The LLM Bottleneck: Why Open-Source Vision LLMs Struggle with Hierarchical Visual Recognition

Supplementary Material

In the **appendices**, we provide all implementation details to promote the reproducibility of our work, more experimental results, and further discussions. 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 discussions by including more related works, limitations, and broader impacts of the work.

A. Curation of the Hierarchical Classification Benchmarks

A.1. Hierarchical Image Classification Benchmarks

Following prior work on hierarchical image classification, we adopted several commonly used hierarchical classification datasets, including ImageNet [10], iNaturalist-2021 [41], CUB-200-2011 [42] and Food-101 [5]. Table 1 summarizes the six taxonomies and four datasets we use to construct the VQA tasks.

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 $\mathrm{Acc}_{\mathrm{leaf}}$ 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 1. To better illustrate the data format, we also provide several examples from different datasets as shown in Figure 2.

A.2. Taxonomy Refinement

To refine the taxnonomy quality, we use Wikipedia and GPT-4o to carefully examine the hierarchical relations in each taxonomy.

Re-Verify the Taxonomies: For example, in the original WordNet hierarchy, "indigo bunting" is misclassified under "finch". However, based on the established taxonomy, it belongs to the cardinal family. We corrected its hierarchical path to: animal → vertebrate → bird → oscine → cardinal → indigo bunting. We detected these errors using GPT-4o and then validated them using reliable taxonomies in Wikipedia.

Remove Ambiguous Paths: For example, in WordNet, "tusker" is assigned to the overly coarse path: animal → vertebrate → mammal → tusker. However, "tusker" is merely a colloquial term for an elephant, and WordNet already includes a more fine-grained and taxonomically accurate path for "elephant": animal → vertebrate → mammal → placental → elephant → African elephant. We removed the "tusker" node, as it lacks specificity and overlaps with the existing, more precise elephant class.

Correct Subordinate Relationships Within the Same Hierarchy Level: In the ImgNet-Artifact dataset, the hierarchy provided by WordNet exhibits notable semantic inconsistencies, particularly where concepts at the same hierarchical level implicitly reflect subordinate relationships. For instance, under the category "device", concepts such as "machine", "instrument", "musical instrument", and "mechanism" are listed as siblings. However, "musical instrument" is a subtype of "instrument", making the latter a hypernym of the former. Treating these as peers can lead to ambiguous or conflicting answers when classifying an image, as both may be considered valid even though one is semantically nested within the other. To resolve these issues, we used GPT-4o to analyze whether sibling category pairs exhibited valid hypernym-hyponym relationships systematically. We refined or removed problematic intermediate nodes and eliminated leaf nodes associated with overly coarse or semantically inconsistent hierarchy paths.

A.3. 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 3.

B. Detailed Experiment Setup and Results

B.1. Models

An overview of the models used in the evaluation experiments is provided in Table 2.

Table 1. Overview of the six taxonomies and four datasets we use to construct the VQA tasks.

Dataset	#Levels	#Leaf Nodes	#Images	Hierarchy Distribution
CUB-200 [42]	4	200	5,794	13-37-124-200
iNat21-Plant [41]	6	4,271	42.71K	5-14-85-286-1702-4271
iNat21-Animal [41]	6	5,388	53.88K	6-27-152-715-2988-5388
ImgNet-Animal [10]	11	397	19.85K	2-10-37-81-123-81-65-41-64-34-2
ImgNet-Artifact [10]	7	491	24.55K	5-40-147-204-162-62-44
Food-101 [5]	4	84	21.00K	6-29-40-24

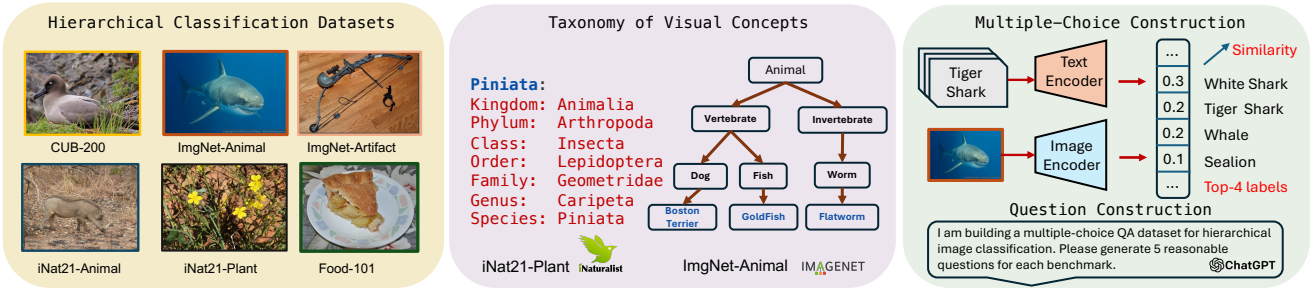

Figure 1. **Overview of hierarchical image classification benchmarks construction process.** Our hierarchical VQA benchmark is built on four datasets and covers six taxonomies. We first obtain the hierarchical structure for each taxonomy (biology standard and WordNet [27] semantics). Then, we use SigLIP [51] to generate four choices for each image based on the image-text similarities, comprising the groundtruth class name and the top three classes returned by SigLIP. Finally, we leverage GPT-4o to generate the corresponding questions.

B.2. Hierarchical Evaluation Metrics

In addition to the metrics introduced in Section 3.2, we report results on three complementary metrics that probe different aspects of hierarchical classification ability.

