Taxonomy-Aware Evaluation of Vision–Language Models

Vésteinn Snæbjarnarson[1,2*] Kevin Du[2] Niklas Stoehr[2] Serge Belongie[1]
Ryan Cotterell[2] Nico Lang[1] Stella Frank[1]
[1]University of Copenhagen [2]ETH Zürich

Abstract

When a vision–language model (VLM) is prompted to identify an entity depicted in an image, it may answer "I see a conifer," rather than the specific label NORWAY SPRUCE. *This raises two issues for evaluation: Firstly, the unconstrained generated text needs to be mapped to the evaluation label space (i.e.,* CONIFER*). Secondly, a useful classification measure should give partial credit to less-specific, but not incorrect, answers (*NORWAY SPRUCE *being a type of* CONIFER*). To meet these requirements, we propose a framework for evaluating unconstrained text predictions such as those generated from a vision–language model against a taxonomy. Specifically, we propose the use of hierarchical precision and recall measures to assess the level of correctness and specificity of predictions with regard to a taxonomy. Experimentally, we first show that existing text similarity measures do not capture taxonomic similarity well. We then develop and compare different methods to map textual VLM predictions onto a taxonomy. This allows us to compute hierarchical similarity measures between the generated text and the ground truth labels. Finally, we analyze modern VLMs on fine-grained visual classification tasks based on our proposed taxonomic evaluation scheme. Data and code are made available at* https://github.com/vesteinn/vlm-eval.

1. Introduction

Modern vision–language models (VLMs) [1, 6, 12, 41, 46, 50] serve as viable one-stop shops for a wide variety of tasks at the intersection of vision and language. Given an image as input, VLMs generate unconstrained text, a key enabler of their flexibility across tasks. This flexibility, however, introduces additional challenges in their evaluation. In this paper, we study the problem of evaluating the unconstrained text generated by VLMs in the context of fine-grained visual categorization (FGVC), and put forward two desiderata when evaluating VLMs on FGVC (Fig. 1):

*Correspondence: vesteinnsnaebjarnarson@gmail.com.

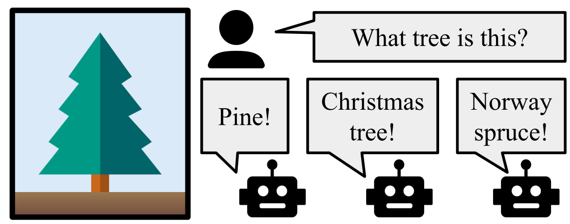

Figure 1. **Vision–language-models (VLMs) as fine-grained classifiers.** VLMs generate text with varying degrees of specificity and similarity to gold-standard label classes. We tackle the problem of aligning these outputs to taxonomic classes.

Evaluation with Textual Awareness. When tasked with classifying entities within an image, VLMs often generate text that does not inherently fit within the predefined set of labels used in conventional evaluations. For example, a model might describe an image with the phrase *pines in the snow*, while the ground-truth label is simply PINE TREE. We suggest that a good evaluation should rely on fuzzy or representation-based matching techniques, which assess the similarity between the generated text and the predefined class labels, rather than whether the VLM generated the predefined class's label exactly. Such approaches have a long history in natural language processing (see, e.g., [42, 58]). We call an evaluation measure for FGVC that is compatible with unconstrained text **textually aware**.

Evaluation with Hierarchical Awareness. Standard evaluation methods for VLM-based entity classification typically treat text generated by a VLM in a binary manner, i.e., it is either completely correct or incorrect [18, 31, 74]. However, many classification tasks involve label sets that are organized hierarchically, such as taxonomic trees or knowledge graphs. When such a structure is available, we suggest that evaluation measures take advantage of the structure, i.e., predictions should be judged based on their semantic

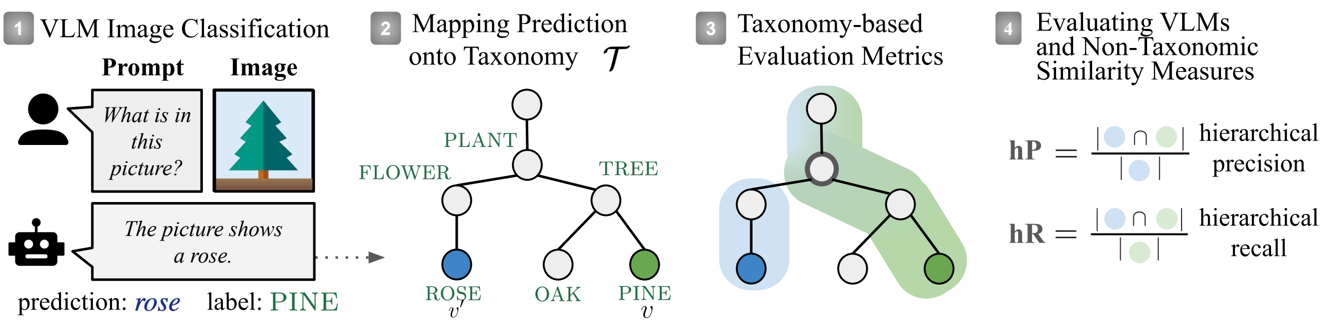

Figure 2. **Framework for taxonomy-aware VLM evaluation.** We propose a method for granular evaluation of the open-ended image classification capabilities of VLMs. To this end, we map model predictions onto a taxonomy, e.g., sourced from a large knowledge graph. We then use hierarchical precision (**hP**) and recall (**hR**) to evaluate how taxonomically accurate and specific a model prediction is. This, in turn, lets us evaluate and compare different VLMs and prompts.

proximity to the correct label within the hierarchy. Fortunately, these structured label spaces are common in FGVC and other real-world classification settings. We call a VLM evaluation measure that is sensitive to taxonomic relationships **hierarchically aware**.

