Free-Grained Hierarchical Visual Recognition

Seulki Park[1] Zilin Wang[1] Stella X. Yu[1,2]
[1]University of Michigan [2]UC Berkeley
{seulki,zilinwan,stellayu}@umich.edu

Abstract

Hierarchical image recognition seeks to predict class labels along a semantic taxonomy, from broad categories to specific ones, typically under the tidy assumption that every training image is fully annotated along its taxonomy path. Reality is messier: A distant bird may be labeled only bird, while a clear close-up may justify bald eagle.

We introduce free-grain training, where labels may appear at any level of the taxonomy and models must learn consistent hierarchical predictions from incomplete, mixed-granularity supervision. We build benchmark datasets with varying label granularity and show that existing hierarchical methods deteriorate sharply in this setting. To make up for missing supervision, we propose two simple solutions: One adds broad text-based supervision that captures visual attributes, and the other treats missing labels at specific taxonomy levels as a semi-supervised learning problem.

We also study free-grained inference, where the model chooses how deep to predict, returning a reliable coarse label when a fine-grained one is uncertain. Together, our task, datasets, and methods move hierarchical recognition closer to the way labels arise in the real world [1].

1. Introduction

Hierarchical classification [5, 7, 16, 26] predicts a **semantic tree** (*Bird* → *Bird of prey* → *Bald eagle*), capturing categories from broad to specific. Predicting the full hierarchy can improve robustness and scalability, encouraging models to generalize across levels and making it easier to extend taxonomies with new parent or child classes. It also supports flexible use: An expert may want *Bald eagle*, while a layman needs only *Bird*. Yet most existing methods [5, 41] assume *complete supervision*, where every training image is annotated along its full taxonomy path (Fig. 1.3).

Real annotations are rarely so tidy. Labels often appear at different levels of the taxonomy for two reasons. One is intrinsic: The image may not contain enough visual evidence to justify a fine-grained label. The other is extrinsic: annotation may be constrained by cost, expertise, or evolving labeling protocols [18, 23]. Thus a distant image or non-expert annotator may provide only a coarse label such

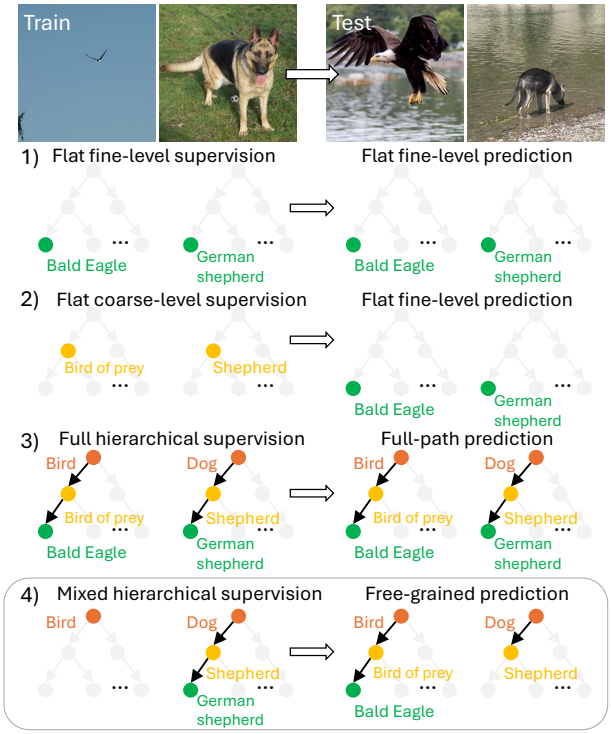


Figure 1. **We propose free-grained hierarchical recognition, where label granularity can freely vary across instances.** Both its training and inference differ from existing settings. **1) Flat fine-grained recognition**: fine labels → fine predictions (fully supervised) [46]. **2) Flat weakly supervised recognition**: coarse labels → fine predictions [13]. Both 1) and 2) operate on a single flat level without modeling cross-level relations. **3) Hierarchical recognition**: full hierarchy → full hierarchy, requiring complete annotations at all levels [26]. **4) Free-grained recognition**: Labels may appear at different levels during training, e.g., a distant bird as *Bird* and a close-up dog as *German shepherd*. At inference, the model predicts at an appropriate level of specificity. The task thus requires learning from mixed, incomplete supervision and predicting hierarchical labels at the right depth.

[1]Our dataset and code is available at FreeGrainLearning.

