task involves extremely fine-grained categories such as the leaf level in iNat21-Plant, which contains 4,271 distinct classes, where even the probing model achieves only 65% accuracy.

### C.3 Text HCA on Large Qwen2.5-VLs

We also evaluated text-only hierarchical classification on the 32B and 72B variants of Qwen2.5-VL [\[4\]](#), with results presented in Table [17](#). The findings align with our observations regarding the scaling law in hierarchical classification: models with a larger number of parameters demonstrate stronger hierarchical visual understanding. However, the highest HCA across all datasets, 92.98% for Qwen2.5-VL-72B still falls short of expectations. This suggests that even with a stronger model, shallower taxonomy, and smaller dataset, the LLMs hierarchical consistency remains suboptimal. Furthermore, the consistently high Pearson correlation coefficients between text-based and visual HCAs reinforce the conclusion that the LLM component is the primary bottleneck in VLLM’s hierarchical visual understanding.

Table 17: (Text) HCA of VLLMs’ LLMs and its correlation  $\rho$  with VLLMs’ (visual) HCA on Qwen2.5-VL-32B and Qwen2.5-VL-72B.

LLM of	iNat21-Animal	iNat21-Plant	ImgNet-Artifact	ImgNet-Animal	CUB-200	$\rho(\text{text, visual})$
Qwen2.5-VL-32B <a href="#">[4]</a>	67.88	72.88	41.91	82.18	90.62	0.9517
Qwen2.5-VL-72B <a href="#">[4]</a>	83.08	87.76	41.19	84.51	92.98	0.9192

### C.4 HCA over Different Taxonomy Depth

To investigate which taxonomy levels contribute the most to performance degradation, we report the HCA across different taxonomy depths for both image-based and text-only hierarchical classification tasks using Qwen2.5-VL-7B, InternVL2.5-8B, and LLaVA-OV-7B on the iNat21-Plant dataset (Table [18](#)). For VLLMs, we recompute HCA by treating upper taxonomy levels as the leaf level. For LLMs, we re-run the experiments by substituting the original leaf-node labels with higher-level labels (e.g., replacing species-level labels at level 6 with genus-level labels at level 5).

The results show that VLLMs consistently perform better as the taxonomy depth becomes shallower, which is expected since the label space decreases. However, a notable drop in performance is observed at level 5 for all models and at level 3 for Qwen2.5-VL-7B and LLaVA-OV-7B. This suggests that these specific levels of the iNat21-Plant taxonomy may represent bottlenecks for the LLMs’ hierarchical reasoning capabilities.

### C.5 Comparison Between Vision-Tuned LLMs and Original LLMs

We present an extended comparison between vision-tuned LLMs and their original counterparts for all 7B/8B open-source VLLMs in Figure [12](#). As shown, with the exception of LLaVA-OV-7B, all other models exhibit improved performance in their vision-tuned versions on at least 3 out of the 5 benchmarks.

### C.6 Linear Probing of Text Features

To quantify the extent to which hierarchical structure is preserved in the residual stream of the LLM, we perform linear probing using text token embeddings from the residual stream (across all decoder

Table 18: HCA of different VLLMs and their LLMs over the iNat21-Plant taxonomy of various depths.

VLLM	Level 6	Level 5	Level 4	Level 3	Level 2	Level 1
Qwen2.5-VL-7B [2]	17.67	35.15	51.86	65.32	90.53	98.81
InternVL2.5-8B [9]	5.66	14.58	28.35	43.03	74.82	90.99
LLaVA-OV-7B [29]	4.62	11.83	25.22	42.30	78.40	96.12
LLM of	Level 6	Level 5	Level 4	Level 3	Level 2	Level 1
Qwen2.5-VL-7B [2]	64.22	60.06	79.78	65.99	99.86	N/A
InternVL2.5-8B [9]	41.15	38.38	65.49	83.76	99.67	N/A
LLaVA-OV-7B [29]	28.49	27.95	55.20	49.46	99.82	N/A