In the technical portion of the paper, we propose a novel evaluation measure for assessing VLMs' performance on FGVC tasks that incorporates both textual and hierarchical awareness. We term such an evaluation strategy **taxonomy-aware**. Fig. 2 illustrates the approach for taxonomy-aware VLM evaluation that we present in this paper. Our measure is based on hierarchical precision and recall [33] after placing strings generated by the VLM on a given taxonomy. To contextualize our taxonomy-aware evaluation scheme within the literature, we first evaluate existing text similarity measures from NLP to assess whether they implicitly capture taxonomic relationships. Our analysis reveals that these measures do not align well with the underlying taxonomic structure, motivating our novel taxonomy-aware evaluation framework that explicitly maps unconstrained text generated from a VLM onto a taxonomy.

In the empirical part of the paper, we explore our taxonomy-aware evaluation scheme on a range of VLMs, varying in size, training strategies, and performance, under our proposed taxonomy-aware evaluation scheme. To do so, we apply diverse prompt templates to control for accuracy, i.e., whether the model generates correct answers, and specificity, i.e., how detailed the response should be and whether to be more general when the model is not confident. In support of our evaluation, we leverage large, carefully constructed taxonomies derived from knowledge graphs extracted from Wikidata [71] and the Linnean Catalogue of Life [30]. These taxonomies span diverse domains, e.g., food, sports, animals, plants, cars, and landmarks, providing a holistic characterization of VLM behavior. Our results demonstrate that our proposed taxonomy-aware evaluation measures effectively quantify varying degrees of specificity that more rigid evaluation measures like accuracy fail to capture. We find that larger commercial models respond more successfully to prompts requesting varying levels of specificity, revealing substantial differences compared to the prompt sensitivity exhibited by smaller, publicly available models. These findings illustrate how taxonomy-aware evaluation can uncover nuanced model behaviors when conventional evaluation measures cannot.

2. Background

We briefly review prior work on evaluating VLMs and textual similarity measures, with emphasis on approaches to visual categorization and hierarchical evaluation. A more general overview of the use of hierarchical labels for visual categorization is given in §17 in the supplement.

2.1. Evaluating VLMs for Classification

Recent work has explored VLM evaluation in open-domain visual recognition and question answering, often as a byproduct of general model development. A dominant approach involves retrieving the most similar target class label to an unconstrained text using maximum inner-product search [15, 31]. The retrieved label can then be directly used for classification or further processed to extract candidate labels. In some cases [26], these extracted labels are compared to the original image using image-to-text similarity methods, such as CLIP [55]. In-context learning for classification [27] leverages large language models and a few example labels to generate or extract an image label from a longer caption. Other approaches rely on models like GPT-4 [50] to score generated answers based on a given question and reference answer [4, 34, 46, 87]. A recent method [26] expands classification label spaces by prompting models to produce increasingly specific labels using taxonomically structured templates extracted from ImageNet [18]. However, final model predictions are still evaluated using string-matching measures. Some multi-task VLM benchmarks include image classification tasks [9, 78], typically using exact-match evaluation or text-based measures borrowed from image captioning; see §17 in supplementary material. The Open-domain Visual Entity (OVEN) benchmark [31] relies on representation-based methods while other work avoids direct classification tasks altogether [84].

2.2. Annotation and Prediction Similarity

Language generation tasks such as image captioning have traditionally been evaluated using measure based on string similarity, e.g., **ROUGE** [42], **BLEU** [53], **METEOR** [5] and **CIDER** [69]. They primarily rely on surface form (string) overlap in prediction and ground truth labels, and as such, do not capture semantic similarities. This led to a family of trained measures that rely on a language model that, given a pair of inputs, predicts a similarity score, typically based on the semantic match. Examples of these are **BERTScore** [82], and **SentenceBERT** [58]. Causal and bi-directional language models have been shown to encode varied hierarchical semantic knowledge in human-interpretable metric spaces [28, 52, 54]. Natural language inference (**NLI**), also known as textual entailment [11, 73], can also capture hierarchical specificity judgments, since sub-categories entail their supercategories, i.e., *This is a field sparrow* entails *This is a bird*. In computer vision, this idea has been adopted as visual entailment [77].

Contrastive vision and language models [55, 81] can be used to compare image and text pairs using top-k retrieval by inner product similarity. This method relies on the intrinsic quality of the representations. While impressive, these have been shown to fail in cases requiring combining complex compositional information that was not seen together at training time [80]. More recently, prompt-based methods for scoring textual similarity have been proposed: For instance, the LLaVA models [46] are evaluated on the COCO[44] image captioning datasets by asking GPT-4 [50] to provide numerical values (1-10) for a range of measures such as relevance and accuracy. Another approach is fine-tuning large language models to act as evaluation judges [87]. This approach has also been claimed to outperform other evaluation methods for machine translation [34]. A more detailed description of the text similarity measures we compare against in this paper is given in the supplement in section §10.

3. Taxonomy-aware Evaluation

Taxonomy-aware classification evaluation is a subtle and understudied problem [33, 35, 62]. It is necessary to assign partial credit to answers that share some taxonomic similarity, e.g., to answers that are within the correct sub-tree, even if they do not match the correct node exactly. **Hierarchical precision** (**hP**) and **hierarchical recall** (**hR**) [33] are two existing measures that suitably assign partial credit. Both take the length of the longest common path from the root node of the taxonomy—equivalent to the size of the **ancestor set**—and normalize by either the length of the predicted path (**hP**) or the target path (**hR**).

We now introduce **hP** and **hR** formally. Let Σ be an alphabet of symbols, and Σ^* be the Kleene closure, i.e., the set of all strings whose symbols are drawn from Σ. We assume the VLM under consideration constitutes a distribution over Σ^* when it is conditioned on an image. We then define a **taxonomy** $\mathcal{T}$ as a triplet $(V, \mathcal{L}, E)$, where V represents a finite set of **nodes**, $\mathcal{L}\colon V \to \Sigma^*$ is a function that maps each node to a string, called a **label**, and $E \subset V \times V$ is a set of **directed edges** between pairs of nodes. Additionally, we require that the directed graph (V, E) forms a directed tree with a unique root $\rho \in V$. The set of nodes encountered along the unique path from the taxonomy tree's root node ρ to a given node v constitutes that node's ancestor set, which we denote $\mathrm{anc}(v)$. Formally, for any node $v \in V$, let $\mathrm{par}(v)$ denote v's unique parent node in the tree; $\mathrm{par}(\rho)$, i.e., with the root note ρ as an argument, is undefined. We then recursively define $\mathrm{anc}(v)$ as:

$$\mathrm{anc}(v) \stackrel{\text{def}}{=} \{\rho\} \qquad \text{(1a, base case; } v = \rho\text{)}$$

$$\mathrm{anc}(v) \stackrel{\text{def}}{=} \{v\} \cup \mathrm{anc}(\mathrm{par}(v)). \qquad \text{(1b, inductive case; } v \neq \rho\text{)}$$