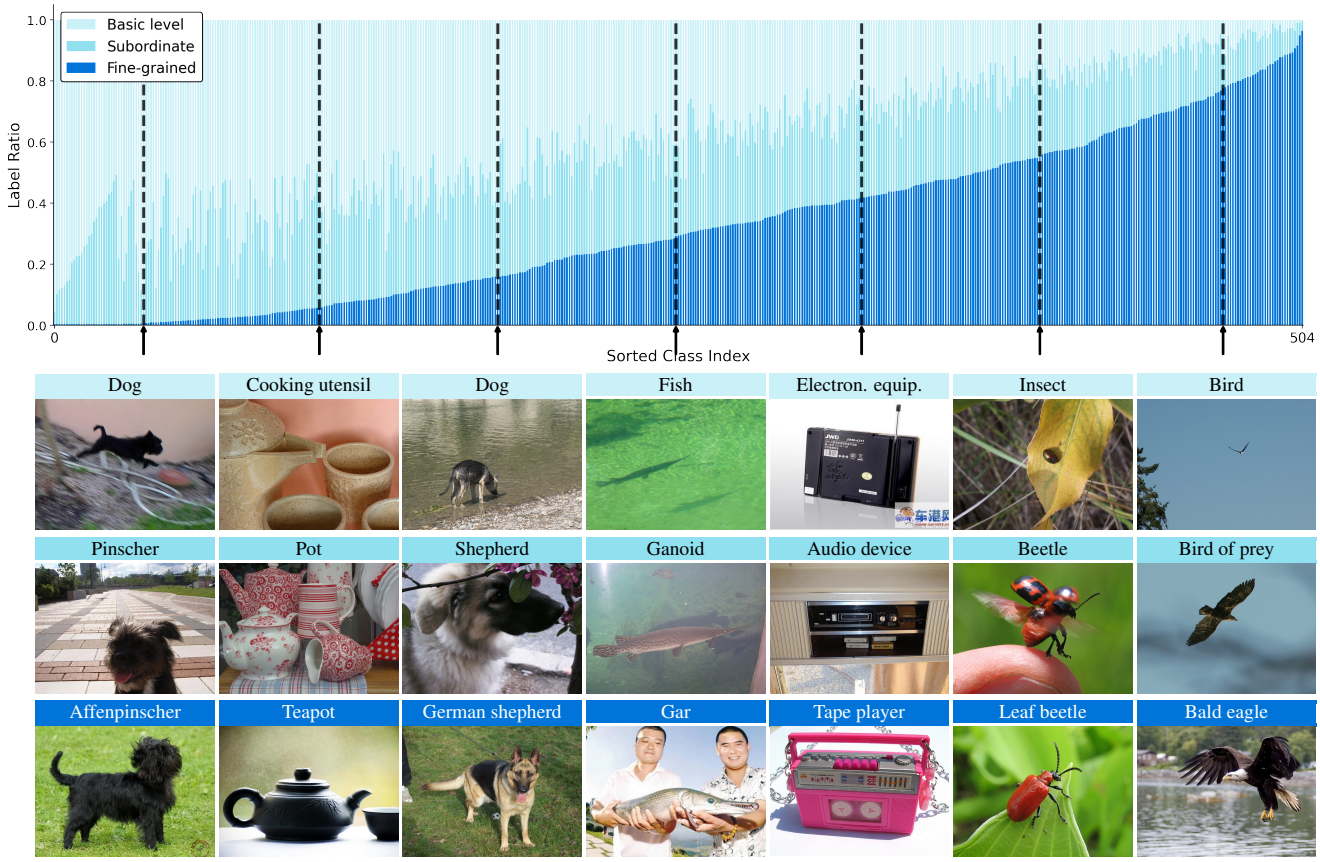


Figure 2. **Our ImageNet-F reflects realistic mixed-granularity supervision, exhibiting both long-tailed fine-grained labels and visual ambiguity. Top:** Distribution of label depth (basic, subordinate, fine) across classes. Fine-grained labels are highly imbalanced, forming a long-tailed pattern where some classes retain many fine labels while others have few. **Bottom:** Examples illustrating confidence-based label assignment using foundation models. (Last column) A distant bird is labeled at the basic level (Bird); a mid-range instance at the subordinate level (Bird of prey); and a clear close-up at the fine-grained level (Bald eagle).

as *Bird*, while a close-up or expert annotation may support a fine-grained label such as *German shepherd* (Fig. 1.4).

To capture this messier reality, we propose *free-grained hierarchical recognition*, where supervision is free to vary in granularity: training labels may appear at any level of a taxonomy, from coarse to fine, and may differ across instances. At inference, the model outputs a free-grained prediction, selecting the deepest label it can reliably support.

The challenge of free-grained learning is to learn from supervision that is *incomplete, uneven, and spread across levels*, while still producing predictions that remain consistent with the taxonomy. This setting differs from conventional ones, where labels are provided at a single level – either fine-grained in the fully supervised case or coarse in the weakly supervised case – without requiring the model to connect information across levels (Fig. 1.1-2).

To support this setting, we construct new benchmarks by adapting existing hierarchical datasets (CUB [43], Aircraft [22], iNat21-mini [39]) to exhibit mixed-granularity supervision. While these benchmarks provide valuable testbeds, they are limited in scale or diversity: CUB and Aircraft are small-scale, and iNat21-mini is confined to a single biological domain. Larger datasets such as ImageNet [32] inherit deep, inconsistent hierarchies from WordNet [10], and are therefore rarely used for prior hierarchical recognition [5, 7, 26]. We thus redesign ImageNet into a clean three-level tree for hierarchical recognition.

On top of these datasets, we construct two complementary variants to capture realistic annotation difficulty and varying label availability. **1) Foundation-based variants** (ImageNet-F, iNat21-mini-F, CUB-F) set label depth by whether large visual foundation models [28, 35] correctly predict each level. Though imperfect as a proxy for human annotation, they provide a practical approximation: Deeper-level errors often coincide with visual ambiguity, yielding realistic mixed-granularity supervision. Thus distant birds may be labeled as *Bird*, mid-range ones as *Bird of prey*, and clear close-ups as *Bald eagle* (Fig. 2). This also induces long-tailed label availability, since fine-grained labels are removed more often for some classes than for others.

2) Randomized variants (CUB-Rand, Aircraft-Rand) assign label depths at varying proportions, enabling systematic evaluation under different levels of label availability. Together, these variants span diverse mixed-granularity scenarios and pose a broad benchmark for free-grain learning.

Our setting is hard for existing hierarchical classifiers. When trained under free-grained supervision, state-of-the-art methods [7, 26] lose up to **40%** in full-path accuracy on iNat21-mini [39], where a prediction counts as correct only if every level of the taxonomy is correct. This sharp deterioration underscores the difficulty of learning from mixed-granularity labels and the need for methods that can handle incomplete supervision more robustly.

To address this, we propose two *free-grained training* methods that compensate for missing supervision in different ways. **1) Text-guided pseudo-attributes** add auxiliary text supervision in the form of image descriptions generated by a vision–language model [9], providing semantic cues about visual attributes that help the model learn discriminative features even when fine-grained labels are absent. **2) Taxonomy-guided semi-supervised learning** (SSL) instead treats missing labels at particular taxonomy levels as unlabeled data, using hierarchical consistency to learn from both labeled and unlabeled examples. Across datasets, each method improves over state-of-the-art hierarchical classifiers by 5–25%, providing strong baselines while also underscoring the remaining difficulty of free-grained learning.