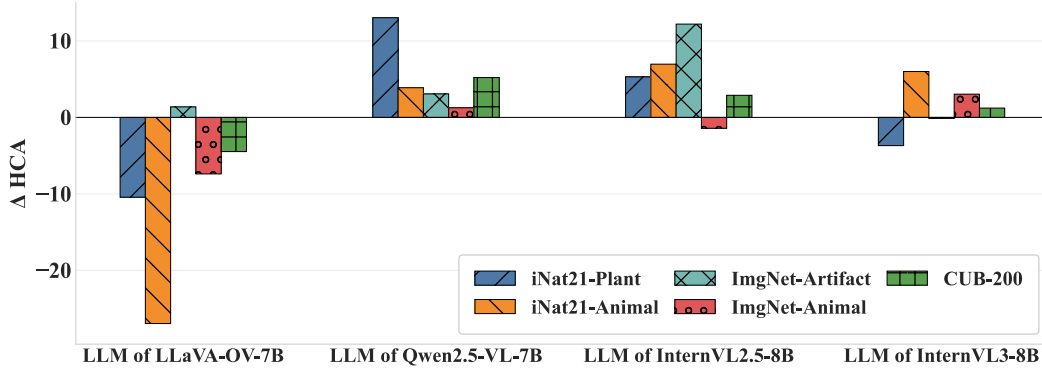


Figure 12: HCA difference between vision-tuned LLMs and their original versions across all 7B/8B open-source VLLMs. ( $\Delta HCA$  = Vision-Tuned HCA - Original HCA.)

layers) of the LLM component in Qwen2.5-VL-7B, evaluated on iNat21-Plant and CUB-200. We adopt three prompt templates (listed in Table 19) that differ in semantic framing, with Prompts 1 and 2 encoding explicit hierarchical information and Prompt 3 capturing it implicitly. Following the setup in Appendix C.2, we apply mean pooling over all text token embeddings and use the same training configuration.

For text probing, we partition the taxonomy from the leaf level using an approximate 3:2 training-to-testing split ratio. This ensures that both sets share all higher-level taxonomy nodes, allowing for a unified label space across training and testing for the linear classifier. Specifically, we use 2,508 leaf nodes for training and 1,763 for testing in iNat21-Plant, and 137 leaf nodes for training and 63 for testing in CUB-200.

Table 19: Prompt templates for text probing queries. For example, {species} = Panthera leo, {hierarchy} = genus, {label} = Panthera.

Prompt ID	Template
Prompt 1	{species} belongs to the {hierarchy} {label}.
Prompt 2	Given the {species}, what is its taxonomic classification at the {hierarchy} level? It belongs to {label}.
Prompt 3	Given the {species}, what is its taxonomic classification at the {hierarchy} level?

## C.7 Hierarchical Text Classification Results on Gemma Models

We report hierarchical text classification performance for the Gemma models [50] evaluated by Park et al. [42] on ImgNet-Animal in Table 20, including both the 2B and 7B variants, as well as their base and instruction-tuned (IT) versions. All Gemma models perform poorly on our hierarchical

benchmarks. Although the base Gemma-7B variant is the strongest among the Gemma family, it still yields the lowest text HCA compared to all other evaluated open-source VLLMs. This result suggests that even when a model exhibits perfect orthogonality in the geometric representation of hierarchical concepts, as reported in [42], it may still lack hierarchical consistency in practice.

Table 20: Hierarchical text classification performance of Gemma models on ImgNet-Animal dataset.

Model	Gemma-2B	Gemma-2B-IT	Gemma-7B	Gemma-7B-IT
HCA	1.11	16.74	39.57	29.22
POR	42.22	70.37	84.98	79.95

## D Supplementary for Section 4

### D.1 Training Data Construction

Following the format of hierarchical image classification benchmarks, we construct visual instruction tuning as a multi-turn question-answering task. Each question is a four-choice multiple-choice query, and each answer is a single letter denoting the correct choice, mirroring the style of the LLaVA instruction-tuning dataset [34]. We adopt the **iNat21-Plant training** split, which contains 4,271 species (leaf nodes). Of these, we allocate 3,771 species nodes for training and hold out 500 species nodes for out-of-domain evaluation. The hierarchy distribution of the training and testing split is depicted in Figure 13. For each leaf node, we sample 10 images from the training set, yielding 37,710 training images in total. Each image is paired with a five-turn conversation that traverses the taxonomy from the class level down to the species (leaf) level. From the unused training images we construct a *validation* split by sampling 3 images per node for *all* 4,271 species, resulting in 12,813 images. This split is used for model selection and early-stopping.

### D.2 Implementation Details

During finetuning, we freeze the parameters of both the vision encoder and the vision-language projector of Qwen2.5-VL-7B, updating only the LLM component using LoRA [22] adapters. We adopt a batch size of 128 and a learning rate of  $5 \times 10^{-5}$ , optimized with AdamW and a warm-up ratio of 0.03. The LoRA configuration consists of a rank of 64, an  $\alpha$  value of 64, and a dropout rate of 0.2. Training is performed for 1 epoch using 4 A6000 GPUs, resulting in a total of 295 steps completed within 1 hour. We report results using the model checkpoint that achieves the best performance on the validation set.