We now discuss how to compare ground-truth and predicted nodes on a taxonomy. Let $\{(v_n^{\text{gt}}, v_n^{\text{pr}})\}_{n=1}^N$ be a set of pairs of nodes where v_n^{gt} is the ground-truth node for the n^{th} image and v_n^{pr} is the node mapped from the string generated from the VLM conditioned on the n^{th} image. (We describe a procedure to create this string–node mapping in §6.) We then define **hP** and **hR** over $\{(v_n^{\text{gt}}, v_n^{\text{pr}})\}_{n=1}^N$ as follows

$$\mathbf{hP} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} \frac{|\mathrm{anc}(v_n^{\text{pr}}) \cap \mathrm{anc}(v_n^{\text{gt}})|}{|\mathrm{anc}(v_n^{\text{pr}})|}, \tag{2a}$$

$$\mathbf{hR} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} \frac{|\mathrm{anc}(v_n^{\text{pr}}) \cap \mathrm{anc}(v_n^{\text{gt}})|}{|\mathrm{anc}(v_n^{\text{gt}})|}. \tag{2b}$$

The hierarchical F1 (**hF**) is then defined as the harmonic mean of **hP** and **hR**,

$$\mathbf{hF} \stackrel{\text{def}}{=} \frac{2 \cdot \mathbf{hP} \cdot \mathbf{hR}}{(\mathbf{hP} + \mathbf{hR})}. \tag{3}$$

Both **hP** and **hR** fall in the interval $[0, 1]$. The maximum is achieved when the target and predicted paths coincide. Less-than-perfect scores capture the extent to which the path deviates from the gold answer.

hP **Hierarchical precision** captures the amount of incorrect information in the prediction, i.e., deviations from the correct path, relative to the number of nodes shared with the correct path.

TRAIN — Prediction: *A mode of transport* — **hP**: 1.00 **hR**: 0.75

TELEPHOTO LENS — Prediction: *A lens cover* — **hP**: 0.80 **hR**: 0.67

PITH HELMET — Prediction: *Neckwear* — **hP**: 0.75 **hR**: 0.50

STIRLING CASTLE — Prediction: *Château de Ranrouët* — **hP**: 0.83 **hR**: 0.83

POOL — Prediction: *High jump* — **hP**: 0.67 **hR**: 0.67

Figure 3. **Illustrative examples.** Hierarchical precision (**hP**) and recall (**hR**) calculations on random label pairs (§5). While **hP** penalizes incorrect labels, i.e., labels that are further away from the target's ancestor set, **hR** penalizes mistakes made higher up on the taxonomy.

hR Hierarchical recall measures the amount of correct information in the prediction, i.e., how much the path from the root to the predicted node overlaps with the path from the root to the ground truth node. Hierarchical recall thus penalizes missing coverage of the target path, especially early deviations. This can be seen as measuring the specificity of the prediction, in particular when the **hP** is 1.

We give an example of the computation of **hP** and **hR** in Fig. 3, sampled from the OVEN datasets. In the case of the picture of TRAIN, we can see how only predicting MODE OF TRANSPORT leads to lower **hR** since it lacks information, while **hP** remains 1 since there is no incorrect information. For the sport POOL and STIRLING CASTLE, we notice how both scores drop, as incorrect nodes are predicted—these responses contain both incorrect information and some partially correct information. In the case of PITH HELMET, we see that the prediction is incorrect yet not very specific, resulting in a lower **hR** than **hP**. We also note that it is not possible to have a lower **hP** than **hR** if the ground truth nodes are leaf nodes.

4. Taxonomically Linked Datasets

To investigate taxonomy-aware evaluation measures for text generated from a VLM, we consider two taxonomies linked to FGVC tasks. First, the Catalogue of Life taxonomy as contained in the iNaturalist21 dataset [30] is a clean, expert-curated taxonomy that maps evolutionary relationships between species and contains exactly 10,000 leaf nodes. The second is a more general taxonomy constructed from the Wikidata [71] knowledge graph, which links entities from FGVC datasets in the Open-domain Visual Entity (OVEN) benchmark [31].

4.1. Extracting Taxonomies

While the iNaturalist21 dataset includes a built-in taxonomy, this is not the case for other FGVC datasets. For those included in the OVEN collection, we utilize Wikidata—a large-scale knowledge graph derived from Wikipedia [71]. Unlike taxonomies, knowledge graphs are not constrained to tree or even directed acyclic graph structures, which necessitates the extraction of a taxonomy that fits the formal definition given in §3. To construct such a tree from Wikidata, we rely on the `subclass of` relation. When multiple paths exist from the root to a given node in the taxonomy, we retain the longest path. In the case of ties, we sample one of the tied paths uniformly at random. To ensure a rooted directed tree, we exclude higher-level abstract classes that introduce cycles. Because OVEN already links its datasets (§4.2) to Wikipedia entries, we can directly map them to Wikidata entities and incorporate them into our taxonomy.

4.2. Entity Linked Datasets of Images

The iNaturalist21 dataset [30] was compiled and curated from the citizen science platform iNaturalist. It contains 2.7 million images spanning 10,000 species. The OVEN dataset [31] aggregates several FGVC datasets, linking them to Wikipedia identifiers and natural language questions to support the evaluation of open-domain visual entity recognition. It includes ImageNet21k-P [61, 63], iNaturalist2017 [29], Cars196 [36], SUN397 [76], Food101 [10], Sports100 [25], Aircraft [47], Oxford Flowers [49], and Google Landmarks v2 [56, 75]. We exclude iNaturalist2017 from our OVEN-based evaluations, as we separately analyze the newer iNaturalist21 dataset, which is aligned with the higher-quality Catalogue of Life taxonomy.

