# Vision LLMs Are Bad at Hierarchical Visual Understanding, and LLMs Are the Bottleneck

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

This paper reveals that many state-of-the-art large language models (LLMs) lack hierarchical knowledge about our visual world, unaware of even well-established biology taxonomies. This shortcoming makes LLMs a bottleneck for vision LLMs' hierarchical visual understanding (e.g., recognizing `Anemone Fish` but not `Vertebrate`). We arrive at these findings using about one million four-choice visual question answering (VQA) tasks constructed from six taxonomies and four image datasets. Interestingly, finetuning a vision LLM using our VQA tasks reaffirms LLMs' bottleneck effect to some extent because the VQA tasks improve the LLM's hierarchical consistency more than the vision LLM's. We conjecture that one cannot make vision LLMs understand visual concepts fully hierarchical until LLMs possess corresponding taxonomy knowledge.

## 1 Introduction

Taxonomy is natural and core in visual understanding. The biology taxonomies cover many objects in our visual world [53]; for example, a `Boston Terrier` belongs to the class of `Terrier`, which is a subtype of `Dog`, under `Mammal`, and ultimately part of the broader category `Animal`, forming a semantic path in the animal taxonomy: `Animal` → `Mammal` → `Dog` → `Terrier` → `Boston Terrier`. ImageNet [13] expands from the WordNet [37] taxonomy. Visual parts [28, 15, 3], attributes [14, 27, 41], and relationships [26] can be grouped hierarchically due to shared characteristics.

A high-performing, general-purpose visual understanding system should map visual inputs to both fine-grained leaf nodes of a taxonomy and coarse-grained inner nodes. Meanwhile, it should label an input hierarchically consistently along the path that traces a leaf up to the root. Figure 1 illustrates a case selected from our experiments that the model predictions lack *hierarchical consistency*, failing to follow the path of `Animal` → `Vertebrate` → `Fish` → `Spiny-finned Fish` → `Anemone Fish`.

Surprisingly, little has been done to assess the hierarchical visual understanding performance of vision large language models (VLLMs) [4, 29, 9, 72, 34, 29], which have the potential to make such a general-purpose vision system. Indeed, VLLMs unify various vision tasks (e.g., visual recognition [13], captioning [8], question answering [2], and retrieval [62]) into one model by anchoring visual encoders [46, 66, 10, 39] to a versatile pretrained LLM [19, 60], typically orders of magnitude bigger, offering integrated interactions with humans that involve images and videos in conjunction with natural language prompts. Comprehensively benchmarking VLLMs is essential for realizing their potential and identifying opportunities for improvements. Extensive benchmarks have recently emerged, such as the bilingual MMBench [36], manually labeled MME [16], and MMMU [64] collected from college exams. We refer readers to [67] for an extensive list.

This work systematically evaluates VLLMs' hierarchical visual understanding capabilities using six taxonomies and four hierarchical image classification datasets. Conventionally, the hierarchical im-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

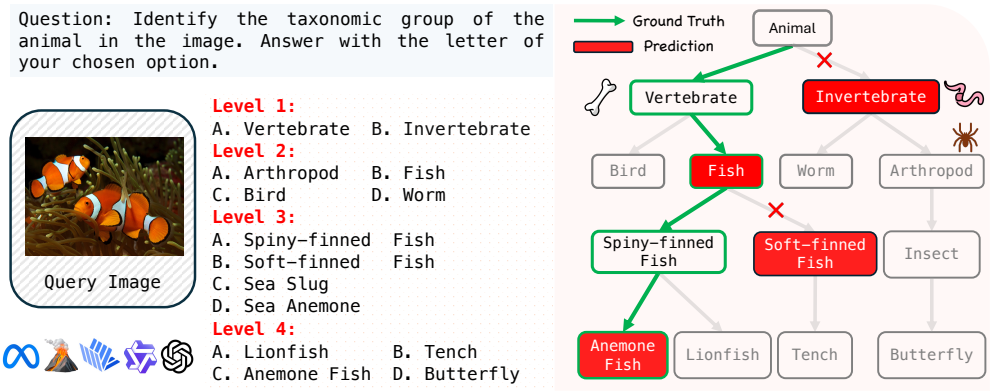


Figure 1: **Left**: Four-choice VQA tasks for evaluating VLLMs' hierarchical visual understanding. **Right**: A VLLM's answers (in red boxes) deviate from the ground truth path (**green arrows**), illustrating its lack of hierarchical consistency.

age classification [47, 44, 58, 61, 59] aims to classify visual inputs into semantically structured categories across multiple levels of specificity, in contrast to flat classification, which treats labels as mutually exclusive and unstructured. We construct about one million four-choice visual question-answering (VQA) tasks from the hierarchical datasets (see Figure 1 for some examples). The tasks traverse all taxonomy levels, and the four choices of an individual task are from the same level. When evaluating VLLMs' performance over these tasks, we stress hierarchical consistency because it is unique to hierarchical visual understanding and crucial for adaptability to users' varying granularity preferences [44, 12, 58].