### D.3 Text-only LoRA Finetuning

Table 21: (Visual) HCA and  $\text{Acc}_{\text{leaf}}$  of Qwen2.5-VL-7B before and after the (text-only) LoRA-finetuning.

Model	iNat21-Animal		iNat21-Plant		ImgNet-Animal		CUB-200	
	HCA	$\text{Acc}_{\text{leaf}}$	HCA	$\text{Acc}_{\text{leaf}}$	HCA	$\text{Acc}_{\text{leaf}}$	HCA	$\text{Acc}_{\text{leaf}}$
Qwen2.5-VL-7B	19.43	41.33	17.67	41.61	56.00	80.01	43.76	65.50
Qwen2.5-VL-7B (LoRA)	22.54	44.71	21.81	42.22	56.67	79.85	44.43	65.15
$\Delta$	+3.11	+3.38	+4.14	+0.61	+0.67	-0.16	+0.67	-0.35

Table 22: (Text) HCA of the LLM of Qwen2.5-VL-7B before and after the (text-only) LoRA-finetuning.

Model	iNat21-Animal	iNat21-Plant	ImgNet-Animal	CUB-200
LLM of Qwen2.5-VL-7B	52.08	64.21	68.14	63.86
LLM of Qwen2.5-VL-7B (LoRA)	62.72	87.67	70.76	67.92
$\Delta$	+10.64	+23.46	+2.62	+4.06

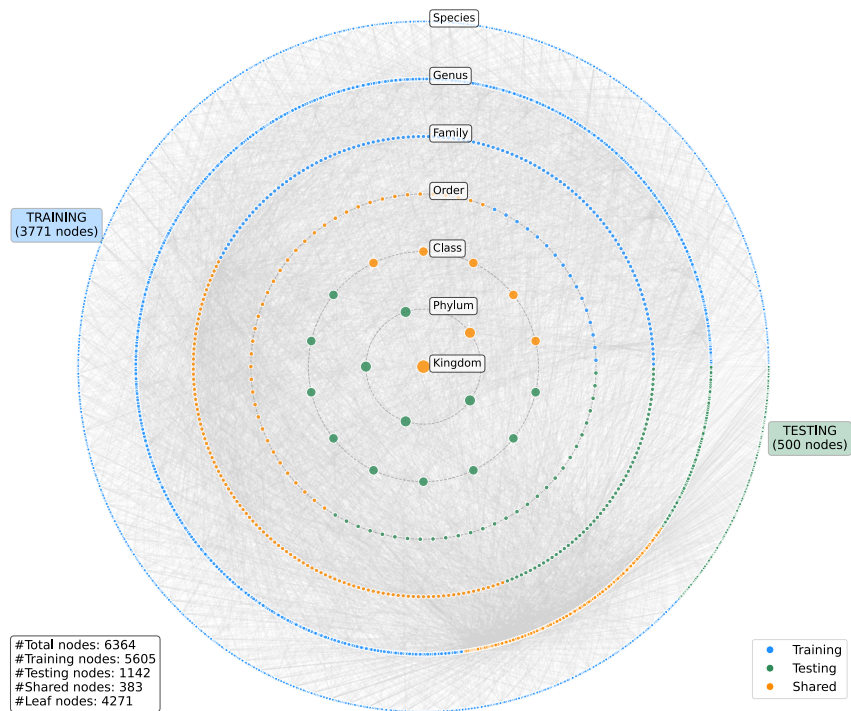


Figure 13: Hierarchy distribution of the iNat21-Plant training and testing splits.

To investigate whether finetuning the LLM with *purely* text supervision can enrich its hierarchical representations, and thereby enhance the VLLM’s hierarchical visual understanding, we create a *text-only* instruction-tuning corpus. Similar to what we did in text-only hierarchical benchmark curation, this dataset is obtained by replacing the image tokens from our visual instruction-tuning corpus by the leaf node label while preserving the multi-turn prompts and their ground-truth answers. For the text-only finetuning, we adopt the same training setup as described in Appendix D.2.