Point-Overlap Ratio (POR) [48]. To provide a more comprehensive evaluation of model performance across the full hierarchy, Yi et al. [48] proposed the point-overlap ratio, defined as:

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

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:

$$\text{S-POR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_i}\max_{1\le a\le b\le L_i}\left[(b-a+1)\prod_{j=a}^{b}\mathbb{1}\left[f_\theta(x_i;\mathcal{Y}_j)=y_j^i\right]\right]. \quad (2)$$

Top Overlap Ratio (TOR). Following Wu et al. [45], TOR measures *local* consistency by treating each pair of adjacent layers as an evaluation unit:

$$\mathrm{TOR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_i-1}\sum_{j=1}^{L_i-1}\mathbb{1}\left[f_\theta(x_i;\mathcal{Y}_j)=y_j^i\right]\mathbb{1}\left[f_\theta(x_i;\mathcal{Y}_{j+1})=y_{j+1}^i\right]. \quad (3)$$

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

B.3. Evaluation Results with All Metrics

We report more comprehensive evaluation results in Table 3 and Table 4. From these tables, we observe that VLLMs

[1]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.

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."

Based on the image, what is the taxonomic classification at the order level?
A. Anseriforme
B. Pelecaniforme
C. Procellariiformes
D. Podicipediformes
Answer with the option's letter from the given choices directly.

How can the bird in this image be categorized taxonomically?
A. Pomarine jaeger
B. Black-footed albatross
C. Laysan albatross
D. Sooty albatross
Answer with the option's letter from the given choices directly.

Given the plant in the image, what is its taxonomic classification at the order level?
A. Gentianales
B. Apiales
C. Cornales
D. Dipsacales
Answer with the option's letter from the given choices directly.

What is the systematic position of the plant in the image in the biological classification hierarchy?
A. Bailey
B. Draba
C. Erysimum
D. Barbarea
Answer with the option's letter from the given choices directly.

Figure 2. Examples of the prompt formats used in our four-choice hierarchical VQA tasks.

System Prompt: "You are a helpful assistant."

Black-footed Albatross

Given the bird in the image, what is its taxonomic classification at the order level?
A. Anseriforme
B. Pelecaniforme
C. Procellariiformes
D. Podicipediformes
Answer with the option's letter from the given choices directly.

Text-Only

Given the **Black-footed Albatross**, what is its taxonomic classification at the order level?
A. Anseriforme
B. Pelecaniforme
C. Procellariiformes
D. Podicipediformes
Answer with the option's letter from the given choices directly.

Figure 3. An example of the text QA construction from the hierarchical VQA.

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.

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.

Table 2. Models used in evaluation experiments and their sources.

Model	Source (Huggingface Repos and API Platform)
LLaVA-OV-7B [22]	`lmms-lab/llava-onevision-qwen2-7b-ov`
LLaVA-OV-1.5-8B [2]	`lmms-lab/LLaVA-OneVision-1.5-8B-Instruct`
InternVL2.5-8B [7]	`OpenGVLab/InternVL2_5-8B`
InternVL3-8B [55]	`OpenGVLab/InternVL3-8B`
Qwen2.5-VL-7B [4]	`Qwen/Qwen2.5-VL-7B-Instruct`
Qwen2.5-VL-32B [4]	`Qwen/Qwen2.5-VL-32B-Instruct`
Qwen2.5-VL-72B [4]	`Qwen/Qwen2.5-VL-72B-Instruct`
Qwen3-VL-8B [3]	`Qwen/Qwen3-VL-8B-Instruct`
Qwen3-VL-32B [3]	`Qwen/Qwen3-VL-32B-Instruct`
Qwen3-Omni-30B-A3B [47]	`Qwen/Qwen3-Omni-30B-A3B-Instruct`
GPT-4o[1] [31]	`OpenAI API`
OpenCLIP [8]	`laion/CLIP-ViT-L-14-laion2B-s32B-b82K`
SigLIP [51]	`google/siglip-so400m-patch14-384`
BioCLIP [39]	`imageomics/bioclip`
BioCLIP2 [14]	`imageomics/bioclip-2`

Table 3. 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
Open-Source VLLMs												
LLaVA-OV-7B	58.46	42.06	35.01	83.56	70.56	72.36	63.36	26.33	44.74	55.50	43.08	37.09
LLaVA-OV-1.5-8B	69.26	55.64	50.74	87.22	80.77	79.55	65.26	29.72	46.48	60.70	49.76	43.68
InternVL2.5-8B	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	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	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	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	89.79	86.20	83.63	92.43	88.74	87.77	67.21	30.86	48.10	80.23	70.92	69.86
Qwen3-VL-8B	80.25	72.99	68.35	90.69	85.04	84.50	64.29	27.42	45.50	72.96	63.06	60.15
Qwen3-VL-32B	86.49	82.32	78.55	91.30	87.19	86.09	68.31	37.30	49.71	72.08	62.27	59.06
Qwen3-Omni-30B-A3B	84.41	79.57	75.26	92.52	88.55	87.78	65.80	29.46	46.75	75.23	65.09	62.93
CLIP Models												
OpenCLIP	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	66.56	45.84	41.79	78.95	48.46	59.24	50.90	16.15	24.77	34.67	16.97	15.00
BioCLIP	83.99	65.63	71.82	55.36	25.01	28.29	29.03	9.13	8.73	69.80	31.13	47.62
BioCLIP2	86.27	64.31	75.52	64.16	31.28	39.75	34.03	7.57	11.12	84.34	54.86	68.64
Proprietary VLLMs												
GPT-4o	94.46	92.29	91.00	93.33	89.16	88.83	70.45	40.42	51.47	79.92	69.37	68.62

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 [9] 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 1 of the main paper, models with higher POR, S POR, and TOR scores tend to exhibit better hierarchical consistency.

B.4. Illustrative Mistakes Made by VLLMs

We visualize some hierarchical prediction errors made by open-source VLLMs in Figure 4.