Our main findings are as follows. First of all, many state-of-the-art VLLMs struggle with our VQA tasks, substantially lacking hierarchical consistency. For example, Qwen2.5-VL-72B [4] makes mistakes over 67% of the hierarchical paths in the iNaturalist [53] taxonomy. Moreover, in our attempt to tracing down the error causes, we find that LLMs are the bottleneck and lack taxonomy knowledge about the visual world. In contrast, the visual encoder and projector modules demonstrate the ability to retain highly discriminative and well-structured visual features. We further show that the LLM embeddings about the visual concepts contain sufficient hierarchical cues and organize them orthogonally, but the model cannot decode them. Finally, finetuning a VLLM using our VQA tasks enhance its LLM's (text) hierarchical consistency more than the VLLM's (visual) hierarchical consistency, reaffirming LLMs' bottleneck effect to some extent.

## 2 VLLMs Lack Hierarchical Consistency in Visual Understanding

We construct six hierarchical image classification benchmarks in a four-choice VQA format to systematically assess VLLMs' accuracy and hierarchical consistency in visual understanding. These benchmarks leverage datasets that inherently exhibit taxonomic structures, either derived from WordNet [37] or grounded in biological classification standards. In what follows, we formally define hierarchical image classification, followed by two evaluation metrics about accuracy and consistency, respectively. We then describe our VQA tasks and the first set of experiment results in this work.

### 2.1 Hierarchical Image Classification: Notations and Problem Statement

General image classification tasks typically assume a flat label space, where each image $x \in \mathcal{X}$ is assigned a class label $y \in \mathcal{Y}$ out of a predefined set $\mathcal{Y}$ of mutually exclusive categories. However, many real-world problems exhibit rich semantic structures, in which labels are naturally organized into a hierarchy $\mathcal{T} = (\mathcal{Y}, \mathcal{E})$ [44, 58, 61, 59], such as a tree or a directed acyclic graph. Here, $\mathcal{E} \subseteq \mathcal{Y} \times \mathcal{Y}$ denotes the set of directed edges representing parent-child relationships, where $(y_i, y_j) \in \mathcal{E}$ indicates that $y_i$ is the parent of $y_j$ in the hierarchy. In hierarchical image classification, the objective is not only to predict the leaf node label $y \in \mathcal{Y}_{leaf} \subseteq \mathcal{Y}$ but also to correctly recover its full ancestral path $(y_0, y_1, \cdots, y_L)$ in $\mathcal{T}$, where $y_0$ denotes the root node and $L$ is the depth of the hierarchy. In this paper, we aim to evaluate VLLMs' hierarchical image classification capabilities, identify their limitations and underlying causes, and enhance their performance based on these insights.

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

| Dataset | # Levels | # Leaf Nodes | # Test Images | Hierarchy Distribution |
|---|---|---|---|---|
| CUB-200-2011 [54] | 4 | 200 | 5,794 | 13-37-124-200 |
| iNaturalist-Plant [53] | 6 | 4,271 | 42.71K | 5-14-85-286-1702-4271 |
| iNaturalist-Animal [53] | 6 | 5,388 | 53.88K | 6-27-152-715-2988-5388 |
| ImageNet-Animal [13] | 11 | 397 | 19.85K | 2-10-37-81-123-81-65-41-64-34-2 |
| ImageNet-Artifact [13] | 8 | 492 | 24.60K | 5-40-149-205-162-62-44 |
| Food-101 [5] | 4 | 84 | 21.00K | 6-29-40-24 |

### 2.2 Two Evaluation Metrics about Accuracy and Consistency, Respectively

For evaluation, we mainly focus on the hierarchical consistency of model predictions [58, 43]. Besides, we are interested in the leaf-level classification accuracy [68, 35, 20], which can be viewed as the upper bound of the hierarchical consistency, detailed below.

**Hierarchical Consistent Accuracy (HCA) [58, 43].** This metric is defined as

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

where $N$ is the number of images in the testing set, $L^i$ denotes the depth of the hierarchy for the $i$-th input $x^i$ and may vary for different tasks in uneven trees, $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ is an image classifier, $\mathcal{Y}_j$ represents the set of labels at the $j$-th layer of the hierarchy, and $\mathbb{1}[\cdot]$ is an indicator function. HCA evaluates whether a model's predictions are consistent with the entire hierarchical path from the root to a leaf node. Specifically, it measures the proportion of samples for which all ancestor nodes along the predicted paths match the ground truth. This is a stricter metric than flat accuracy and serves as our primary evaluation criterion for hierarchical classification.

**Leaf-Level Accuracy** $\mathrm{Acc}_{\mathrm{leaf}}$ **[68, 35, 20].** It cares about the predictions at the most fine-grained level of a taxonomy:

$$\mathrm{Acc}_{\mathrm{leaf}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left[f_\theta\left(x^i;\mathcal{Y}_L\right)=y_L^i\right]. \tag{2}$$

Interestingly, $\mathrm{Acc}_{\mathrm{leaf}}$ upper-bounds HCA because correctly assigning a leaf label $y_L$ to an input $x$ contributes to $\mathrm{Acc}_{\mathrm{leaf}}$, but it does not increase HCA unless the model makes no mistake over all nodes in the path $(y_0, y_1, \cdots, y_L)$ connecting the leaf label to the root.

### 2.3 VQA Tasks Derived from Hierarchical Image Classification Datasets

VLLMs are the image classifiers $f_\theta$ in equations (1) and (2), and one can use language prompts to steer their output to a particular taxonomy level. More concretely, we formalize a VQA task for each image given a desired taxonomy level, $(x^i, \mathcal{Y}_j), i = 1, 2, \cdots, N, j = 1, 2, \cdots, L^i$, as follows.