We also study *free-grained inference*, in which the model adaptively chooses how deep to predict. This is motivated by a simple practical point: A correct coarse prediction is often preferable to an incorrect fine-grained one. We consider two strategies: **1) confidence-based**, selecting the deepest label with sufficient confidence, and **2) consistency-based**, selecting the deepest level that maintains hierarchical consistency. We find that the latter yields more reliable and deeper correct predictions.

Contributions. 1) We introduce free-grained hierarchical visual recognition, where training labels may appear at any taxonomy level and inference adaptively chooses prediction depth. **2)** We build benchmark datasets for this setting, including foundation-based variants for realistic annotation difficulty and randomized variants for controlled label availability. **3)** We propose text-guided pseudo-attributes and taxonomy-guided SSL, both of which outperform prior hierarchical classifiers under mixed-granularity supervision. **4)** We show that, for free-grained inference, consistency-based inference yields more reliable and deeper correct predictions than confidence-based inference.

2. Related Work

Hierarchical classification predicts the full taxonomy path for each image, requiring accurate level-wise predictions while also encouraging parent–child consistency across the hierarchy [5, 7, 26, 41]. Meanwhile, some methods use the hierarchy as an auxiliary signal for flat (fine-grained) classification, for example, to regularize feature learning [49] or to reduce the severity of fine-grained mistakes [12, 17]. Importantly, all these approaches *assume complete hierarchical supervision* is available for every training sample.

Long-tailed and semi-/weakly-supervised recognition address distinct real-world challenges [20, 25, 30, 44], typically operating at a *single* granularity, using either fine-grained labels alone or coarse labels alone. In contrast, our free-grained learning introduces a new, unexplored problem: learning from *mixed-granularity* labels within a hierarchy. This naturally brings together challenges from multiple areas, including class imbalance within each level, imbalance across different levels of the hierarchy, weak/semi-supervision, and the need to maintain hierarchical consistency, all within a single unified framework. Unlike prior work that addresses these challenges in isolation, our setting requires handling them jointly under mixed-granularity supervision (Table 1). See more related works in Appendix I.

Table 1. **Our task is more practical and challenging.** Free-grained learning reflects real-world annotation, where each image may have fine (F) or coarse (C) labels, and models must predict a taxonomy-consistent hierarchy. It jointly introduces class imbalance (Cls. Imb.) and level imbalance (Lvl. Imb.), along with weak and partial supervision—factors mostly studied in isolation. Evaluation considers both accuracy (Acc.) and consistency (Con.).

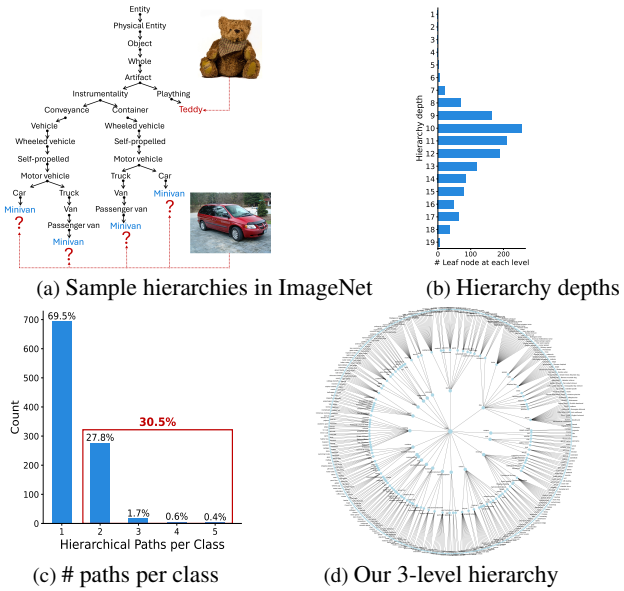
	Input		Output		Labels	Imbalance		Metrics	
Tasks	F.	C.	F.	C.	Avail.	Cls.	Lvl.	Acc.	Con.
Long-tailed recog.	✓	✗	✓	✗	All	✓	✗	✓	✗
Semi-supervised	✓	✗	✓	✗	*Partial*	✗	✗	✓	✗
Weakly-supervised	✗	✓	✓	✗	All	✗	✗	✓	✗
Hierarchical recog.	✓	✓	✓	✓	All	✗	✗	✓	✓
Free-grained recog.	✓	✓	✓	✓	*Partial*	✓	✓	✓	✓

3. Benchmarks for Free-Grained Recognition

We adapt existing hierarchical benchmarks to our free-grained setting, but prior datasets are often small-scale (e.g., CUB, Aircraft) or domain-specific (e.g., iNat21-mini), as shown in Table 2. To enable a large-scale and diverse benchmark, we reorganize ImageNet into a clean three-level hierarchy (ImageNet-3L), as its original WordNet taxonomy is irregular (Fig. 3). This simplified structure supports mixed-granularity prediction without unnecessary

Table 2. **We convert existing hierarchical benchmarks into free-grain versions.** Since ImageNet's taxonomy is inconsistent, we newly curate a consistent three-level hierarchy, ImageNet-3L.

Dataset	**#levels**	**#classes per level**	**#train**	**#test**
CUB	3	13-38-200	5,994	5,794
Aircraft	3	30-70-100	6,667	3,333
iNat21-mini	8	3-11-13-51-273-1103-4884-10000	500,000	100,000
ImageNet	5-19	- 1000	1,281,167	50,000
ImageNet-3L	**3**	**20-127-505**	**645,480**	**25,250**



(a) Sample hierarchies in ImageNet (b) Hierarchy depths

(c) # paths per class (d) Our 3-level hierarchy

Figure 3. **We curate ImageNet-3L as a benchmark for hierarchical classification. (a)** Sample hierarchies reveal two issues: 1) some classes have multiple valid paths (e.g., *Minivan*), and 2) classes at the same depth can have mismatched specificity (e.g., *Teddy* vs. *Conveyance*). **(b)** ImageNet classes span widely varying depths (5–19 levels), often exceeding 10, highlighting inconsistency in hierarchy depth. **(c)** 30% of classes have multiple valid hierarchical paths, introducing ambiguity in evaluation. **(d)** We construct a coherent 3-level taxonomy, inspired by cognitive psychology [31]: *basic* for general recognition, *subordinate* for contextual specificity, and *fine-grained* for specialized distinctions.

complexity. We first describe this restructuring (Sec. 3.1), then introduce foundation-based variants that mimic real-world annotation patterns (Sec. 3.2) and randomized variants for controlled evaluation (Sec. 3.3).