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

As shown in Table 5, using random choices significantly improves the model's fine-grained accuracy-reaching up to

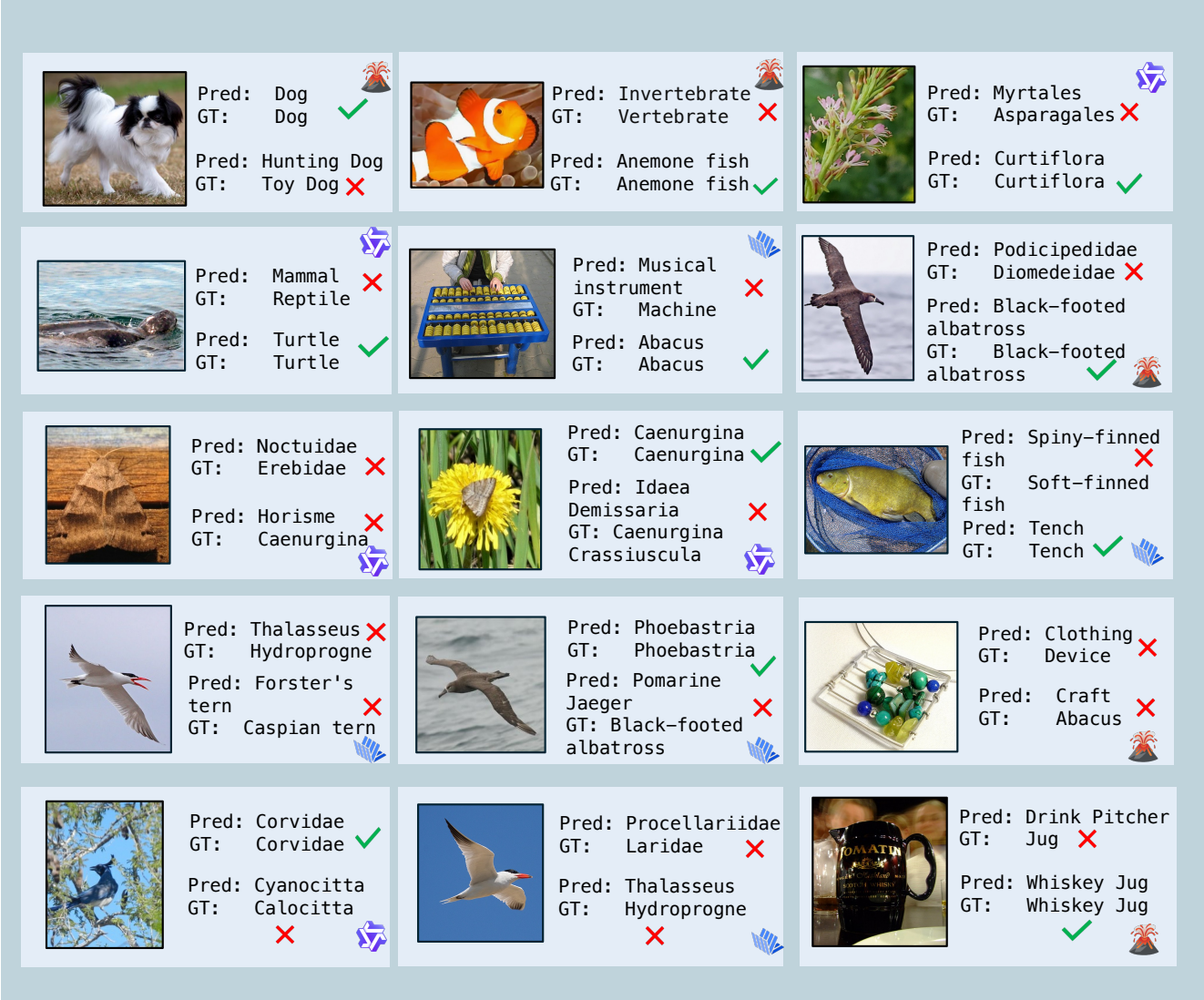

Figure 4. **Error Examples of the hierarchical predictions of VLLMs.** Examples are drawn from different VLLMs (Qwen2.5-VL-7B, InternVL2.5-8B and LLaVA-OV-7B) to reflect the diverse error modes observed across taxonomic levels.

90% for Qwen2.5-VL. However, even with random choices, the gap between Acc_{leaf} 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_{leaf}. 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 6), where no answer choices are provided. In this setting, model performance drops significantly, particularly on the iNat21-Plant benchmark, where the model struggles to generate correct answers. This results in very low fine-grained accuracy and HCA.

B.7. Food-101 Results

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

Table 4. Evaluation results across all VLMs on iNat21-Animal with all metrics reported.

Model	Acc_{leaf}	HCA	POR	S-POR	TOR
Open-source VLLMs					
LLaVA-OV-7B	26.47	4.53	60.31	45.96	45.53
LLaVA-OV-1.5-8B	27.08	8.42	66.35	58.71	53.57
InternVL2.5-8B	27.65	8.52	66.26	57.07	53.50
InternVL3-8B	35.40	11.93	69.00	59.13	55.55
Qwen2.5-VL-7B	41.66	19.73	74.80	66.92	63.71
Qwen2.5-VL-32B	46.98	26.90	78.38	72.09	68.93
Qwen2.5-VL-72B	54.20	35.73	81.76	76.05	73.55
Qwen3-VL-8B	39.28	19.54	75.39	69.08	65.23
Qwen3-VL-32B	51.74	32.44	80.69	74.70	72.08
Qwen3-Omni-30B-A3B	46.93	27.25	78.76	72.50	69.58
CLIP Models					
OpenCLIP	23.53	1.04	41.11	19.02	21.12
SigLIP	12.71	2.15	38.24	38.24	33.95
BioCLIP	88.13	17.61	72.46	29.10	55.77
BioCLIP2	95.94	41.84	85.89	56.13	73.34
Proprietary VLLM					
GPT-4o	63.79	42.95	84.25	77.74	76.15

B.8. HierarCaps Results

We further experiment with HierarCaps [1] under this work's settings, and the results are presented in Table 8. Both CLIPs and VLLMs have poor hierarchical consistency and are especially worse on abstract and shorter captions at higher levels of the hierarchy (i.e., the first two levels). This is likely due to the noisy nature of the top-level labels in the HierarCaps dataset, where each image can be associated with multiple coarse-grained captions. For example, the image with the caption "A table with three plates with food and a man and a woman both holding utensil near plates" can reasonably be classified under both "persons" and "table" categories at the first layer of the hierarchy. We also compute the HCA for the last two levels (Level 3 and Level 4), which contain more concrete and longer captions. However, there remains a noticeable gap between the HCA and the leaf node accuracy (Level 4). In many cases, the model can correctly classify at level 4 while still struggling at level 3, suggesting that the model does not fully understand the hierarchical relationship across longer textual contexts.