**VQA Tasks.** We derive approximately one million four-choice VQA tasks and six taxonomies from four hierarchical image classification datasets [54, 53, 13, 5] to evaluate VLLMs in a closed-set setting. This setting mitigates the challenge of open-set generation, which involves a prohibitively large prediction space [68] and ambiguous prediction granularity. We test different VQA prompts (provided in Appendix C), and they generally follow this format:

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

Arguably, the four-choice VQA tasks are easier than the conventional hierarchical image classification, whose label space is orders of magnitude bigger than four. To compensate this difference, we make sure the four choices are from the same level of a taxonomy and use "confusing labels" in the VQA tasks. Specifically, we use SigLIP [66] to compute the cosine similarity scores between an image and all text labels other than the ground truth (at a particular taxonomy level), selecting

Table 2: The hierarchical consistent accuracy (HCA) and leaf-level accuracy $\mathrm{Acc}_{\mathrm{leaf}}$ of six open-source VLLMs, two CLIP-style models, and the proprietary GPT-4o.

| Model | iNat21-Animal | | iNat21-Plant | | ImgNet-Artifact | | ImgNet-Animal | | CUB-200 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HCA | $\mathrm{Acc}_{\mathrm{leaf}}$ | HCA | $\mathrm{Acc}_{\mathrm{leaf}}$ | HCA | $\mathrm{Acc}_{\mathrm{leaf}}$ | HCA | $\mathrm{Acc}_{\mathrm{leaf}}$ | HCA | $\mathrm{Acc}_{\mathrm{leaf}}$ |
| **Open-Source VLLMs** | | | | | | | | | | |
| LLaVA-OV-7B [29] | 4.53 | 26.47 | 4.46 | 27.51 | 17.15 | 80.77 | 34.36 | 65.50 | 11.51 | 44.23 |
| InternVL2.5-8B [9] | 8.52 | 27.65 | 5.56 | 28.36 | 21.42 | 78.07 | 37.82 | 65.19 | 22.07 | 45.56 |
| InternVL3-8B [72] | 11.93 | 35.40 | 8.68 | 36.39 | 17.87 | 77.50 | 42.31 | 69.41 | 25.75 | 50.52 |
| Qwen2.5-VL-7B [4] | 19.43 | 41.33 | 17.67 | 41.61 | 16.47 | 85.20 | 56.00 | 80.01 | 43.76 | 65.50 |
| Qwen2.5-VL-32B [4] | 26.90 | 46.98 | 24.64 | 48.57 | 26.30 | 84.51 | 62.23 | 80.48 | 56.80 | 69.00 |
| Qwen2.5-VL-72B [4] | 35.73 | 54.20 | 32.82 | 55.00 | 21.08 | 85.61 | 64.08 | 80.52 | 66.36 | 75.04 |
| **CLIP Models** | | | | | | | | | | |
| OpenCLIP [10] | 1.04 | 23.53 | 0.19 | 28.12 | 9.11 | 83.64 | 12.57 | 81.14 | 4.31 | 80.39 |
| SigLIP [66] | 2.15 | 12.71 | 0.46 | 18.84 | 6.41 | 87.19 | 24.40 | 86.85 | 23.18 | 73.84 |
| **Proprietary VLLM** | | | | | | | | | | |
| GPT-4o [1] | 42.95 | 63.79 | 35.53 | 62.95 | 27.57 | 86.05 | 67.69 | 85.50 | 81.96 | 87.25 |

the top three most similar labels as the distracting VQA choices. Besides, we provide the results of randomly sampled choices in Appendix B.

**Hierarchical Image Classification Datasets.** Table 1 summarizes the six taxonomies and four datasets we use to construct the VQA tasks. CUB-200-2011 (CUB-200) [54] is a fine-grained bird dataset containing 200 species. We prompt GPT-4o [1] to map each class to a four-level taxonomy: `Order` $\rightarrow$ `Family` $\rightarrow$ `Genus` $\rightarrow$ `Specie`. To ensure taxonomic accuracy, we cross-validate the generated hierarchy using corresponding entries from Wikipedia. In addition, we incorporate the iNaturalist-2021 (iNat21) dataset [53], a large-scale collection with species-level annotations spanning various biological taxa. We separate it into two taxonomies, Plant and Animal, comprising 4,271 and 5,388 leaf nodes, respectively, and six levels. Both CUB-200 and iNat21 provide well-established biological taxonomies with even hierarchical depths. To increase structural diversity, we also experiment with ImageNet-1K (ImgNet) [13], whose leaf labels are coarser-grained than iNat21 and CUB-200. ImgNet is built upon the WordNet [37]. We extract two relatively well-structured subsets from ImgNet: ImgNet-Animal and ImgNet-Artifact, following [58]. We further refine these subsets to improve label quality and semantic consistency. Food-101 [5] is about food classification, and its hierarchy is constructed based on the recent work of Liang and Davis [32].

### 2.4 Experiments and Findings

We mainly study state-of-the-art open-source VLLMs: The Qwen2.5-VL [4] models of 7B, 32B, and 72B parameters, InternVL2.5-8B [9], InternVL3-8B [72], and LLaVA-OV-7B [29]. Meanwhile, we include the proprietary GPT-4o's results for reference; in general, GPT-4o slightly outperforms Qwen-2.5-VL-72B, but the main findings below still apply. Finally, we experiment with two CLIP-style [46] models, SigLIP-SO400M [66] and OpenCLIP-L [10], following the experiment protocol in [46] except that the candidate labels for each test image are restricted to the same four choices as fed to VLLMs. Table 2 shows the results about the models' hierarchical consistency (HCA) and leaf-level accuracy ($\mathrm{Acc}_{\mathrm{leaf}}$) on iNat21, ImgNet, and CUB-200. The Food-101 results are in Appendix B to save space in the main text. We draw the following conclusions.