3.1. Constructing ImageNet-3L

As in Fig. 3, the WordNet hierarchy [10] is noisy and inconsistent, making ImageNet unsuitable for full-path evaluation. To address this, we simplify the WordNet hierarchy by removing overly abstract nodes (e.g., *Entity*, *Whole*) and restructuring it into a consistent three-level taxonomy. This design is guided by categorization principles [31], where the *basic* level is the most natural and visually distinctive.

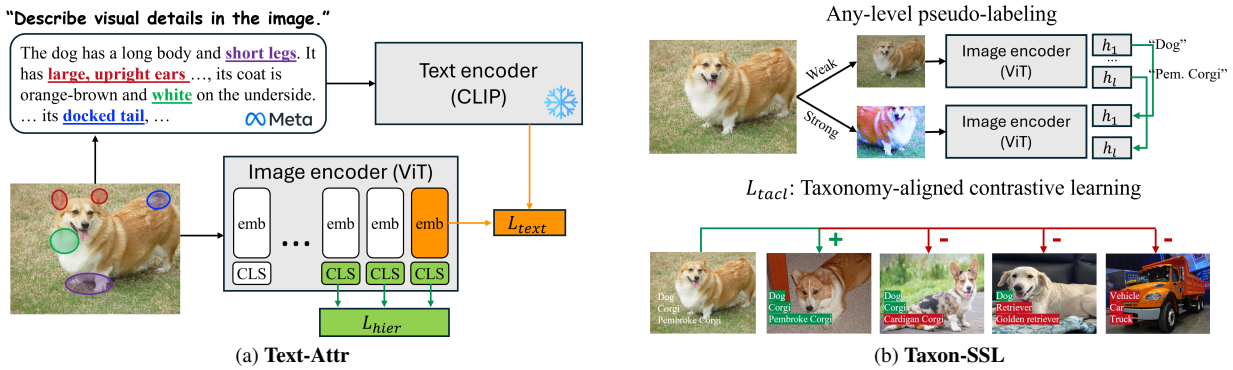
We anchor categories at a shared *basic* level (e.g., *vehicle, dog*) and organize subordinate and fine-grained categories under it. When multiple candidates exist, we select those that best match the granularity of existing basic classes. We further apply additional design principles, described below, with full details provided in the appendix A.
1) *Enforce meaningful structure:* We remove paths where each node has only one child, since coarse labels fully determine the fine labels. Branches with fewer than three levels are also excluded. **2) *Maximize within-group diversity:*** Among subordinate candidates under each basic class, we favor those with richer fine-grained subclasses, for example choosing *parrot* (4 children) over *cockatoo* (1 child). **3) *Refine vague categories:*** Ambiguous groups such as *Women's Clothing* are reorganized into precise, functionally grounded categories (e.g., *Underwear*) to improve clarity. **4) *Validate with language models and human review:*** We use large language models (ChatGPT [1]) to suggest refinements, with all decisions manually reviewed for semantic consistency. Applying this curation process to ImageNet-1k yields a structured benchmark of 20 basic, 127 subordinate, and 505 fine-grained classes, ImageNet-3L, ensuring every branch supports meaningful hierarchical prediction (a complete list is provided in Appendix B).

3.2. Foundation-based Pruning

To build realistic free-grain training sets, we prune hierarchical labels using large vision–language models: CLIP [28] for ImageNet-F and BioCLIP [35] for iNat21-mini-F and CUB-F. While these models are not designed to measure ambiguity, their zero-shot confidences indicate when fine-grained labels are less reliable, whether due to limited visual detail, annotation difficulty, or inconsistent expertise. Fig. 2 shows that this results in mixed-granularity supervision patterns that often follow such visual ambiguity, with distant or less discernible instances labeled more coarsely and clearer ones labeled more finely.

We adopt CLIP's prompt-ensemble strategy (e.g., *a photo of a [class]*) and assign labels from coarse to fine based on prediction correctness: **(1)** We always retain the *basic* label. **(2)** If the subordinate prediction is correct, we retain the *subordinate* label. **(3)** If both subordinate and fine-grained predictions are correct, we retain the *fine-grained* label. This defines the deepest available label for each image, with higher-level labels assumed from the given taxonomy. Since relying solely on foundation models' predictions can produce biased label distributions, we further remove a portion of subordinate labels based on the fine-grained removal rate per class, introducing more challenging supervision.

This pruning only removes labels and introduces no additional semantic information; it may even increase difficulty by discarding labels of harder examples. While we use CLIP and BioCLIP, similar pruning can be performed with other foundation models or ensembles (e.g., removing labels consistently mispredicted across models).
1) ImageNet-F. After pruning, 32.6% of images retain all three levels (Basic + Subordinate + Fine-grained), 28.0% retain two (Basic + Subordinate), and 39.4% retain only the Basic. Each class keeps the same number of images as in ImageNet; imbalance arises only from label granularity.
2) iNat21-mini-F. BioCLIP, a biology foundation model, performs well on species-level prediction but struggles with coarser labels. This mismatch enables substantial pruning:



(a) **Text-Attr** (b) **Taxon-SSL**

Figure 4. **Overview of the proposed methods.** (a) Text-Attr enriches feature representations using semantic cues from images, compensating for missing labels and capturing shared attributes across levels. (b) Taxon-SSL handles missing-level labels by treating them as unlabeled and learns from visual consistency through augmented views. Both methods offer distinct benefits for our challenging task.