B.9. Performance Analysis across Different Datasets

Among all datasets, CUB-200 exhibits the smallest gap between HCA and leaf-level accuracy, which can be attributed to its shallow hierarchy (only four levels vs. six levels in iNat21-Plant and iNat21-Animal), making the task relatively simple compared to other datasets. ImgNet-Animal and ImgNet-Artifact have the deepest hierarchies. However, their leaf nodes generally correspond to basic-level concepts, which makes the leaf-level classification easier for VLLMs, resulting in high leaf accuracy. ImgNet-Artifact is the most challenging dataset on which all models yield low hierarchical consistency scores, probably for two main reasons. (1) Unlike the well-defined biological hierarchies in iNat-21 and ImgNet-Animal, the intermediate nodes in WordNet, which were used to construct ImageNet, are relatively abstract and vague, making hierarchical discrimination difficult. (2) Unlike the animal images, images in the ImgNet-Artifact dataset often contain multiple objects of different classes. When queried about higher-level categories, the model may mistakenly associate the question with another non-central object in the scene.

B.10. Could the Poor Hierarchical Consistency Originate from the Four-choice VQA Format?

In general, VQA benchmarks [21, 26] adopt a multiple-choice question format, with four-choice questions comprising the majority. Current open source VLLMs [4, 7, 22, 55] have already demonstrated strong performance on these general VQA benchmarks. Therefore, the poor performance observed in our setting is unlikely to be caused by the question format or prompt design, but rather by the limitations of the VLLMs themselves. A more comprehensive analysis of the effects of prompt design and question formats on the hierarchical understanding of VLLMs is provided in Appendix C.

C. Supplementary Materials for Section 4 in the Main Paper

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 9), comprising both hierarchy-aware (Hierarchical) and general formulations (General). For CUB-200, 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 10, averaging performance separately over general (General Prompts) and hierarchy-aware prompts (Hierarchy Prompts). Overall, hierarchy-aware prompts outperform general prompts on CUB-200 and iNat21-Plant.

C.1.2. Chain of Thought (CoT)

To examine whether Chain-of-Thought reasoning improves hierarchical inference, we follow [19, 44]. Concretely, we

Table 5. Hierarchical evaluation results (image) on CUB-200 and iNat21-Plant benchmarks with random choices.

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

Table 6. HCA and leaf-level accuracy Acc_{leaf} 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_{leaf}	HCA	Acc_{leaf}	HCA	Acc_{leaf}	HCA	Acc_{leaf}	HCA	Acc_{leaf}
Qwen2.5-VL-7B	0.01	8.33	0.11	3.88	N/A	7.43	N/A	15.46	9.39	22.02

Table 7. Evaluation results across all VLMs on Food-101 with all metrics reported.

Model	Acc_{leaf}	HCA	POR	S-POR	TOR
Open-source VLLMs					
LLaVA-OV-7B	88.80	46.45	77.30	57.70	57.06
LLaVA-OV-1.5-8B	87.86	43.42	75.60	55.20	53.82
InternVL2.5-8B	84.76	41.18	72.91	52.77	51.85
InternVL3-8B	84.88	37.95	71.26	49.11	48.96
Qwen2.5-VL-7B	90.51	43.11	75.15	53.48	53.13
Qwen3-VL-8B	91.27	51.76	79.57	60.61	60.95
Qwen2.5-VL-32B	89.13	47.32	76.99	56.83	57.93
Qwen2.5-VL-72B	92.02	52.00	80.46	60.95	62.33
Qwen3-VL-8B	91.27	51.76	79.57	60.61	60.95
Qwen3-VL-32B	89.92	45.53	77.12	55.89	56.73
Qwen3-Omni-30B-A3B	94.71	46.81	78.10	55.38	58.22
CLIP Models					
OpenCLIP	93.89	37.53	72.73	49.76	44.83
SigLIP	97.17	42.49	73.68	50.03	51.69
BioCLIP	27.82	2.48	26.29	10.81	7.25
BioCLIP2	75.89	14.10	45.19	20.81	18.39
Proprietary VLLM					
GPT-4o	95.67	55.60	82.97	63.03	66.79