5. The Woes of Existing Measures

In this section, we evaluate whether non-taxonomic text similarity measures already encode sufficient taxonomic information, e.g., through label overlap or representational similarity. To investigate this, we conduct a targeted synthetic experiment comparing the taxonomic similarity measures (**hP** and **hR**), with textual and multimodal similarity measures from §2.2 commonly used to assess the quality of language model outputs. These measures are both string-based and representation-based.

	Measure	iNat21 τ-hP	iNat21 τ-hR	OVEN τ-hP	OVEN τ-hR
	EM	0.01	0.07	0.01	0.01
	Contained	0.02	0.09	0.02	0.03
string	ROUGE	0.18	0.41	0.17	0.21
string	METEOR	0.18	0.40	0.18	0.22
rep.	BERTscore	0.01	0.31	0.27	0.18
rep.	SentBERT	0.02	0.25	0.34	0.31
rep.	NLI	0.08	-0.20	-0.17	0.19
	CLIP-t2t	0.14	0.46	0.25	0.15
	CLIP-i2t	**0.35**	**0.49**	**0.35**	**0.34**

Table 1. **Correlation of text similarity measures with hierarchical metrics.** Kendall's τ correlations between taxonomic metrics **hP**, **hR**, and the different similarity measures on the synthetic evaluation sets. All correlations have a p-value < 0.0001.

Experimental Setup. We construct two synthetic evaluation sets to determine to what extent text similarity measures reflect taxonomic structure. We consider the taxonomies from iNaturalist21 and the Wikidata (see §4) to generate controlled pairs of labels corresponding to pairs of nodes in each taxonomy. A reference node v and a candidate node v' are chosen such that v is always a leaf node and v' is at some given distance from v. The distance between two nodes in the taxonomy is the length of the shortest undirected path between the two nodes in the taxonomy. To ensure a balanced dataset, we sample the node pairs such that their pairwise distance is uniformly distributed up to a maximum distance of 7, depending on the depth of the node. To evaluate **hP**, we sample 100,000 reference–candidate node pairs from the taxonomies. To evaluate **hR** in a controlled fashion, however, we additionally require that $v' \in \text{anc}(v)$, i.e., that $\textbf{hP} = 1$. This allows us to isolate the sensitivity of the similarity measures at decreasing levels of specificity.[1] For each pair, we compute both the taxonomy-aware score and text-based measures between the corresponding labels $\mathcal{L}(v)$ and $\mathcal{L}(v')$. We then measure the rank correlation between the taxonomy-aware score and various text-based measures using Kendall's τ, which evaluates whether the similarity measure reflects taxonomic proximity under controlled conditions. By construction, we use the node label $\mathcal{L}(v')$ as the prediction to avoid the problem of mapping unconstrained output onto the taxonomy. All of the measures we consider are text-based (see §10 in the supplement for details), except the **CLIP** image-to-text score, where we sample an image for the reference node from the corresponding dataset and measure **CLIP** model similarity between the image and the candidate label. We choose high-performing models that are not prohibitively large for the model-based comparisons.[2]

[1] Evaluating **hP** while controlling for **hR** is only possible for non-leaf reference nodes. Since our datasets always have leaf reference nodes, we did not explore this setting.

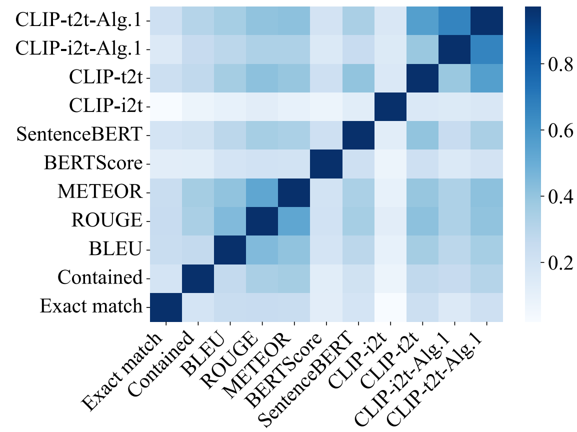

Figure 4. **Agreement in node placement using different similarity measures.** We observe variation in node placement for all measures. Only **METEOR** and **ROUGE**, and the **CLIP** variations when combined with Alg. 1 share predictions frequently.

Results. Table 1 presents correlations between **hP** and **hR** on one hand and the text similarity measures from §2.2 and **CLIP-i2t** on the other, for the iNaturalist21 and Wikidata labels. We observe that surface-level measures (**Exact Match (EM)**, **Contained**) exhibit a low correlation with the taxonomy-aware measures **hP** and **hR**. **ROUGE** and **METEOR**, which are based on n-gram overlap, capture taxonomic similarity more effectively. This is presumably due to label overlap, e.g., GRAY SEAL is a type of SEAL. The highest correlation is achieved by **CLIP** image-to-text (**CLIP-i2t**) for both datasets, followed by **CLIP** text-to-text (**CLIP-t2t**), except for **hP** on the iNaturalist21 dataset. In that case, **CLIP-t2t** fails to capture deviations from the correct taxonomic path. This suggests that **CLIP's** image features may encode more taxonomic information than its text representations, which can struggle due to the low-frequency vocabulary inherent to this task. **NLI** is an outlier, exhibiting a negative correlation in most cases. While NLI captures taxonomic similarity for closely related labels, it performs poorly on more distantly related pairs. This is visualized in Fig. 8 in the supplement.

Summary. Overall, we observe that representation-based measures capture some notions of taxonomic distance along the path from the leaf node to the root (**hR**), but are, for the most part, poorly equipped to measure taxonomic distances across different subtrees (**hP**).

[2] All models are downloaded from. huggingface.co, i.e., **BERTScore** (microsoft/deberta-large-mnli), **SentenceBERT** (sentence-transformers/all-mpnet-base-v2), **NLI** (MoritzLaurer/mDeBERTa-v3-base-mnli-xnli) and **CLIP** (apple/DFN5B-CLIP-ViT-H-14).