22.5% of images retain all three levels (Order + Family + Species), 28.0% retain two, and 49.5% retain only Order.
3) CUB-F. With the same procedure, 31.5% of images keep 3 levels, 23.3% two (Order, Family), 45.2% only Order.

3.3. Randomized Pruning

To control label availability, we construct randomized variants, CUB-Rand and Aircraft-Rand, by randomly pruning labels from CUB [43] and Aircraft [22]. Unlike realistic pruning, this design systematically varies supervision and simulates *extreme* sparsity (e.g., only 10% fine-grained labels), enabling stress-testing of model robustness across diverse label distributions. Although random removal is independent of image difficulty, it reflects practical factors such as annotator expertise, cost, or task-specific constraints. We denote availability as a-b-c, where a% of basic, b% of subordinate, and c% of fine-grained labels are retained (e.g., 100-50-10 retains 10% fine-grained labels and 40% subordinate-only labels).

4. Free-Grain Learning Methods

We define our free-grained hierarchical recognition problem and introduce two ways to handle missing supervision.

4.1. Problem Setup

In free-grained hierarchical classification, the goal is to predict labels across all levels of a taxonomy from training data with mixed granularity. Each image is annotated at a certain level, and all coarser labels are assumed to be available while finer ones are missing; the coarsest label is always given. The model is trained to produce consistent predictions across the full hierarchy.

Free-grained Hierarchical Loss. To adapt prior hierarchical recognition methods to the free-grained setting, we modify their hierarchical supervision by applying the loss only at levels with available labels. Given hierarchical labels $y_1, \ldots, y_L$ across L levels, the loss is defined as:

$$\mathcal{L}_{\text{hier}} = \sum_{l=1}^{L} \mathbb{1}_{\{y_l \text{ exists}\}} \cdot \mathcal{L}(f_l(x), y_l), \quad (1)$$

where $f_l(x)$ is the prediction at level l, and $\mathcal{L}$ denotes a classification loss (e.g., cross-entropy).

4.2. Semantic Guidance with Text Attributes

Our semantic guidance approach is motivated by the observation that while class labels differ across hierarchical levels (e.g., *Dog* → *Corgi* → *Pembroke*), many visual attributes, such as "tail length" or "ear shape", remain consistent (Fig. 4a). To capture these shared semantic cues, we use image descriptions as auxiliary supervision. While recent large language models (LLM) (e.g., ChatGPT [1])-based approaches such as FineR [19] also use vision-language model (VLM)-generated text, their purpose is different: they feed these cues into an LLM for training-free fine-grained class reasoning, whereas we use text as supervision to train image representations to capture visual attributes shared across hierarchical levels.

Specifically, given an input image x, we use a frozen vision-language model (VLM), Llama-3.2-11B [9], to generate a language description d_x, using the prompt: "*Describe visual details in the image.*" This produces descriptions containing phrases such as "short legs" or "pointed ears," which we encode into a text embedding z_x^t using CLIP's text encoder [28]. We cap generation at 100 tokens, while CLIP accepts 77 tokens; longer descriptions are truncated during encoding. Although truncation discards some details, our method focuses on shared semantic cues (e.g., "short legs," "brown markings") rather than exhaustive captions, making it robust to this limitation. In parallel, we obtain the image embedding z_x^v from the image encoder, and align it with the text embedding z_x^t using a contrastive loss:

$$\mathcal{L}_{\text{text}} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(\text{sim}(z_i^v, z_i^t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_i^v, z_j^t)/\tau)} \right), \quad (2)$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity and τ is a temperature parameter. This loss guides the encoder to capture salient, label-independent traits shared across levels. Although not explicitly predicting attributes, aligning image features with text induces intermediate representations, which we call pseudo-attributes. This model-agnostic method can be applied to any architecture.

4.3. Visual Guidance with SSL

We adopt a semi-supervised formulation because missing-grain labels can be treated as unlabeled data. CHMatch [44] shows that coarse labels can improve pseudo-labeling, but it is limited to a two-level (coarse–fine) setting and focuses on refining fine-grained predictions. We generalize this to arbitrary multi-level taxonomies by 1) generating pseudo-labels at *every level*, and 2) enforcing *cross-level consistency* so that predictions remain valid along the hierarchy.

1) Multi-level pseudo-labeling. Following CHMatch, we decouple the classifier f into a shared feature extractor f_{feat} and level-specific heads $\{h_l\}_{l\in\mathcal{S}_x}$, where each head predicts labels at a different taxonomy level. The supervised loss is computed using Eq. 1, applying supervision only at levels with available labels. Pseudo-labels at each level are generated from the predictions of the corresponding head given a weakly augmented input $W(x)$.

2) Taxonomy-aligned feature learning. A key challenge is that pseudo-labels at different levels may be inconsistent (e.g., two samples share a coarse label but differ at fine levels). To address this, we only treat pairs as *reliable positives* when they agree across *all levels*.

For each mini-batch, we build level-wise affinity graphs W^l based on pseudo-label agreement: $W^l_{ij} = 1$ if images i and j share the same pseudo-label at level l, and 0 otherwise. We then define a taxonomy-aligned affinity:

$$W_{ij} = \begin{cases} 1 & \text{if } W^1_{ij} = \cdots = W^L_{ij} = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

This enforces that two samples are considered similar *only if they are consistent across the entire hierarchy*, effectively filtering out noisy or partially incorrect pseudo-labels.

Contrastive objective. Using W, we pull together positive pairs ($W_{ij} = 1$) and push apart negative pairs ($W_{ij} = 0$). The taxonomy-aligned contrastive loss is:

$$\mathcal{L}_{\text{tacl}} = -\frac{1}{\sum_j W_{ij}} \log \frac{\sum_j W_{ij} \exp\left((g(f(x_i)) \cdot g(f(x_j))')/t\right)}{\sum_j (1 - W_{ij}) \exp\left((g(f(x_i)) \cdot g(f(x_j))')/t\right)}, \tag{4}$$

where $g(f(x_i))$ is the projected feature of image i, and t is a temperature. This objective encourages samples with consistent hierarchical semantics to form tight clusters in the feature space, while separating those that disagree at any level. See more details in the Appendix H.