**VLLMs Lack Hierarchical Consistency in Visual Understanding.** Regardless of the leaf-level accuracy, all open-source VLLMs, CLIP models, and GPT-4o lack hierarchical consistency because their HCA is significantly lower than $\mathrm{Acc}_{\mathrm{leaf}}$ (up to 99.3% relatively). The gaps on iNat21-Plant are especially big (e.g., 32.82 vs. 55.00 for Qwen2.5-VL-72B and 35.53 vs. 62.95 for GPT-4o). While one might expect better results on ImgNet, neither open-source VLLMs nor GPT-4o can make their HCA match $\mathrm{Acc}_{\mathrm{leaf}}$ — more than 20% decrease for all models, indicating that VLLMs make many mistakes along the paths from the taxonomies' roots to the leaf nodes even when they are correct over the leaves.

**Fine-Grained Visual Recognition Remains Challenging for VLLMs.** While VLLMs and CLIP models perform moderately well on ImgNet, they struggle with fine-grained object recognition; on

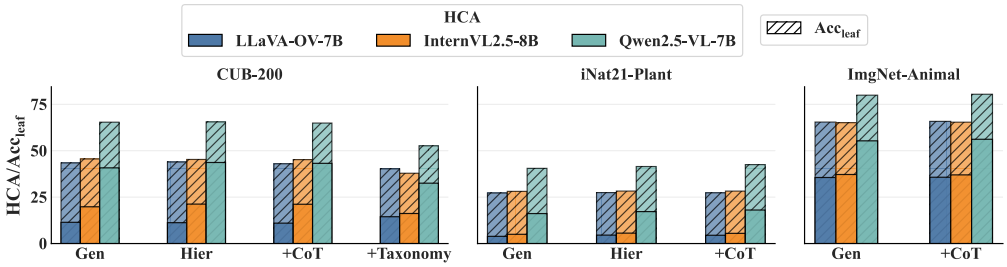


Figure 2: Prompt variants and their effects on VLLMs' hierarchical consistency (HCA) and fine-grained recognition $\mathrm{Acc}_{\mathrm{leaf}}$ (**Gen**: general prompts, **Hier**: hierarchical prompts, **+CoT**: prompts with Chain-of-Thought reasoning, **+Taxonomy**: prompts that include an explicit taxonomy in the JSON format. Please see Appendix C for details and examples.).

the iNat21 dataset, even the best-performing GPT-4o gives rise to only 63% leaf-level accuracy, far from its 86% on ImgNet. Notably, InternVL2.5 and LLaVA-OV's results (about 27%) on iNat21 are only slightly above random guess (25%), and the CLIP models are barely on par with random guess. In contrast, a small task-specialized model [23] leads to 61.56% leaf-level accuracy on iNat21, and some models [11, 69] achieve 93% accuracy on CUB-200, outperforming all the general-purpose VLLMs in our experiments. These findings are consistent with the recent work [17, 68, 20, 63] that recognizes the limitation of VLLMs on (fine-grained) image classification.

**Scaling Laws Works for Hierarchical Visual Understanding.** Both hierarchical consistency and leaf-level accuracy improve as the size of the Qwen2.5-VL series of models increases. Moreover, the gap between HCA and $\mathrm{Acc}_{\mathrm{leaf}}$ progressively narrows. However, the largest models (Qwen2.5-VL-72B and GPT-4o) are still unsatisfactory in terms of both hierarchical consistency and fine-grained recognition, especially on the iNat21 benchmark.

**Qwen2.5-VLs Are Among the Most Powerful Open-Source VLLMs.** LLaVA-OV-7B's hierarchical consistency and leaf-level accuracy are below InternVLs and Qwen2.5-VLs. InternVL3-8B improves upon InternVL2.5-8B, but it is still under par with Qwen2.5-VL-7B.

# 3 Why Are VLLMs Poor at Hierarchical Image Classification?

We systematically investigate potential causes of VLLMs' low performance on hierarchical visual understanding. We first extensively study prompt variations in Section 3.1 and reveal that some prompts can lead to marginally better results than the rest, but the results remain generally bad. We then examine VLLMs' visual encoders and subsequent visual tokens to see whether and where essential visual information is lost when it forwards through VLLMs (Section 3.2). Interestingly, the discriminative cues in the visual tokens are maintained across various stages of the VLLM architectures, leading to about the same hierarchical image classification results immediately after the visual encoder, after the projection to the language token space, and at the very last layer of an LLM. Finally and surprisingly, we find that the generally believed powerful LLMs, even the one with 72B parameters in our experiments, lack basic taxonomy knowledge and are likely responsible for VLLMs' poor performance on hierarchical visual understanding! (We believe this conclusion is true for open-source VLLMs, but we urge readers not to extrapolate it to proprietary LLMs because we could not probe their intermediate embeddings.)