Measure	hP	hR	hF	Accuracy
Exact Match	0.37	0.42	0.39	17.5%
Contained	0.49	0.50	0.49	24.3%
BLEU	0.71	0.53	0.54	24.3%
METEOR	0.62	0.65	0.63	29.1%
ROUGE	0.62	0.66	0.63	30.0%
BERTScore	0.45	0.52	0.48	11.3%
SentBERT	0.58	0.67	0.62	20.0%
CLIP-t2t	0.72	0.79	0.75	30.3%
CLIP-i2t	0.58	0.71	0.23	5.3%
CLIP-i2t+Alg. 1	**0.80**	0.81	**0.80**	44.0%
CLIP-t2t+Alg. 1	0.79	**0.82**	**0.80**	**47.1%**

Table 2. **Evaluating taxonomy mapping with human annotation.** We report the performance of text similarity measures (directly) and CLIP image-to-text similarity for mapping VLM text predictions to the taxonomy using a subset of iNaturalist21 samples that were manually mapped to the correct nodes.

6. Mapping VLM Predictions to Taxonomies

The key challenge for taxonomy-aware evaluation of unconstrained text generated from VLMs is mapping the text generated by a VLM onto nodes in a given taxonomy. A baseline approach to this mapping is to use text similarity measures (e.g., those in §2.2) to calculate the best match between the generated text and the labeled nodes in a taxonomy. We also develop a heuristic approach that performs this mapping using a general similarity measure as a subroutine, which makes using **CLIP text-to-image** similarity feasible. Specifically, this algorithm combines string-based matching with scores from a given similarity measure to position predictions onto the taxonomy. We describe this procedure in words below and give pseudocode in §11 (**Alg. 1** supplement).

Our goal is to map each VLM-generated prediction string to a node in the taxonomy using a series of increasingly flexible matching strategies. Generally, we attempt exact matches first, then allow for partial overlaps or ambiguous cases, always preferring more specific nodes when multiple candidates are viable. Before matching, we normalize all predictions and node labels by lowercasing, replacing dashes with spaces, and stripping other punctuation. We then compute similarity scores between the prediction and all node labels using the **CLIP** similarity measures. More specifically, the mapping proceeds in stages. We first check whether any of the top-k ($k = 10$) highest-scoring nodes, according to the **CLIP** measure applied to their labels, appear verbatim in the prediction string; if so, we return the corresponding node. If this fails, we search for verbatim matches against all other labels in the taxonomy. If no full-string matches are found, we repeat these checks using n-gram overlap for $n \in \{4, 3, 2\}$. If multiple candidates remain with similar scores, we apply a fallback based on score ambiguity: if the difference between the top two scores, and between the top and k^{th} score, are both below threshold values, we treat the case as ambiguous. We then search for a common ancestor among the top candidates. If any node appears sufficiently often (at least four times) in the ancestor sets of multiple nodes in the top-k highest-scoring nodes we return the most specific such node, defined as the one deepest in the taxonomy. If none of the prior criteria are met, we fall back to returning the single highest-scoring candidate. Thus, the algorithm always returns a node in the taxonomy. If multiple nodes match the above criteria, we return the most specific one, i.e., the node deepest in the taxonomy. The "shared ancestor" step is particularly beneficial for the CLIP image-to-text similarities since we do not have images corresponding to higher-order taxonomic concepts.

6.1. Mapping Quality

To evaluate the quality of the taxonomic node proposals we measure their agreement against manually annotated node positions. We compare the similarity measures described in §2.2 (and Supplement §10) directly against all candidate node labels. We then do the same using Alg. 1, combined with the **CLIP** measures. To construct the test set, we gather 416 unique VLM outputs for the iNaturalist21 dataset, randomly sampled from the generations described in §7. For each output, we manually find the best matching taxonomic node given the string, e.g., matching *This is a bird* to AVES/BIRDS.

Results. Results are shown in Tab. 2. We find string-matching measures (**Exact Match** and **Contained**) do fairly poorly, while other surface-level measures (**BLEU**, **METEOR**, and **ROUGE**) perform slightly better. The representation-based approaches do not perform better in general, except for **CLIP text-to-text**. Fig. 4 shows that the string-level measures have considerable diversity in their predictions, while the representation-based measures show more diversity. The best performance is found with **CLIP text-to-text** and **CLIP image-to-text** using the algorithm described above (Alg. 1), which results in the same **hF** score but slightly improved exact node match accuracy score for **CLIP text-to-text + Alg. 1**.

7. Evaluating and Comparing VLMs

We compare one closed model and seven publicly available VLMs on fine-grained visual classification using our taxonomy-aware evaluation scheme. To map (unconstrained) strings generated from a VLM to nodes in the taxonomy, we use the best-performing similarity method (**CLIP-t2t** + Alg. 1), enabling the computation of **hP**, **hR**, and **hF**. These scores offer insight into the level of accuracy (**hP**) and specificity (**hR**) of VLM predictions for a given taxonomy. The resulting rankings differ significantly from

Model Name	# Parameters	Training Pipeline	OpenVLM Avg.
GPT-4 [50]	Unknown	Unknown	63.1
ILmXC2 [23]	7B	Pretr., instr.-tuning	62.2
OLM12B [79]	12B	Pretr., instr.-tuning, RLHF	54.6
QVLChat [3]	9.6B	Pretr., instr.-tuning	51.3
OLM3B [32]	3B	Pretr., instr.-tuning	47.8
LLaVA [45]	7.2B	Pretr., instr.-tuning	45.9
QwenVL [3]	9.6B	Pretr.	21.7
Fuyu [6]	8B	Pretr.	N/A

Table 3. **Overview of evaluated VLMs.** Details on the models we evaluate, including number of parameters, training method, and average performance on a large vision–language benchmark.

those based on standard text similarity measures, highlighting the taxonomy-aware measure's ability to capture previously unmeasured aspects of model performance.