append the phrase "`Let's think step by step.`" to the end of each question prompt followed the work [52]. The results are presented in Table 10, where no significant improvement is observed when building CoT upon the *Hierarchical Prompt*. Apart from the simple "`Let's think step by step.`"prompt, we also evaluated a biologically grounded chain-of-thought prompting strategy on the iNat21-Plant and iNat21-Animal datasets, which feature more comprehensive and standardized taxonomies. Specifically, we incorporated the biological reasoning 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.
```

The hierarchical reasoning example in the system

Table 8. Hierarchical evaluation results on HierarCaps (Level Accuracy and HCA).

Model	Level 4	Level 3	Level 2	Level 1	HCA (All)	HCA (Last two levels)
OpenCLIP	68.80	50.70	28.40	17.10	5.70	45.30
SigLIP	72.90	52.50	26.80	15.50	5.70	49.20
LLaVA-OV-7B	78.60	60.70	32.60	22.20	9.10	55.60
Qwen2.5-VL-7B	77.10	58.40	30.50	21.10	7.10	54.00

Table 9. Prompt templates used across datasets. Placeholders: (i) **CUB-200**: `level` ∈ {`order`, `family`, `genus`, `species`}; (ii) **iNat21**: `object` ∈ {`animal`, `plant`}, `level` ∈ {`kingdom`, `phylum`, `class`, `order`, `family`, `genus`, `species`}; (iii) **ImgNet**: `class` ∈ {`animal`, `artifact`}.

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

prompt for iNat21-Plant is adapted accordingly using a representative example from the iNat21-Plant taxonomy.

We report the evaluation results of Qwen2.5-VL-7B in Table 11. Notably, incorporating the biological chain-of-thought does not yield performance improvements and even underperforms compared to the simple chain-of-thought prompting strategy, as shown in Table 10.

C.1.3. Taxonomy as Context

The taxonomy is encoded as a JSON dictionary that maps each leaf node to the ordered list of its ancestors up to the root. We provide this structure verbatim at the beginning of the prompt by concatenating "`Here's a taxonomy: `" + {`Taxonomy JSON`} + {`original prompt`}. This supplies the model with the full taxonomic context. We report results on representative open-source VLLMs using the CUB-200 dataset in Table 10. Surprisingly, explicitly providing the taxonomy as context to VLLMs does not improve performance; instead, it leads to a degradation in HCA. This may be attributed to the additional taxonomy consuming a portion of the model's attention capacity, thereby reducing the attention available for visual tokens. In addition, we include a text-only evaluation where each prompt is contextualized with the full taxonomy. The results are summarized in Table 12. Notably, even when the explicit textual taxonomy is provided, the text-only HCA reaches only 74.82%, which remains substantially below our expectations for LLMs.

Table 12. (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

C.1.4. Few-shot VQA

We conducted additional experiments on the CUB-200 dataset using Qwen2.5-VL-7B with few-shot prompting ranging from 1 to 5 shots. We use level-specific QA pairs as few-shot examples to evaluate performance at each hier-

Table 10. 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_{leaf}	HCA	Acc_{leaf}	HCA	Acc_{leaf}
LLaVA-OV-7B	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 [44]	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	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 [44]	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	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 [44]	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

Table 11. Biological chain-of-thought results on iNat21-Plant and iNat21-Animal using Qwen2.5-VL-7B.

Model	iNat21-Plant		iNat21-Animal	
	HCA	Acc_{leaf}	HCA	Acc_{leaf}
Qwen2.5-VL-7B	15.42	40.96	15.39	40.47

archy level. An example is provided as follows:

```
Based on taxonomy, where does the <leaf label>
(e.g., Black-footed Albatross)
fall in terms of <level> (e.g., order)?

A. <Ground Truth> (e.g., procellariiformes)
B. <Similar Choice> (e.g., apodiformes)
C. <Similar Choice> (e.g., podicipediformes)
D. <Similar Choice> (e.g., pelecaniformes)

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

Results are reported in Table 13. Interestingly, we observe no performance improvement across different numbers of few-shot examples.

Table 13. HCA performance across different few-shot settings.

#Few-shot	0	1	2	3	4	5
HCA	66.26	64.83	65.57	65.50	66.10	65.46

C.1.5. Questions with Binary Answer

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

In this scenario, if none or multiple "Yes" responses appear among the four statements, we consider the model uncertain at the current hierarchy level and mark the prediction as incorrect. A prediction is counted as valid only when exactly one "Yes" answer is returned out of the four questions. Based on this criterion, we assess the answers and report the

results on all metrics on CUB-200 using Qwen2.5-VL-7B in Table 14. Compared with the original four choice question answering setting, this scenario exhibits a significant performance drop, with approximately 27% degradation in HCA and 13% in leaf level accuracy. This result is expected, as the model no longer has access to contrasting choices within a single forward pass. In uncertain cases, the absence of explicit alternatives makes it more prone to errors, whereas the four choice setting can implicitly guide the model toward a correct selection by constraining the label space.

Table 14. Hierarchical evaluation results using binary QA format on CUB-200.

Model	HCA	Acc_{leaf}	POR	S-POR	TOR
Qwen2.5-VL-7B	16.22	51.71	63.23	41.37	42.60

C.2. Linear Probing of Visual Features

For linear probing experiments on image features, we use Qwen-2.5VL-7B and retrieve image token embeddings from three checkpoints in the pipeline: (i) vision encoder output, (ii) projector output, and (iii) residual stream of the final layer of LLM. We evaluate two pooling heuristics: mean pooling across all image tokens versus selecting the final image token, and observe that mean pooling consistently outperforms the final-token alternative. Accordingly, all results in Section 4.2 are reported with mean-pooled representations, echoing the empirical findings of Zhang et al. [52].

We train a linear classifier on the training sets of CUB-200 and iNat21-Plant using a batch size of 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 5. 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 15. 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 LLM's 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.

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 16). 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 6. As shown, with the exception of LLaVA-OV-7B and InternVL3-8B, all other models exhibit improved performance in their vision-tuned versions on at least 4 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 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 17) that