The evaluation results on hierarchical VQA benchmarks are shown in Table D.1, and the corresponding results on hierarchical text-only QA benchmarks are presented in Table D.2. As seen in Table D.1, although the improvements are modest, the model shows consistent gains in the four evaluated benchmarks, with an increase of 4.14 in HCA on iNat21-Plant and 3.11 on iNat21-Animal. This suggests that enhancing the hierarchical understanding of LLM in the language space can also benefit the hierarchical visual reasoning of VLLM, reinforcing our earlier finding that the LLM component is a key bottleneck.

For the text-only results in Table D.2, the model achieves performance that is, on average, comparable to the vision instruction-tuned model. Notably, the performance gains on iNat21-Plant and CUB-200 exceed those of the vision-tuned model (Table 5), whereas the improvements are smaller on iNat21-Animal and ImgNet-Animal.

#### D.4 Evaluation on General VQA Benchmarks

We report the evaluation results of our vision instruction-tuned model on three general VQA benchmarks: MME [16], MMBench [36], and SEED-Bench [30], as shown in Table D.3. Notably, our hierarchically enhanced VLLM demonstrates no degradation in general-purpose performance and even achieves improvements on MME and MMBench. These results suggest that our finetuned model can serve both as a specialized assistant for users interested in taxonomy and as a general-purpose VLLM for broader applications.

Table 23: Performance comparison between the original Qwen2.5-VL-7B (OG) and our (vision) LoRA-tuned variant (LoRA) on three general VLLM benchmarks.

Model	MME	MMBench	SEED-Bench
OG	2306	82.04	<b>75.95</b>
LoRA	<b>2345</b>	<b>83.25</b>	75.93

We also report results for the text instruction-tuned model in Table 24. Consistent with the vision instruction-tuned performance, we observe no loss of generalization ability. This further confirms that our hierarchical fine-tuning datasets are helpful and can be seamlessly integrated into both VLLM and LLM instruction-tuning pipelines.

Table 24: Performance comparison between the original Qwen2.5-VL-7B (OG) and our (text) LoRA-tuned variant (LoRA) on three general VLLM benchmarks.

Model	MME	MMBench	SEED-Bench
OG	2306	82.04	<b>75.95</b>
LoRA	<b>2357</b>	<b>82.39</b>	75.84

## E Limitations

While we have identified that the bottleneck in VLLM’s hierarchical visual understanding lies in the LLM component, the underlying cause of LLMs’ lack of hierarchical consistency in the language space remains an open question. Given the vast, highly structured corpora used during pre-training, one might expect stronger hierarchical representations to emerge from LLMs naturally. Unfortunately, our computational budget precludes training an LLM from scratch to verify this hypothesis. We, therefore, leave to future work the investigation of pre-training strategies that inject *explicit* hierarchical knowledge, an avenue that could clarify the underlying cause and potentially close the remaining performance gap.

Moreover, our study focused on hierarchical image classification due to limited resources. However, hierarchical visual understanding is broader, including video, 3D, and other visual modalities and more diverse taxonomies. We conjecture that state-of-the-art VLLMs would still perform poorly in those scenarios, but the causes could be different from our findings. LLMs probably would remain the weak point in those scenarios, and it is possible that the visual encoder or projector would be equally responsible.

Finally, we made a bold hypothesis that one cannot make VLLMs understand visual concepts fully hierarchical until LLMs possess corresponding taxonomy knowledge. It could overly blame LLMs, although we have supported this hypothesis with a systematic empirical investigation and the strong correlations between LLMs’ taxonomy knowledge and the corresponding VLLMs’ hierarchical visual understanding performance. Some post-training and test-time computation methods could work well without explicitly improving LLMs’ taxonomy knowledge.

## F Broader Impacts and Ethics Statement

Accurate hierarchical visual reasoning is critical in applications where coarse- and fine-grained decisions coexist-e.g., biodiversity monitoring, medical diagnostics, autonomous driving, and content moderation. Our study uncovers a systematic weakness in current VLLMs: they often predict plausible fine-grained labels while violating higher-level taxonomic structure. Deploying such models without qualification could, for example, mislead ecological surveys, propagate medical misdiagnosing, or bias downstream decision-making pipelines that rely on hierarchical consistency for error checking.

By pinpointing the LLM component as the bottleneck in VLLM’s hierarchical visual understanding and demonstrating that modest multimodal finetuning already improves textual taxonomy knowledge, our findings encourage the community to (i) incorporate explicit hierarchical objectives during LLM pre-training, (ii) curate multimodal corpora with reliable taxonomic annotations, and (iii)



1164 develop evaluation metrics that penalize hierarchical inconsistency. These steps could yield models  
 1165 that are safer and more trustworthy in real-world, hierarchy-rich settings.