Experimental Setup. We compare eight VLMs that differ in model size, training configuration (pretrained-only, instruction-tuned, and preference-aligned [51]), and performance on the Open VLM Leaderboard [16], a collection of 13 VL benchmarks with 38 models. We summarize model sizes, training configuration, and average performance on the Open VLM leaderboard in Table 3. For both the OVEN-sourced FGVC datasets and iNaturalist21, we randomly sample 6000 images from the validation set. For iNaturalist21, we use the prompt *What species is this?*, while for OVEN the question is included in the dataset and varies by object type. To understand how well we can control **hR** and **hP** for instruction-tuned models, we construct two different sets of instructions: (1) one containing a minimal instruction for answering in a specified format, and (2) one containing an instruction to be specific. We use slight variations of the prompt formatting to fit each model (as specified by the model developers); the exact prompts are provided in the supplement (Tab. 7 and Tab. 8). Because non-instruction-tuned models do not respond meaningfully to instructions, we prompt them with a straightforward question–answer template *Q:* {} *A:*. We use an orange-colored {} to indicate a variable in the template.

7.1. Experimental Results

How does the ranking of VLM performance change when considering taxonomy-aware measures? In Fig. 5, we rank eight VLMs based on their performance on FGVC according to the different similarity measures corresponding to the values given in Tab. 4 and Tab. 5 (§14 in supplement). We find that several text similarity measures yield similar rankings when used as direct evaluation criteria. In contrast, the taxonomy-aware measures lead to notable changes in model rankings. For instance, LLaVA ranks lowest under **Exact Match** for

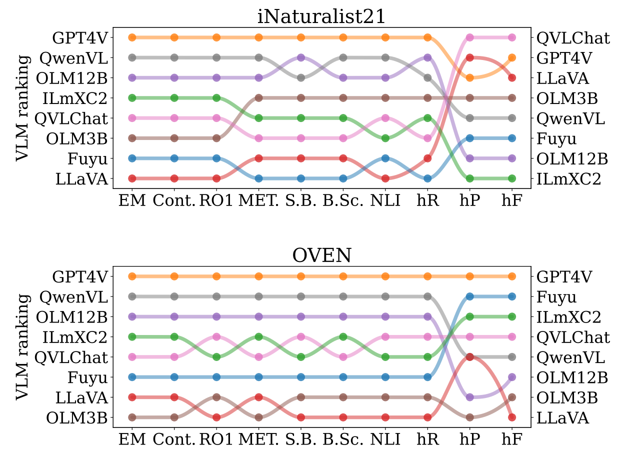

Figure 5. **Ranking of VLMs for the iNaturalist21 (top) and OVEN (bottom) datasets.** We evaluate the ranking of VLMs (vertical axis) based on different evaluation measures (horizontal axis). The best model is shown in the top row. On the left, we see the model names ranked by **exact match** and on the right ranked by the hierarchical F1 **hF**. See §14, Tab. 4 in supplement for exact numbers.

iNaturalist21, yet achieves high hierarchical precision (**hP**) and, consequently, a strong hierarchical F1 (**hF**). This indicates that LLaVA deviates from the correct taxonomy path less frequently than other models. However, its low hierarchical recall (**hR**) suggests that its predictions are often not specific. Another example is GPT-4, which ranks highest on all measures except **hP** and **hF** for iNaturalist21. Compared to QVLChat, which is ranked first in both **hP** and **hF**, we observe that GPT-4 makes more specific but erroneous predictions. Our taxonomy-aware measure offers a complementary perspective to traditional metrics by revealing performance differences that standard similarity-based measures overlook.

Do taxonomy-aware measures capture prompt sensitivity? We prompt the VLMs using two distinct prompts, as described earlier, to examine how their predictions shift in terms of hierarchical precision (**hP**) and recall (**hR**), shown jointly in Fig. 7. Fuyu and QwenVL are not instruction-tuned and therefore do not respond differently, while LLaVA shows only minimal changes. In contrast, the other models clearly adjust their predictions based on the different instructions in the prompts. On the iNaturalist21 dataset, all models except GPT-4 exhibit a trade-off between **hP** and **hR**: higher precision tends to come at the cost of recall. GPT-4, however, offers both more accurate and more fine-grained classifications when prompted appropriately. On the OVEN dataset, results are more mixed: roughly half of the models improve on both axes with prompt changes.

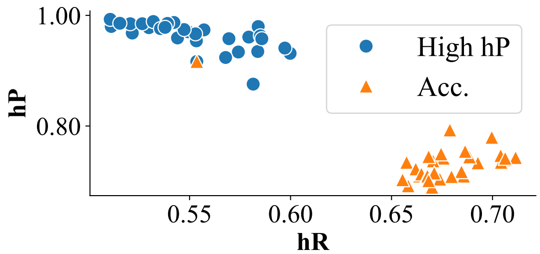

Figure 6. **Prompt tuning results for bird classifier.** For *High hP* (●), we target prompts that prioritize **hP**. For *Acc* (▲) we optimize for prompts that give higher binary accuracy. This demonstrates how **hP** and **hR** can help tune a VLM application.

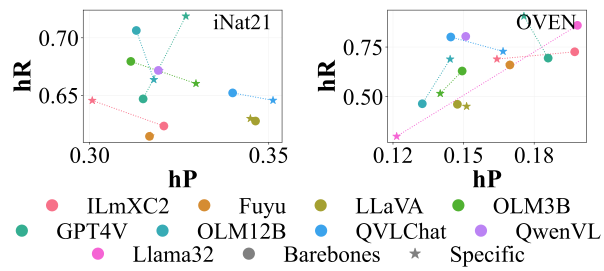

Figure 7. **Prompt sensitivity captured by hR and hP.** We compare two prompt versions, a minimal barebones version and one that asks for a taxonomically specific prediction.

These findings suggest that each application may prioritize different target values for **hP** and **hR**, and that models can adapt substantially to meet those needs. In such cases, prompt design becomes a powerful paradigm for tuning model behavior as further explored in the next section.