5. Experiments

In this section, we first describe the experimental setup (Sec. 5.1). We then present results on our free-grain benchmarks (Sec. 5.2) and provide further analysis of the proposed methods (Sec. 5.3). Finally, we compare free-grained inference methods (Sec. 5.4).

5.1. Experimental Setup

1) Dataset: We conduct experiments using our proposed *ImageNet-F*, *iNat21-mini-F*, and *CUB-F* datasets, along with the synthetic *CUB-Rand* and *Aircraft-Rand* datasets.

2) Evaluation metrics: Following [26], we evaluate accuracy and consistency: *(1) Level-accuracy*: Top-1 accuracy at each level. *(2) Tree-based InConsistency Error rate (TICE)*: Proportion of samples with inconsistent predictions in the hierarchy (lower is better): TICE $= \frac{n_{\text{ic}}}{N}$, where N is the total number of samples and n_{ic} is the number of inconsistent predictions. *(3) Full-Path Accuracy (FPA)*: Proportion of samples correctly predicted at all levels (primary metric): FPA $= \frac{n_{\text{ac}}}{N}$, where n_{ac} is the number of samples correct at all hierarchy levels.

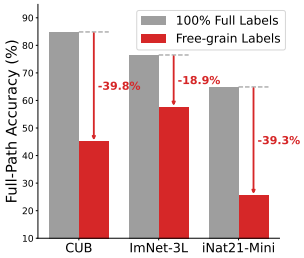
3) Comparison Methods: We adapt two strong and relevant hierarchical classifiers to the free-grained setting for comparison. *(1) Hierarchical Residual Network (HRN)* [7]: the first to handle supervision at both subordinate and fine-grained levels by maximizing marginal probabilities within the tree-constrained space. *(2) H-CAST* [26]: the current SOTA, encouraging consistent visual grouping across taxonomy levels. Originally trained with full supervision, we adapt it to this setting via the level-wise loss in Eq. 1, using only available labels.

4) Implementation: We use H-ViT, a ViT-Small-based hierarchical classifier, as the backbone for evaluating both Text-Attr and Taxon-SSL. To evaluate its compatibility across architectures, we also apply Text-Attr to H-CAST [26], a state-of-the-art hierarchical model with comparable capacity. HRN [7] is evaluated with its original ResNet-50 backbone, which has over twice the parameters. All models are trained for 100 epochs, except for ImageNet-F, which is trained for 200 due to its larger scale. Full architectural and training details are in the appendix G.

5.2. Benchmarking Results

Result 1: Performance drop under free-grained learning. The prior hierarchical SOTA, H-CAST, degrades sharply under mixed-granularity labels on both CUB and iNat21-mini. Fig. 5a shows that the full-path accuracy drops from 84.9% to 45.1% on CUB-F and from 64.9% to 25.6% on iNat21-mini-F. This highlights the challenge of mixed-granularity labels and imbalanced supervision across the hierarchy, and the need for methods that handle them.

Result 2: Performance on ImageNet-F. As shown in Ta-



(a) **H-CAST: full vs. free-grain**

Dataset	ImageNet-F (20-127-505)					iNat21-mini-F (273 - 1,103 - 10,000)				
	FPA(↑)	fine.(↑)	sub.(↑)	basic(↑)	TICE(↓)	FPA(↑)	spec.(↑)	fam.(↑)	order(↑)	TICE(↓)
HRN [7]	37.79	38.73	55.73	78.65	46.69	17.03	25.43	46.51	70.20	53.81
H-CAST [26]	57.59	59.02	82.69	93.53	21.81	25.63	28.61	67.20	83.62	47.17
Taxon-SSL	48.40	52.34	65.74	82.96	19.87	31.74	**37.11**	69.53	82.02	37.31
Taxon-SSL + Text-Attr	49.65	53.43	66.43	83.56	18.81	**31.93**	37.08	69.76	82.20	**37.04**
Text-Attr (H-ViT)	55.48	59.05	77.95	89.45	24.02	27.88	32.07	68.27	80.49	46.35
Text-Attr (H-CAST)	**63.20**	**64.91**	**84.47**	**93.56**	**18.58**	29.74	32.37	**71.79**	**85.99**	44.63

(b) **Method comparison on ImageNet-F and iNat21-mini-F.**

Figure 5. **(a) Transitioning from fully labeled data to our free-grain setting results in a substantial drop in Full-Path Accuracy, highlighting the difficulty of the task.** SOTA H-CAST drops by 19–40 percentage points across datasets. **(b) Our methods effectively improve performance under free-grain supervision, with behavior depending on data characteristics.** Conventional hierarchical methods such as HRN [7] and H-CAST [26] degrade significantly under incomplete supervision. In contrast, Text-Attr (H-CAST) performs strongly on ImageNet-F, where rich visual cues support text-guided learning, while Taxon-SSL is more effective on iNat21-mini-F, where fine-grained classes have similar appearances. Combining both (Taxon-SSL + Text-Attr) yields consistent but modest gains across datasets.

ble 5b, existing hierarchical methods degrade sharply under free-grained learning: HRN reaches only 37.8% FPA, while H-CAST performs better at 57.6% but still struggles with missing labels. Text-Attr (H-ViT) achieves 55.5% without relying on H-CAST's visual grouping, and integrating it into H-CAST further improves performance to 63.2%, demonstrating the effectiveness of semantic-guided pseudo-attribute learning at scale. Taxon-SSL improves over HRN by leveraging visual guidance but remains less effective than Text-Attr methods, whose strong performance benefits from the abundance and diversity of ImageNet-F for reliable visual–semantic alignment.

Result 3: Performance on iNat21-mini-F. In Table 5b, on the large-scale iNat21-mini-F dataset, which contains many classes (10,000), conventional hierarchical methods perform poorly (17.0% for HRN, 25.63% for H-CAST). Taxon-SSL achieves the best performance (31.9% FPA), highlighting the benefits of structural label propagation under limited per-class supervision. Text-Attr methods perform slightly lower (27.9–30.0% FPA), likely due to restricted textual diversity in this fine-grained biological domain, yet still outperform conventional baselines.