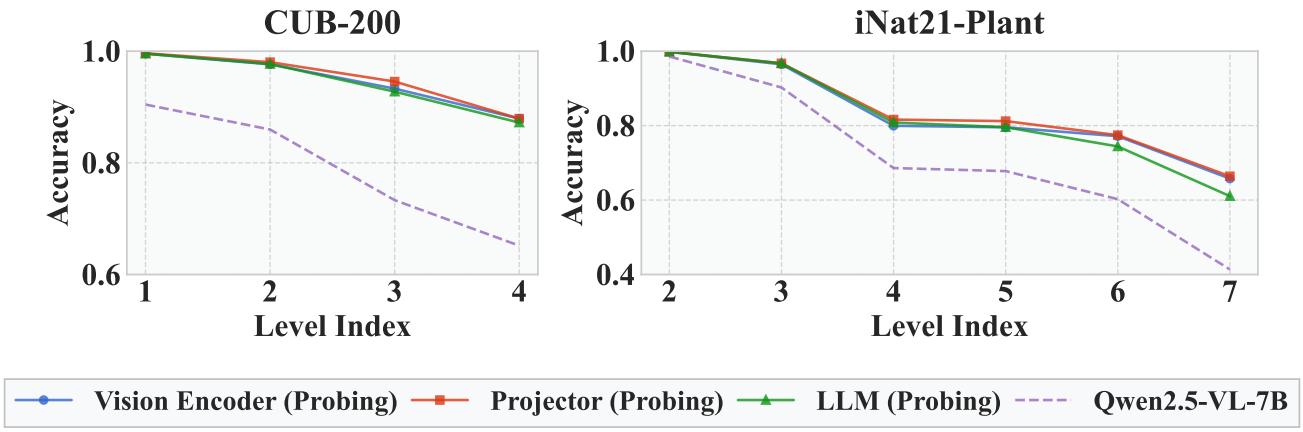

Figure 5. Level-by-level linear probing accuracy on CUB-200 [42] and iNat21-Plant [41] using Qwen2.5-VL-7B [4]. High performance obtained from features taken at the vision encoder, vision projector, and LLM shows that discriminative visual information is preserved end-to-end throughout the VLLM.

Table 15. (Text) HCA of VLLMs' LLMs and its correlation ρ 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	ρ(text,visual)
Qwen2.5-VL-32B	67.88	72.88	41.91	82.18	90.62	0.9517
Qwen2.5-VL-72B	83.08	87.76	41.19	84.51	92.98	0.9192

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.

C.7. Hierarchical Text Classification Results of Gemma Models

We report hierarchical text classification performance for the Gemma models [40] evaluated by Park et al. [33] on ImgNet-Animal in Table 18, 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 [33], it may still lack hierarchical consistency in practice.

D. Supplementary Materials for Section 5 in the Main Paper

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 [24]. 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 7. 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.

Table 16. 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	17.67	35.15	51.86	65.32	90.53	98.81
InternVL2.5-8B	5.66	14.58	28.35	43.03	74.82	90.99
LLaVA-OV-7B	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	64.22	60.06	79.78	65.99	99.86	N/A
InternVL2.5-8B	41.15	38.38	65.49	83.76	99.67	N/A
LLaVA-OV-7B	28.49	27.95	55.20	49.46	99.82	N/A

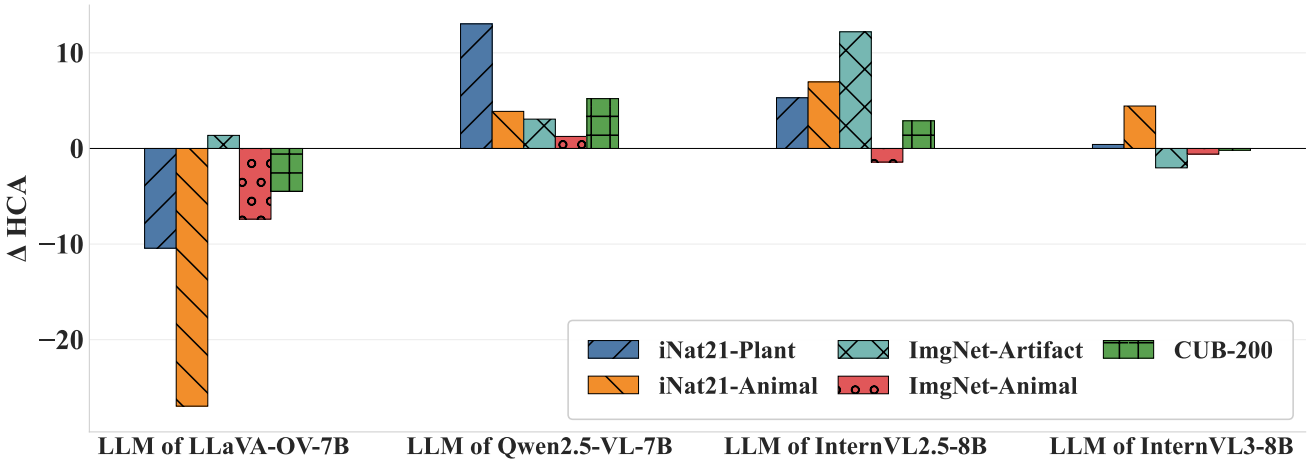

Figure 6. HCA difference between vision-tuned LLMs and their original versions across all 7B/8B open-source VLLMs. (ΔHCA = Vision-Tuned HCA - Original HCA.)

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 [18] adapters. We adopt a batch size of 128 and a learning rate of 5×10^{-5}, optimized with AdamW and a warm-up ratio of 0.03. The LoRA configuration consists of a rank of 64, an α 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

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 hierachical 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 19, and the corresponding results on hierarchical text-only QA benchmarks are presented in Table 20. As seen in Table 19, 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 20, 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 4 of the main paper), whereas the improvements are smaller on iNat21-Animal and ImgNet-Animal.