## 3.1 Language Prompts Are *Not* the Bottleneck

Prompt engineering often comes as a remedy for boosting VLLMs' performance in different applications [6, 55, 68, 58]. Could it also rescue VLLMs on our hierarchical visual understanding tasks? We strive to test prompt variants comprehensively. We specify the taxonomy levels in the prompts for CUB-200 [54] and iNat21 [53], whose taxonomies are grounded in biology. We even add CUB-200's complete taxonomy as a JSON file to the prompts. For the other datasets with more generic taxonomies, we test general and chain-of-thought [24, 57] prompts derived from the template in Section 2.3. Appendix C provides all prompts in detail, and Figure 2 shows the results of some high-performing prompts. We can see from the results that the prompt design alone is insufficient to improve VLLMs' hierarchical consistency or leaf-level accuracy.

Table 3: (Text) HCA of VLLMs' LLMs and its correlation $\rho$ with VLLMs' (visual) HCA

| **LLM of** | iNat21-Animal | iNat21-Plant | ImgNet-Artifact | ImgNet-Animal | CUB-200 | $\rho$(text,visual) |
|---|---|---|---|---|---|---|
| LLaVA-OV-7B [29] | 11.56 | 28.49 | 29.27 | 56.93 | 33.45 | 0.9116 |
| InternVL2.5-8B [9] | 38.15 | 41.15 | 35.32 | 66.11 | 49.11 | 0.8832 |
| InternVL3-8B [72] | 54.20 | 47.49 | 31.86 | 69.92 | 59.87 | 0.9030 |
| Qwen2.5-VL-7B [4] | 52.08 | 64.21 | 35.06 | 68.14 | 63.86 | 0.8640 |
| GPT-4o [1] | 96.85 | 96.70 | 42.31 | 89.56 | 98.81 | 0.7980 |

### 3.2 Visual Embeddings Are *Not* the Bottleneck

The open-source VLLMs in this work vary in specific implementations, but their core components are the same: A visual encoder mapping images to embeddings, a projector translating visual embeddings into the language token space, and an LLM. If the hierarchical structure and discriminativeness are lost before the visual embeddings reach LLMs, the overall VLLMs would inevitably perform poorly on our hierarchical visual understanding tasks. Hence, it is crucial to examine the visual embeddings. We train three linear classifiers per taxonomy level to respectively probe the visual encoder, projector, and last layer of an LLM, where the image representations are an average of the visual tokens. Further details and results of the probing are provided in Appendix C.

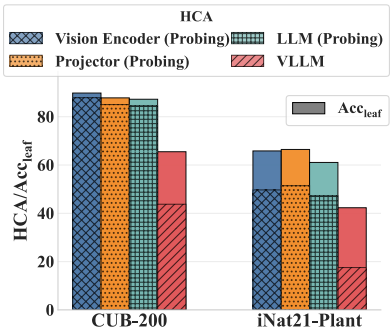


Figure 3: Qwen2.5-VL-7B vs. linearly probing the visual tokens at various stages of Qwen2.5-VL-7B on CUB-200 and iNat21-Plant.

Figure 3 shows the probing results of Qwen2.5-VL-7B over CUB-200 [54] and iNat21-Plant [53]. Remarkably, the linear classifiers outperform Qwen2.5-VL-7B all around. They achieve not only higher leaf-level accuracy than Qwen2.5-VL but also much better hierarchical consistency, even though the classifiers of different taxonomy levels are independently trained. Moreover, the linear probing results remain about the same at different stages of the forward propagation (i.e., immediately after the visual encoder, projector, and last layer of the VLLM), indicating that the visual tokens remain discriminative and structurally rich throughout different LLM layers. These results are a strong defense for the visual embeddings: They carry sufficient hierarchical and discriminative cues and should not be blamed for VLLMs' poor hierarchical visual understanding performance.

### 3.3 *LLMs Are the Bottleneck* in VLLMs' Hierarchical Visual Understanding

The huge discrepancy between the results of linearly probing visual tokens and VLLM performance in Figure 3 propels us to investigate other potential causes of VLLMs' low hierarchical consistency beyond the visual embeddings, and we find that the influential LLMs are the bottleneck.

#### 3.3.1 Open-Souce VLLMs' LLMs Lack Taxonomy Knowledge

We separate LLMs from open-source VLLMs and examine how much they know about the taxonomies used in our experiments. Mechanically, we reformulate our VQA tasks to a text-only version by replacing the images with their corresponding leaf labels:

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

This process results in about 0.7 million QA tasks after deduplication. We use them to assess LLMs and report the (text) HCA results in Table 3 — we use (text/visual) HCA to refer to LLMs/VLLMs' performance on text/visual QA tasks for clarity. We find that Qwen2.5-VL-7B's LLM achieves only 63.86% (text) HCA on CUB-200, whose taxonomy comprises merely four levels. The LLMs of LLaVA-OV and InternVL-2.5 give rise to even lower (text) HCAs on CUB-200 (33% and 49%). One might wonder if these low (text) HCAs are due to that the biology taxonomy underlying CUB-200 is too specific for general LLMs.

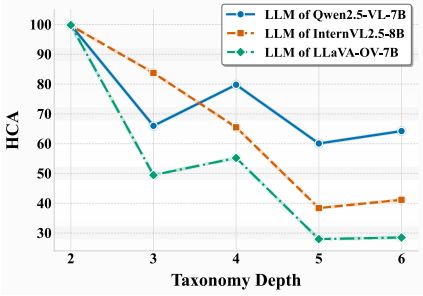


Figure 4: Text HCA of different VLLMs' LLMs over the iNat21-Plant taxonomies of various depths.