Additional results and ablations. We report additional results on CUB-F (Sec. D.1) and randomized variants with varying (limited) label availability (Sec. D.2). Across these settings, conventional hierarchical methods degrade under mixed-granularity supervision, while our approaches remain effective and robust. We further evaluate robustness under the original (unrefined) WordNet hierarchy, which exhibits irregular depth and inconsistent granularity (Sec. C). In addition, we conduct ablations on text encoders, Text-Attr features, training strategies, and architecture design (Sec. F), validating each component's contribution.

5.3. Further Analysis

How do methods behave with varying label availability? Text-attr excels with sparse labels, Taxon-SSL with moderate label availability. We analyze class-wise performance under imbalanced fine-grained label availability on ImageNet-F. To isolate effects, we compare Text-Attr (H-ViT) and Taxon-SSL with identical ViT-small backbones, excluding H-CAST modules. Fig. 6 plots per-class accuracy, where the x-axis shows Text-Attr performance and the y-axis shows Taxon-SSL performance; the diagonal indicates equal performance. Text-Attr (H-ViT) tends to outperform in label-scarce classes, appearing below the diagonal, by leveraging textual descriptions as additional supervision, while Taxon-SSL performs better for classes with more training samples, appearing above the diagonal by propagating consistency across missing levels. We provide additional t-SNE [21] analysis in Appendix E.

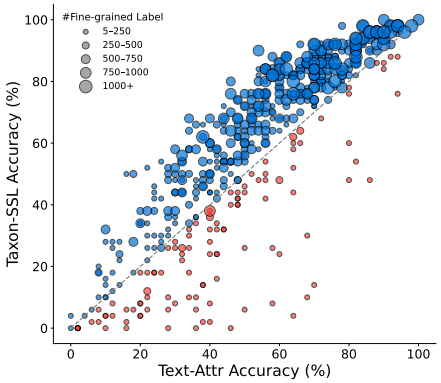


Figure 6. **Text-Attr is more effective under scarce fine-grained supervision, while Taxon-SSL performs better with more training data.** Each circle represents a class in ImageNet-F, with Text-Attr fine-grained accuracy on the x-axis and Taxon-SSL accuracy on the y-axis. The diagonal marks equal performance: points below (●) favor Text-Attr, and points above (●) favor Taxon-SSL. Circle size indicates the number of available training samples per class. Smaller circles tend to lie below the diagonal, showing the advantage of Text-Attr under limited data by leveraging textual guidance, whereas larger circles more often lie above it, indicating that Taxon-SSL benefits from richer supervision.

How does external semantic guidance help? External se-

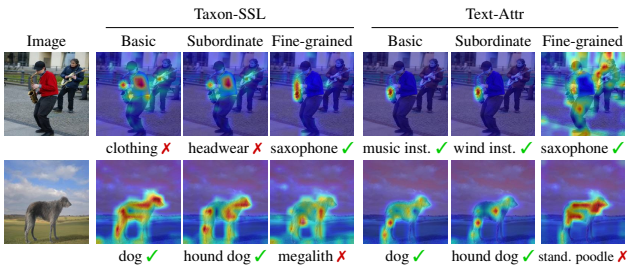


Figure 7. **Text-Attr improves semantic focus under diverse large-scale data.** (**Row 1**) In a multi-object image, Taxon-SSL assigns inconsistent labels ("*clothing*" at the basic level, "*saxophone*" at the fine-grained level), while Text-Attr (H-ViT) correctly predicts "*musical instrument*" by focusing on the relevant object. (**Row 2**) When both fail at the fine-grained level, Taxon-SSL outputs an unrelated class ("*megalith*"), whereas Text-Attr (H-ViT) chooses a semantically closer one ("*poodle*"). This shows that text-derived attributes help the model attend to meaningful regions and maintain semantic plausibility, on a large-scale ImageNet-F dataset with diverse categories and sparse labels. Green/Red denote correct/incorrect predictions.

mantic guidance helps the model attend to semantically relevant features and improves hierarchical consistency. To assess this effect, we compare saliency maps [6] from Taxon-SSL and Text-Attr (H-ViT) (Fig. 7). In Row 1, with multiple objects, Taxon-SSL focuses on a human shoulder and misclassifies the image, violating the hierarchy, while Text-Attr attends to the instrument and predicts correctly. In Row 2, when both fail at the fine-grained level, Taxon-SSL predicts an unrelated class, whereas Text-Attr selects a visually similar dog by focusing on curly fur and body shape. These results show that text-derived semantic cues guide attention toward meaningful features across label granularities, while Taxon-SSL may drift to visually salient but semantically irrelevant regions under sparse or ambiguous supervision.

5.4. Free-grained Inference

Free-grained inference matters in practice, since a correct coarse label is often preferable to an incorrect fine-grained one. We compare confidence-based and consistency-based stopping. Confidence-based stopping halts when the softmax probability falls below $\tau = 0.9$ (chosen from $[0.85, 0.99]$), but often stops prematurely because probability is split across similar sibling classes (Fig. 8).

Consistency-based stopping halts only when taxonomy constraints are violated, requires no threshold tuning, and more reliably reaches deeper correct levels. It therefore yields more reliable and deeper correct predictions. Under this rule, Text-Attr (H-CAST) produces the most taxonomy-consistent outputs (Fig. 9), reaching deeper correct levels while avoiding inconsistent fine-grained predictions. This suggests that stronger hierarchical consistency leads to more effective free-grained inference.