Table 17. 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?`

Table 18. 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.4. Evaluation on General VQA Benchmarks

We report the evaluation results of our vision instruction-tuned model on three general VQA benchmarks: MME [11], MMBench [26], and SEED-Bench [21], as shown in Table 21. 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 21. 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	**75.95**
LoRA	**2345**	**83.25**	75.93

We also report results for the text instruction-tuned model in Table 22. 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 22. 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	**75.95**
LoRA	**2357**	**82.39**	75.84

E. Detailed Discussion of Related Works

Hierarchical classification. Hierarchical classification [20, 36] involves assigning labels from a structured semantic hierarchy rather than from a flat label space lacking relational structure. In the vision domain, hierarchical image classification aims to improve visual consistency across coarse-to-fine categories, thereby enhancing overall classification performance. Recent work has introduced structural priors into visual models through hierarchical loss functions, multi-level supervision, and taxonomy-aligned embeddings [6, 34, 37, 48, 50]. Beyond the visual domain, hierarchical classification has also been extensively explored in the language domain [43, 53, 54]. Similar to approaches developed for enhancing hierarchical consistency in vision models, prior work has focused on injecting hierarchical information into language encoders to improve the structure-awareness of text embeddings. He et al. [17] retrained transformer-based language models in hyperbolic space, resulting in improved modeling of hierarchical knowledge. Another line of research aims to understand how hierarchical structures are inherently encoded within pre-trained language models. Nikishina et al. [29] provide a comprehensive analysis of transformer-based models for the task of hypernymy prediction, evaluating their ability to infer IS-A relations. Lin and Ng [23] study whether pre-trained BERT models capture the transitivity of IS-A relations in WordNet. To more rigorously assess such capabilities, He et al. [16] propose ONTOLAMA, a benchmark and evaluation framework targeting subsumption inference within ontologies. Similarly, Moskvoretskii et al. [28] evaluate the WordNet-based lexical-semantic reasoning ability of the LLaMA-2-7B model through the TaxoLLaMA framework.

Hierarchical classification with VLMs. Existing studies have shown that CLIP models [35] struggle to maintain

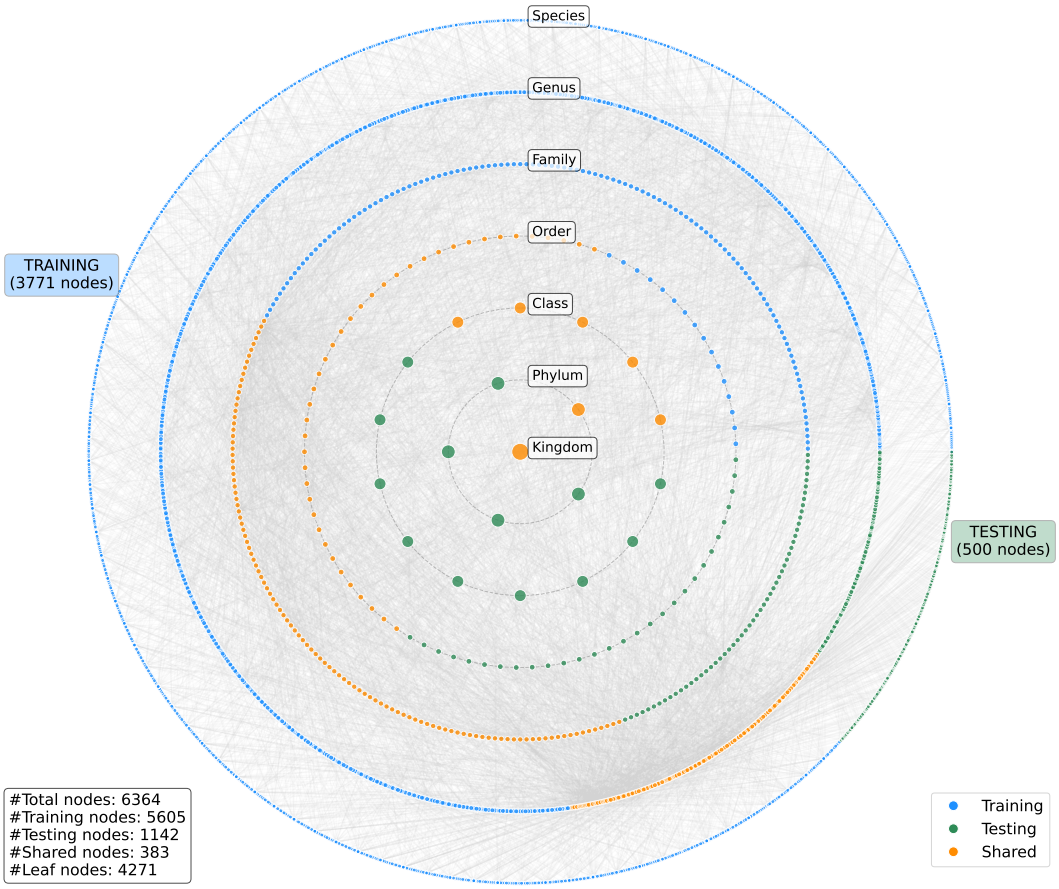

Figure 7. Hierarchy distribution of the iNat21-Plant training and testing splits.

Table 19. (Visual) HCA and Acc_{leaf} of Qwen2.5-VL-7B before and after the (text-only) LoRA-finetuning.

Model	iNat21-Animal		iNat21-Plant		ImgNet-Animal		CUB-200	
	HCA	Acc_{leaf}	HCA	Acc_{leaf}	HCA	Acc_{leaf}	HCA	Acc_{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
Δ	+3.11	+3.38	+4.14	+0.61	+0.67	-0.16	+0.67	-0.35

semantic consistency across taxonomic levels [13, 32, 45, 46]. ProTect [45] evaluated the CLIP model across different levels of semantic granularity and proposed a hierarchy-consistent prompt tuning method. HyCoCLIP [32] leveraged the inherent hierarchical nature of hyperbolic embeddings to improve the hierarchical structuring of CLIP representations. HGCLIP [46] further advanced this direction by combining CLIP with graph-based representation learning to better exploit the hierarchical class structure. By leveraging the hierarchy information, CHiLS [30] improves the zero-shot classification accuracy of the CLIP model.