7.2. An Application: High-Precision Classification

To demonstrate the usefulness of **hR** and **hP** in tuning prompts for a specific application, we prompt Llama 3.2 Vision (11B) [65] to classify birds. The application goal is to minimize the amount of wrong information in the prediction, which is equivalent to maximizing hierarchical precision but hard to formulate in terms of exact match binary classification accuracy. We use prompt-tuning with **hP** and **hR** as feedback signals, and compare to the same procedure but with accuracy as the feedback signal. We start with a basic prompt *What is this an image of? Answer in the format 'A: <answer>.'* and task ChatGPT [50] to improve on the prompt for 30 iterations. A system prompt describes the scores and the goal of developing a conservative classifier that gives a less specific prediction if uncertain (see §15 in the supplement for more details). In each iteration, we include past prompts along with their resulting **hR** and **hP** scores, or the accuracy score, in a few-shot learning setup. We tune the prompts with 20 images and evaluate on 200 images of birds from iNaturalist21. The results are shown in Fig. 6. While this is a toy example, we see how the taxonomy-aware measures not only enable a higher granularity of model performance analysis, but can also facilitate targeted application development.

8. Discussion and Limitations

Evaluating unconstrained text generated from a VLM is inherently challenging: there are countless entities to describe and even more ways to describe them. As the computer vision community shifts from closed-set recognition to open-ended recognition, evaluation increasingly resembles an NLP task. In this paper, we introduce structure into VLM evaluation by leveraging curated taxonomies and established hierarchical evaluation measures. Despite this, several limitations remain. First, placing unconstrained text on a taxonomy is a low-resource problem; there are no large-scale datasets available to evaluate our mapping approach, nor sufficient data to train a dedicated classifier that links VLM outputs to taxonomy nodes. To address this, we repurpose existing text similarity measures, along with CLIP, to perform the mapping. While our evaluation shows that this approach is non-trivial and introduces errors that affect downstream VLM evaluation, we are optimistic about future improvements, such as training on synthetic data generated by language models. Second, taxonomies, especially those derived from Wikidata, are inherently noisy. Even expert-curated taxonomies like the one used for iNaturalist21 can have uneven subtree granularity, complicating interpretations based on globally averaged measures. One potential solution would be to employ expert input to assign weighted edges to reflect semantic distances between nodes.

9. Conclusion

Text similarity measures from NLP are a natural place to look for evaluation of unconstrained VLM output. Additionally, many FGVC tasks come with side information about the class structure, namely taxonomies, which text similarity measures do not fully capture, as we have shown in this paper. To solve this gap, we have proposed a framework for using hierarchical evaluation measures to evaluate VLMs for FGVC, enabling a more fine-grained view of the specificity and accuracy of these models.

Acknowledgements

Vésteinn Snæbjarnarson, Serge Belongie, Nico Lang, and Stella Frank are supported by the Pioneer Centre for AI, DNRF grant number P1. Niklas Stoehr is supported by the Swiss Data Science Center (SDSC) PhD Fellowship.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A visual language model for few-shot learning. In *Neural Information Processing Systems*, pages 23716–23736, 2022. 1

[2] Majed A Alkhamees, Mohammed A Alnuem, Saleh M Al-Saleem, and Abdulrakeeb M Al-Ssulami. A semantic metric for concepts similarity in knowledge graphs. *Journal of Information Science*, 49:778–791, 2023. 4

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv*, 2023. 7

[4] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. TouchStone: Evaluating vision-language models by language models. *arXiv*, 2308.16890, 2023. 2

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 3, 2

[6] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Tasirlar. Fuyu-8b: A multimodal architecture for AI agents, 2023. 1, 7

[7] Walter Bennette, Nathaniel Hofmann, Nathaniel Wilson, and Tyler Witter. Hierarchical open-set recognition for automatic target recognition. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021. 4

[8] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Conference on Computer Vision and Pattern Recognition*, pages 12506–12515, 2020. 4

[9] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. VisIT-Bench: A benchmark for vision-language instruction following inspired by real-world use. In *Neural Information Processing Systems*, 2023. 3

[10] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – Mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461, 2014. 4

[11] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing*, 2015. 3, 2

[12] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv*, 2023. 1

[13] Elijah Cole, Kimberly Wilber, Grant Van Horn, Xuan Yang, Marco Fornoni, Pietro Perona, Serge Belongie, Andrew Howard, and Oisin Mac Aodha. On label granularity and object localization. In *European Conference on Computer Vision*, pages 604–620, 2022. 4

[14] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Conference on Computer Vision and Pattern Recognition*, pages 14755–14764, 2022. 4

[15] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. In *Neural Information Processing Systems*, 2023. 2

[16] OpenCompass Contributors. OpenCompass: A universal evaluation platform for foundation models, 2023. 7

[17] Riccardo De Lutio, Yihang She, Stefano D'Aronco, Stefania Russo, Philipp Brun, Jan D Wegner, and Konrad Schindler. Digital taxonomist: Identifying plant species in community scientists' photographs. *ISPRS journal of photogrammetry and remote sensing*, 182:112–121, 2021. 4

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 2

[19] Jia Deng, Jonathan Krause, Alexander C. Berg, and Li Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 3450–3457, 2012. 4

[20] Xiwen Dengxiong and Yu Kong. Ancestor search: Generalized open set recognition via hyperbolic side information learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4003–4012, 2023. 4

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019. 2

[22] Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *Conference on Computer Vision and Pattern Recognition*, pages 836–837, 2020. 4

[23] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv*, 2401.16420, 2024. 7

[24] S. Doveh, A. Arbelle, S. Harary, E. Schwartz, R. Herzig, R. Giryes, R. Feris, R. Panda, S. Ullman, and L. Karlinsky. Teaching structured Vision & Language concepts to Vision & Language models. In *Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 4

[25] Kaggle (Gerry). 100 sports image classification, 2022. 4

[26] Simon Ging, Maria Alejandra Bravo, and Thomas Brox. Open-ended VQA benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. In *International Conference on Learning Representations*, 2024. 2

[27] Sherzod Hakimov and David Schlangen. Images in language space: Exploring the suitability of large language models for vision & language tasks. In *Findings of the Association for Computational Linguistics*, pages 14196–14210, 2023. 2

[28] Yuan He, Zhangdie Yuan, Jiaoyan Chen, and Ian Horrocks. Language models as hierarchy encoders. In *Neural Information Processing Systems*, 2024. 3