1166 Potential downsides include the amplification of existing taxonomic biases or misclassifications if  
 1167 the training data encode culturally or scientifically outdated hierarchies. Researchers should there-  
 1168 fore audit hierarchies for regional or disciplinary bias, publish data-curation protocols, and, where  
 1169 feasible, provide mechanisms for community feedback and correction.

1170 Overall, we believe that exposing and remedying hierarchical blind spots in VLLMs will enable  
 1171 more reliable AI systems and support scientific, environmental, and industrial domains that depend  
 1172 on structured semantic understanding.

## 1173 G Detailed Related Works

1174 **Hierarchical classification.** Hierarchical classification [47, 25] involves assigning labels from a  
 1175 structured semantic hierarchy rather than from a flat label space lacking relational structure. In the  
 1176 vision domain, hierarchical image classification aims to improve visual consistency across coarse-to-  
 1177 fine categories, thereby enhancing overall classification performance. Recent work has introduced  
 1178 structural priors into visual models through hierarchical loss functions, multi-level supervision, and  
 1179 taxonomy-aligned embeddings [61, 43, 65, 48, 7]. Beyond the visual domain, hierarchical classifica-  
 1180 tion has also been extensively explored in the language domain [70, 56, 71]. Similar to approaches  
 1181 developed for enhancing hierarchical consistency in vision models, prior work has focused on in-  
 1182 jecting hierarchical information into language encoders to improve the structure-awareness of text  
 1183 embeddings. Another line of research aims to understand how hierarchical structures are inherently  
 1184 encoded within pre-trained language models. He et al. [21] retrained transformer-based language  
 1185 models in hyperbolic space, resulting in improved modeling of hierarchical knowledge.

1186 **Hierarchical classification with VLMs.** Existing studies have shown that CLIP models [46] strug-  
 1187 gle to maintain semantic consistency across taxonomic levels [58, 59, 40, 18]. ProTect [58] evaluated  
 1188 the CLIP model across different levels of semantic granularity and proposed a hierarchy-consistent  
 1189 prompt tuning method. HyCoCLIP [40] leveraged the inherent hierarchical nature of hyperbolic  
 1190 embeddings to improve the hierarchical structuring of CLIP representations. HGCLIP [59] further  
 1191 advanced this direction by combining CLIP with graph-based representation learning to better ex-  
 1192 ploit the hierarchical class structure. By leveraging the hierarchy information, CHiLS [38] improves  
 1193 the zero-shot classification accuracy of the CLIP model.

1194 **Classification with VLMs.** While VLMs [4, 72, 29, 9] have demonstrated strong performance  
 1195 across a wide range of tasks, their effectiveness in visual classification, particularly for fine-grained  
 1196 and subordinate-level recognition remains suboptimal [68, 65, 20, 63, 12, 63]. Zhang et al. [68] iden-  
 1197 tified the limitations of current VLLMs in classification tasks and introduced ImageWikiQA, a new  
 1198 benchmark focused on object recognition. Building on this, Liu et al. [35] evaluated a broader range  
 1199 of recent VLLMs, highlighting that models such as Qwen2-VL have achieved notable improvements  
 1200 in classification accuracy, largely due to language model advances and the use of more diverse train-  
 1201 ing data. He et al. [20] further investigated the causes of poor fine-grained classification performance,  
 1202 attributing it primarily to the absence of sufficient category names during training. To better assess  
 1203 the classification capabilities of vision-language models, Geigle et al. [17] proposed FOCI, a bench-  
 1204 mark derived from five popular classification datasets. Yu et al. [63] introduced a comprehensive  
 1205 fine-grained classification benchmark and demonstrated that the performance of VLLMs steadily de-  
 1206 clines as category granularity becomes finer. Beyond closed-set evaluation, Conti et al. [12] explored  
 1207 the open-world classification abilities of VLLMs from a broader perspective. To better evaluate the  
 1208 VLLM in an open-ended format, Snæbjarnarson et al. [49] proposed to evaluate the unconstrained  
 1209 text predictions in a taxonomy manner instead of the exact string matching. In contrast to previous  
 1210 work, we provide a more comprehensive evaluation of classification ability across different levels of  
 1211 semantic abstraction, enabling a finer analysis of hierarchical consistency in VLLMs.