However, Table 3 further reveals that the LLMs also cannot perform well on ImgNet's general taxonomies. Besides, we progressively simplify our QA tasks by chopping the iNat21-Plant taxonomy level by level. Figure 4 plots the (text) HCA results, which increase as the taxonomy becomes shallower (and, correspondingly, the leaf nodes are less fine-grained). Still, they are below 90% regardless of the taxonomies' depths. There are noticeable drops at Levels 3 and 5 for Qwen2.5-VL and LLaVA-OV's LLMs, implying that they pose more challenges than the other levels for the LLMs' hierarchical reasoning. These results are surprising to a large degree, given the recent success of LLMs over various benchmarks and domains [1, 50, 60, 33, 45].

**Correlation between (text) HCA and $\mathrm{Acc}_{\mathrm{leaf}}$-scaled (visual) HCA.** An LLM's low (text) HCA undoubtedly discounts its corresponding VLLM's hierarchical consistency on visual inputs. We can quantify this notion using Pearson's correlation coefficient. Since the (text) HCA's corresponding leaf-level accuracy is 100% — we replaced images with their ground-truth leaf labels when making the text QA tasks, we normalize (visual) HCA by $1/\mathrm{Acc}_{\mathrm{leaf}}$. The last column in Table 3 shows that the correlation between (text) HCA and $\mathrm{Acc}_{\mathrm{leaf}}$-scaled (visual) HCA is as high as 0.9116.

*A note about GPT-4o's (text) HCA.* The analyses above apply to only open-source VLLMs because we cannot separate LLMs from the proprietary GPT-4o. Unlike the open-source LLMs' low (text) HCA, GPT-4o's (text) HCA scores are as high as 98.81. Hence, the LLM part is not GPT-4o's bottleneck in hierarchical visual understanding; instead, there are other possible causes of GPT-4o's hierarchical inconsistency about the visual world.

### 3.3.2 Why Are LLMs Poor at Hierarchical *Text* Classification?

In what follows, we present some preliminary quests into why and where LLMs fail at the seemingly simple hierarchical four-choice text classification tasks. We rule out the vision-language tuning that anchors visual encoders to pretrained LLMs and conclude that the language decoders are responsible for LLMs' lack of taxonomy knowledge.

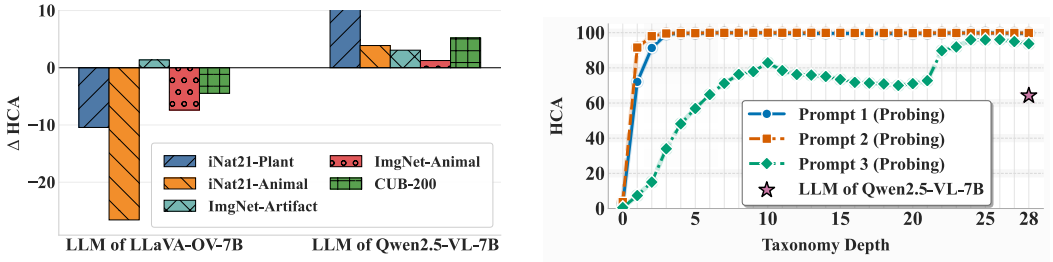


Figure 5: **Left**: (Text) HCA difference between vision-language-tuned LLMs and the original ones. **Right**: (Text) HCA of linearly probing different layers of Qwen-2.5-VL-7B's LLM on iNat21-Plant.

**Vision-Language Tuning Is *Not* the Reason.** Acute readers likely have noted that our previous LLM results are about the LLM parts of VLLMs, not the "true" standalone LLMs. Does the vision-language tuning, which is needed when one connects a visual encoder with an LLM, compromise LLMs and potentially induce catastrophic forgetting of taxonomy knowledge?

We answer this question by studying the original LLMs from which VLLMs are initialized, using the same text-only hierarchical classification setup described in Section 3.3. Figure 5 (Left) compares LLaVA-OV-7B and Qwen2.5-VL-7B's LLMs with their corresponding original LLMs. First of all, we see that the original LLMs are on par with or even worse than their vision-tuned counterparts, indicates that the standalone LLMs still lack a strong grasp of taxonomy knowledge. Interestingly, Qwen2.5-VL's LLM actually outperforms its original LLM on all taxonomies; in other words, the vision-language tuning actually enhances the LLM's (text) hierarchical consistency. In contrast, LLaVA-OV's vision-language tuning weakens the LLM's (text) HCA.

**LLMs Encode Hierarchical Structures Effectively but Cannot Decode Them Sufficiently.** Next, we shift attention to the LLM embeddings of the concepts in our taxonomies — if the embeddings do not provide sufficient hierarchical structural cues, there is little chance LLMs can decode them. To this end, we convert a taxonomy into language prompts of three variants:

**Prompt 1:** `<leaf node label> (e.g., Blue Jay) belongs to the <hierarchy> (e.g., Order) of <ground truth> (e.g., Passeriformes).`
**Prompt 2:** `Given the <leaf node label>, what is its taxonomic classification at the <hierarchy> level? It belongs to <ground truth>.`
**Prompt 3:** `Given the <leaf node label>, what is its taxonomic classification at the <hierarchy> level?`

We then train a linear classifier for each taxonomy level to probe the average embedding of the language tokens in every layer of an LLM. Figure 5 (Right) summarizes the (text) HCA results of Qwen2.5-VL-7B's LLM on iNat21-Plant: The text embeddings give rise to highly hierarchically consistent linear probes. Especially for Prompt 3, with the ground-truth hierarchy labels withheld, the linear probes that receive only the leaf node embeddings can still achieve near-perfect hierarchical consistency in the LLM's deeper layers. In other words, the specialized linear probes can decode the taxonomy knowledge significantly better than the general-purpose LLM.