[29] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexx Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 4

[30] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin MacAodha. Benchmarking representation learning for natural world image collections. In *Conference on Computer Vision and Pattern Recognition*, pages 12879–12888, 2021. 2, 4

[31] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12065–12075, 2023. 1, 2, 3, 4

[32] Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, dahai li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *Conference on Language Modeling*, 2024. 7

[33] Svetlana Kiritchenko, Stan Matwin, Richard Nock, and Fazel Famili. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Canadian AI*, 2006. 2, 3

[34] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Conference of the European Association for Machine Translation*, pages 193–203, 2023. 2, 3

[35] Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: A unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3): 820–865, 2015. 3

[36] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *International Conference on Computer Vision Workshops*, pages 554–561, 2013. 4

[37] Suren Kumar and Rui Zheng. Hierarchical category detector for clothing recognition from visual data. In *International Conference on Computer Vision Workshops*, pages 2306–2312, 2017. 4

[38] Nico Lang, Vésteinn Snæbjarnarson, Elijah Cole, Oisin Mac Aodha, Christian Igel, and Serge Belongie. From coarse to fine-grained open-set recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 17804–17814, 2024. 4

[39] Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, and Honglak Lee. Hierarchical novelty detection for visual object recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 1034–1042, 2018. 4

[40] Clément Lefebvre and Niklas Stoehr. Rethinking the event coding pipeline with prompt entailment. In *EACL Workshop Fact Extraction and VERification*, 2023. 2

[41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 1

[42] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004. 1, 3, 2

[43] Dekang Lin. An information-theoretic definition of similarity. In *International Conference on Machine Learning*, page 296–304, 1998. 4

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 3

[45] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Conference on Computer Vision and Pattern Recognition*, 2023. 7

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Neural Information Processing Systems*, 2023. 1, 2, 3

[47] Subhransu Maji, Esa Rahtu, Juho Kannala, Mathew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv*, 1306.5151, 2013. 4

[48] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *Computer Vision and Pattern Recognition*, 2016. 4

[49] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 4

[50] OpenAI. GPT-4 technical report, 2023. 1, 2, 3, 7, 8

[51] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda

Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Neural Information Processing Systems*, 2022. 7

[52] Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. Large language models and knowledge graphs: Opportunities and challenges. *Transactions on Graph Data and Knowledge*, 2023. 3

[53] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting on Association for Computational Linguistics*, page 311–318, 2002. 3, 2

[54] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Conference on Empirical Methods in Natural Language Processing*, pages 2463–2473, 2019. 3

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 3

[56] Elias Ramzi, Nicolas Audebert, Clement Rambour, Andre Araujo, Xavier Bitot, and Nicolas Thome. Optimization of rank losses for image retrieval. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4

[57] Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees GM Snoek. SelEx: Self-expertise in fine-grained generalized category discovery. In *European Conference on Computer Vision*, 2024. 4

[58] Nils Reimers and Iryna Gurevych. Sentence-BERT: sentence embeddings using siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3980–3990, 2019. 1, 3, 2

[59] Philip Resnik. A class-based approach to lexical discovery. In *Annual Meeting of the Association for Computational Linguistics*, pages 327–329, 1992. 4

[60] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conference on Artificial Intelligence*, page 448–453, 1995. 4

[61] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 4

[62] Kevin Riehl, Michael Neunteufel, and Martin Hemberg. Hierarchical confusion matrix for classification performance evaluation. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(5):1394–1412, 2023. 3

[63] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 4

[64] Paolo Simeone, Raúl Santos-Rodríguez, Matt McVicar, Jefrey Lijffijt, and Tijl De Bie. Hierarchical novelty detection. In *Advances in Intelligent Data Analysis*, pages 310–321, 2017. 4

[65] Hugo Touvron, Aaron Grattafiori, Michael Auli, et al. The Llama 3 herd of models. *arXiv*, 2407.21783, 2024. 8

[66] Mehmet Ozgur Turkoglu, Stefano D'Aronco, Gregor Perich, Frank Liebisch, Constantin Streit, Konrad Schindler, and Jan Dirk Wegner. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment*, 264:112603, 2021. 4

[67] Grant Van Horn, Steve Branson, Scott Loarie, Serge Belongie, and Pietro Perona. Lean multiclass crowdsourcing. In *Conference on Computer Vision and Pattern Recognition*, 2018. 4

[68] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. 4

[69] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Conference on Computer Vision and Pattern Recognition*, 2015. 3

[70] Nakul Verma, Dhruv Mahajan, Sundararajan Sellamanickam, and Vinod Nair. Learning hierarchical similarity metrics. In *Conference on Computer Vision and Pattern Recognition*, pages 2280–2287, 2012. 4

[71] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. 2, 4

[72] Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. Evaluating open-QA evaluation. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2

[73] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *arXiv*, 2104.14690, 2021. 3, 2

[74] Xiu-Shen Wei, Yi-Zhen Song, Oisin Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(12):8927–8948, 2022. 1

[75] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jacjk Sim. Google Landmarks Dataset v2: A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Conference on Computer Vision and Pattern Recognition*, 2020. 4

[76] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 4

[77] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding, 2019. 3, 2

[78] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Fanqing Meng, Siyuan Huang, Meng Lei, Ping Luo, and Yu Qiao. LVLM-eHub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[79] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Conference on Computer Vision and Pattern Recognition*, 2024. 7

[80] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 3

[81] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision*, pages 11941–11952, 2023. 3

[82] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. 3, 2

[83] Bingchen Zhao, Nico Lang, Serge Belongie, and Oisin Mac Aodha. Labeled data selection for category discovery. In *European Conference on Computer Vision*, 2024. 4

[84] Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. VLUE: A multi-task multi-dimension benchmark for evaluating vision-language pre-training. In *International Conference on Machine Learning*, pages 27395–27411, 2022. 3

[85] Ganggao Zhu and Carlos A. Iglesias. Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85, 2017. 4

[86] Ganggao Zhu and Carlos A. Iglesias. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications*, 101:8–24, 2018. 4

[87] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, 2025. 2, 3