Classification with VLMs. While VLMs [4, 7, 22, 55] have demonstrated strong performance across a wide range of tasks, their effectiveness in visual classification, particularly for fine-grained and subordinate-level recognition—remains suboptimal [9, 15, 25, 49, 49, 52]. Zhang et al. [52] identified the limitations of current VLLMs in classification tasks and introduced ImageWikiQA, a new benchmark focused on object recognition. Building on this, Liu et al. [25] evaluated a broader range of recent VLLMs,

Table 20. (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
Δ	+10.64	+23.46	+2.62	+4.06

highlighting that models such as Qwen2-VL have achieved notable improvements in classification accuracy, largely due to language model advances and the use of more diverse training data. He et al. [15] further investigated the causes of poor fine-grained classification performance, attributing it primarily to the absence of sufficient category names during training. To better assess the classification capabilities of vision-language models, Geigle et al. [12] proposed FOCI, a benchmark derived from five popular classification datasets. Yu et al. [49] introduced a comprehensive fine-grained classification benchmark and demonstrated that the performance of VLLMs steadily declines as category granularity becomes finer. Beyond closed-set evaluation, Conti et al. [9] explored the open-world classification abilities of VLLMs from a broader perspective. To better evaluate the VLLM in an open-ended format,Snæbjarnarson et al. [38] proposed to evaluate the unconstrained text predictions in a taxonomy manner instead of the exact string matching. In contrast to previous work, we provide a more comprehensive evaluation of classification ability across different levels of semantic abstraction, enabling a finer analysis of hierarchical consistency in VLLMs.

F. Limitations

While we have identified that the bottleneck in VLLM's hierarchical visual recognition 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 recognition 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 recognition performance. Some post-training and test-time computation methods could work well without explicitly improving LLMs' taxonomy knowledge.

G. Broader Impacts

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

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

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

References

[1] Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations. In *European Conference on Computer Vision*, pages 220–238. Springer, 2024. 6

[2] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 4

[3] 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*, 2025. 4

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 6, 10, 11, 14

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 1, 2

[6] Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4858–4867, 2022. 13

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4, 6, 14

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023. 4

[9] Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers. *arXiv preprint arXiv:2503.21851*, 2025. 4, 14, 15

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

[11] 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. *arXiv preprint arXiv:2306.13394*, 2023. 13

[12] Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. *arXiv preprint arXiv:2406.14496*, 2024. 15

[13] 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. 14

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

[15] Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. *arXiv preprint arXiv:2501.15140*, 2025. 14, 15

[16] Yuan He, Jiaoyan Chen, Ernesto Jimenez-Ruiz, Hang Dong, and Ian Horrocks. Language model analysis for ontology subsumption inference. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3439–3453, 2023. 13

[17] Yuan He, Moy Yuan, Jiaoyan Chen, and Ian Horrocks. Language models as hierarchy encoders. *Advances in Neural Information Processing Systems*, 37:14690–14711, 2024. 13

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 12

[19] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 6

[20] Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29:820–865, 2015. 13

[21] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 6, 13

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. 4, 6, 14

[23] Ruixi Lin and Hwee Tou Ng. Does bert know that the is-a relation is transitive? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 94–99, 2022. 13

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 11

[25] Huan Liu, Lingyu Xiao, Jiangjiang Liu, Xiaofan Li, Ze Feng, Sen Yang, and Jingdong Wang. Revisiting mllms:

An in-depth analysis of image classification abilities. *arXiv preprint arXiv:2412.16418*, 2024. 14

[26] 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*, pages 216–233. Springer, 2024. 6, 13

[27] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[28] Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. Taxollama: Wordnet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2331–2350, 2024. 13

[29] Irina Nikishina, Polina Chernomorchenko, Anastasiia Demidova, Alexander Panchenko, and Chris Biemann. Predicting terms in is-a relations with pre-trained transformers. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 134–148, 2023. 13

[30] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 14

[31] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4

[32] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*, 2024. 14

[33] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*, 2024. 11

[34] Seulki Park, Youren Zhang, X Yu Stella, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification. In *The Thirteenth International Conference on Learning Representations*, 2025. 13

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 13

[36] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72, 2011. 13

[37] Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. *Advances in Neural Information Processing Systems*, 37:91220–91259, 2024. 13

[38] Vésteinn Snæbjarnarson, Kevin Du, Niklas Stoehr, Serge Belongie, Ryan Cotterell, Nico Lang, and Stella Frank. Taxonomy-aware evaluation of vision-language models. *arXiv preprint arXiv:2504.05457*, 2025. 15

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

[40] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 11

[41] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Computer Vision and Pattern Recognition*, 2021. 1, 2, 11

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2, 11

[43] Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. *arXiv preprint arXiv:2203.03825*, 2022. 13

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 6, 9

[45] Tz-Ying Wu, Chih-Hui Ho, and Nuno Vasconcelos. Protect: Prompt tuning for taxonomic open set classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16531–16540, 2024. 2, 14

[46] Peng Xia, Xingtong Yu, Ming Hu, Lie Ju, Zhiyong Wang, Peibo Duan, and Zongyuan Ge. Hgclip: exploring vision-language models with graph representations for hierarchical understanding. *arXiv preprint arXiv:2311.14064*, 2023. 14

[47] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 4

[48] Kai Yi, Xiaoqian Shen, Yunhao Gou, and Mohamed Elhoseiny. Exploring hierarchical graph representation for large-scale zero-shot image classification. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022. 2, 13

[49] Hong-Tao Yu, Xiu-Shen Wei, Yuxin Peng, and Serge Belongie. Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation. *arXiv preprint arXiv:2504.14988*, 2025. 14, 15

[50] Siqi Zeng, Sixian Du, Makoto Yamada, and Han Zhao. Learning structured representations by embedding class hierarchy with fast optimal transport. *arXiv preprint arXiv:2410.03052*, 2024. 13

[51] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2, 4

[52] 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 classi-

fication? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7, 10, 14

[53] Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1106–1117, 2020. 13

[54] Juncheng Zhou, Lijuan Zhang, Yachen He, Rongli Fan, Lei Zhang, and Jian Wan. A novel negative sample generation method for contrastive learning in hierarchical text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5645–5655, 2025. 13

[55] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 4, 6, 14