**LLMs' Hierarchical Orthogonality Does Not Guarantee Hierarchical Consistency.** Park et al. [42] recently predicted that LLMs represent hierarchical relations orthogonally in the representation space, e.g., `animal` is orthogonal to `bird`$-$`mammal`. They validated the prediction using Gemma [51] and LLaMA [19], and we further verify it in Figure 6 using both the original Qwen2.5-7B and the one after vision-language tuning. This pleasant geometric interpretation is, unfortunately, shadowed by the poor performance of Gemma and Qwen2.5-7B on our taxonomy QA tasks — we report the Gemma results in Appendix C. We argue that more fine-grained analyses of the LLM representation are required to establish a relationship between LLMs' hierarchical consistency and geometry.

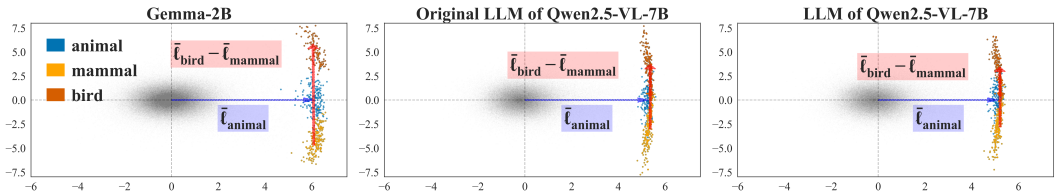


Figure 6: Hierarchical semantics are encoded as orthogonality in different LLMs' representation spaces (figures drawn following [42]).

# 4 LLMs Gain More Hierarchical Consistency than VLLMs from Finetuning

Could we improve the VLLMs' hierarchical visual understanding capabilities via finetuning using our VQA tasks built upon taxonomies? Likely, no, because LLMs are the bottleneck: The LLMs' hierarchical consistency over text-only tasks is so bad (Table 3) that we conjecture this shortcoming can only be fixed in the pretraining stage rather than the "tail patching" finetuning stage.

Still, the following presents some LoRA-finetuning [22] experiments with Qwen2.5-VL-7B, the best-performing 7B VLLM in our previous experiments, mainly for two reasons. One is to see how much finetuning could help, even though we believe pretraining instead of finetuning should be the rescue to VLLMs' hierarchical inconsistency. The other is further to investigate the interplay between VLLMs and their LLMs — interestingly, our results reaffirm that LLMs are the bottleneck for VLLMs' hierarchical visual understanding because LLMs' performance gain from the finetuning upper-bounds VLLMs'. Our finetuning data consists of VQA tasks constructed from iNat21-Plant's training set, covering 3,771 species nodes in the taxonomy instead of the full 4,271 species nodes. We then evaluate the finetuned model's improvement on iNat21-Plant, its generalization to other hierarchical visual understanding datasets, and how well it maintains the general vision-language capabilities. Please see Appendix D for more details on the training.

**Results and Discussion.** Tables 4 shows that finetuning Qwen2.5-VL using the VQA tasks that partially cover the iNat21-Plant taxonomy delivers improvements on both iNat21-Plant and other datasets. On iNat21-Plant, HCA rises from 17.67 to 29.34 ($+11.67$ absolute gain), while $\text{Acc}_{\text{leaf}}$

Table 4: (Visual) HCA and $\text{Acc}_{\text{leaf}}$ of Qwen2.5-VL-7B before and after the 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) | 23.38 | 45.00 | 29.34 | 47.66 | 58.62 | 80.28 | 46.17 | 67.12 |
| $\Delta$ | +3.95 | +3.67 | +11.67 | +6.05 | +2.62 | +0.27 | +2.41 | +1.62 |

Table 5: (Text) HCA of the LLM of Qwen2.5-VL-7B before and after the 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) | 65.63 | 84.87 | 72.39 | 66.15 |
| $\Delta$ | +13.55 | +20.66 | +4.25 | +2.29 |

gains 6.05. The HCA on ImageNet-Animal increases from 56.00 to 58.62 and on CUB-200 from 43.76 to 46.17. More interestingly, Table 5 indicates that the LLM's (text) HCA increases more from the finetuning than Qwen2.5-VL's (visual) HCA (e.g., 20.66 vs. 11.67 on iNat21-Plant and 4.25 vs. 2.62 on ImgNet-Animal). To some extent, this finding reaffirms that LLMs are the bottleneck of VLLMs' hierarchical visual understanding, and one has to improve LLMs' (text) taxonomy knowledge to boost VLLMs' (visual) hierarchical consistency. Besides, our results demonstrate that vision-language training can benefit both VLLMs and their LLMs, aligning with some recent advocates for improving LLMs using multimodal data beyond language only [31, 52]. Appendix D reports more results and discussion, including that the finetuned model does not lose its general capability tested on MME [16], MMBench [36], and SEED-Bench [30].

## 5 Related Work

Hierarchical classification [47, 25] enables many applications. It is vital for a comprehensive understanding of the visual world [61, 43, 65, 48, 7, 44] and many language concepts [70, 56, 71, 21]. Several recent studies have revisited this longstanding problem and shown that CLIP-style [46] models lack consistency across taxonomic levels [58, 18]. Wu et al. [58] evaluate CLIP under multiple levels of semantic granularity and introduce a hierarchy-consistent prompt tuning method. Pal et al. [40] enhance CLIP's hierarchical representations by embedding them to a hyperbolic space. Xia et al. [59] further extend this direction by incorporating graph-based representation learning. Novack et al. [38] use hierarchical information to improve zero-shot classification accuracy. Zhang et al. [68] first identified the limitations of current VLLMs in fine-grained image classification. Building on this, Liu et al. [35] further assess a broader range of VLLMs. He et al. [20] point out a potential cause, the scarcity of image class names in pretraining. Beyond closed-set evaluation [63, 17], Conti et al. [12] benchmark VLLMs' open-world classification, while Snæbjarnarson et al. [49] propose to evaluate VLLMs' open-set predictions using a taxonomic similarity rather than exact string matching. However, to the best of our knowledge, no prior work has examined VLLMs under the hierarchical visual understanding context.

## 6 Conclusion

This work presents a systematic evaluation of state-of-the-art VLLMs's hierarchical visual understanding performance. We find that both open-source VLLMs and the proprietary GPT-4o give rise to low hierarchical consistency over six taxonomies of visual concepts. Probing results reveal that the visual and text embeddings carry rich hierarchical and discriminative cues, whereas the LLMs fail to decode them, implying LLMs are the bottleneck. Finetuning on hierarchical VQA tasks improves VLLMs' hierarchical consistency on visual inputs while preserving their performance on general VQA tasks. Intriguingly, the finetuning benefits the LLMs (text) hierarchical consistency more than the corresponding VLLM's (visual) hierarchical measure. Ingesting the taxonomy-knowledge gap to LLMs, likely during pretraining rather than post-hoc patching, is a promising path toward VLLMs that reason coherently across different levels of semantic granularity about the visual world.

# References


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

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

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

[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.

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

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] 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.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[9] 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.

[10] 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.

[11] Po-Yung Chou, Yu-Yung Kao, and Cheng-Hung Lin. Fine-grained visual classification with high-temperature refinement and background suppression. *arXiv preprint arXiv:2303.06442*, 2023.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[20] 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.

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

[22] 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.

[23] Pranav Jeevan, Kavitha Viswanathan, Amit Sethi, et al. Wavemix: A resource-efficient neural network for image analysis. *arXiv preprint arXiv:2205.14375*, 2022.

[24] 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.

[25] 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.

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

[27] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009.

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

[29] 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. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

[30] 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.

[31] Yunxin Li, Baotian Hu, Wei Wang, Xiaochun Cao, and Min Zhang. Towards vision enhancing llms: Empowering multimodal knowledge storage and sharing in llms. *arXiv preprint arXiv:2311.15759*, 2023.

[32] Tong Liang and Jim Davis. Making better mistakes in clip-based zero-shot classification with hierarchy-aware language prompts. *arXiv preprint arXiv:2503.02248*, 2025.

[33] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

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

[35] 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.

[36] 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.

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

[38] 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.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[40] 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.

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

[42] 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.

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

[44] 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.

[45] Qwen Team. Qwen3 technical report. https://github.com/QwenLM/Qwen3/blob/main/Qwen3_Technical_Report.pdf, 2025. Last accessed: 14 May 2025.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[51] 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.

[52] Haoqin Tu, Bingchen Zhao, Chen Wei, and Cihang Xie. Sight beyond text: Multi-modal training enhances LLMs in truthfulness and ethics. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=2Zl0zc7f08.

[53] 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.

[54] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Jul 2011.

[55] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5749–5757, 2024.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[61] 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.

[62] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

[63] 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.

[64] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[65] 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.

[66] 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.

[67] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. URL https://arxiv.org/abs/2407.12772.

[68] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MwmmBg1VYg.

[69] Zhicheng Zhang, Hao Tang, and Jinhui Tang. Multi-scale activation, selection, and aggregation: Exploring diverse cues for fine-grained bird recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10385–10393, 2025.

[70] 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.

[71] 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.

[72] 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.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims are further disscussed in Section 2, Section 3 and Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Discussed in Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed descriptions of evaluation benchmarks, model architectures, training protocols, and evaluation metrics to ensure reproducibility of the main results in the paper. The code will be released upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data will be made publicly available upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer:[Yes]

Justification: We provide detailed descriptions of data curation procedures, model architectures, evaluation metrics, and training protocols in corresponding sections. The code for all experiments will be released upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We did not include error bars or statistical significance analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Discussed in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Discussed in Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit all external assets used in this work, including datasets, pre-trained models, and code repositories. Licenses and terms of use are reviewed and respected, with appropriate citations provided in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed descriptions of the hierarchical VQA benchmark curation process, model architecture, training protocols, and evaluation metrics in the corresponding sections. All data, code, and model checkpoints will be released upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our evaluation involves vision large language models (VLLMs) with large language model (LLM) backbones as a core component. We describe how the LLM is used for hierarchical classification and detail its role in the architecture, prompting, and evaluation.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.