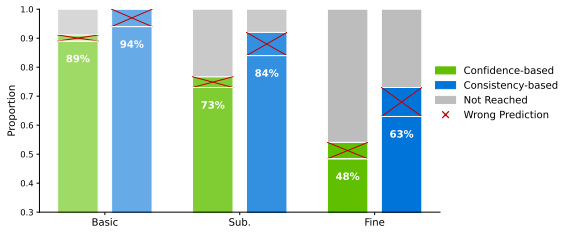


Figure 8. **Consistency-based stopping (blue) yields more reliable and deeper correct predictions than confidence-based stopping (green).** The figure compares confidence- and consistency-based stopping on ImageNet-F dataset using Text-Attr (H-CAST). Bars show the proportion of samples reaching each level (Basic → Subordinate → Fine). Gray ("Not Reached") indicates early stopping at a coarser level, and red crosses mark incorrect predictions. Confidence-based rules often stop early, failing to reach deeper levels due to probability splitting among similar classes, whereas consistency-based stopping more often reaches deeper correct predictions.

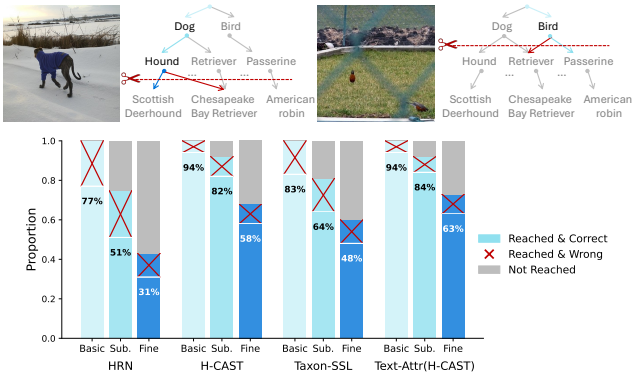


Figure 9. **Consistency-based free-grained inference reliably stops at appropriate levels for low-confidence samples on ImageNet-F. Top:** The model stops at the appropriate level, subordinate (e.g., *Hound*, left) or basic (e.g., *Bird*, right), when deeper predictions become inconsistent, yielding more reliable outputs. **Bottom:** Inference stops when finer-level predictions conflict with their coarser ancestors. On ImageNet-F, Text-Attr (H-CAST) descends deeper while maintaining correctness, whereas HRN stops earlier and produces fewer fine-level predictions.

6. Conclusion

We introduce free-grained training and inference for hierarchical recognition, where models learn from labels of varying granularity while maintaining taxonomy consistency. We present diverse benchmarks and two methods that strongly outperform prior work, while highlighting the difficulty of free-grained learning and the need for more robust approaches. **Limitations:** Class- and level-wise imbalance are not addressed and are left for future work. Additionally, better pruning based not just on CLIP could be explored to approximate intrinsic annotation difficulty.

Acknowledgements

This project was supported, in part, by NSF 2215542, NSF 2313151, and Bosch gift funds to S. Yu at UC Berkeley and the University of Michigan, with additional compute support from NAIRR Pilot (CIS250430, CIS240431).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 5, 12, 28

[2] Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. CUDA: Curriculum of data augmentation for long-tailed recognition. In *The Eleventh International Conference on Learning Representations*, 2023. 28

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 28

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 2019. 28

[5] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": fine-grained, or not. In *CVPR*, 2021. 1, 2, 3, 18, 28

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 8

[7] Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *CVPR*, 2022. 1, 2, 3, 6, 7, 18, 20, 21, 26, 28

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 26

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 5

[10] Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998. 2, 4, 28

[11] Ashima Garg, Shaurya Bagga, Yashvardhan Singh, and Saket Anand. Hiermatch: Leveraging label hierarchies for improving semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 28

[12] Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In *European Conference on Computer Vision*, 2022. 3, 28

[13] Matej Grcic, Artyom Gadetsky, and Maria Brbic. Fine-grained classes and how to find them. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 1, 28

[14] Taegil Ha, Seulki Park, and Jin Young Choi. Novel regularization via logit weight repulsion for long-tailed classification. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*, 2023. 28

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 28

[16] Juan Jiang, Jingmin Yang, Wenjie Zhang, and Hongbin Zhang. Hierarchical multi-granularity classification based on bidirectional knowledge transfer. *Multimedia Systems*, 2024. 1, 28

[17] Shyamgopal Karthik, Ameya Prabhu, Puneet K. Dokania, and Vineet Gandhi. No cost likelihood manipulation at test time for making better mistakes in deep networks. In *International Conference on Learning Representations*, 2021. 3, 28

[18] Dong-Jin Kim, Zhongqi Miao, Yunhui Guo, and X Yu Stella. Modeling semantic correlation and hierarchy for real-world wildlife recognition. *IEEE Signal Processing Letters*, 2023. 1, 28

[19] Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 5, 28

[20] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 28

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 7, 22

[22] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 2, 5, 28

[23] Zhongqi Miao, Stella X Yu, Kyle L Landolt, Mark D Koneff, Timothy P White, Luke J Fara, Enrika J Hlavacek, Bradley A Pickens, Travis J Harrison, and Wayne M Getz. Challenges and solutions for automated avian recognition in aerial imagery. *Remote Sensing in Ecology and Conservation*, 2023. 1

[24] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 28

[25] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 28

[26] Seulki Park, Youren Zhang, Stella X. Yu, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 3, 6, 7, 18, 20, 21, 26, 28

[27] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 28

[28] 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*. PMLR, 2021. 2, 4, 5, 28

[29] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, 2020. 28

[30] Joshua Robinson, Stefanie Jegelka, and Suvrit Sra. Strength from weakness: Fast learning using weak supervision. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 3, 28

[31] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 1976. 4, 12, 28

[32] 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. *IJCV*, 2015. 2, 28

[33] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 28

[34] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 2020. 28

[35] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Bioclip: A vision foundation model for the tree of life. In *CVPR*, 2024. 2, 4

[36] Yuwen Tan, Yuan Qing, and Boqing Gong. Vision llms are bad at hierarchical visual understanding, and llms are the bottleneck. *arXiv preprint arXiv:2505.24840*, 2025. 28

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017. 28

[38] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European conference on computer vision*. Springer, 2022. 28

[39] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 2, 3, 28

[40] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. 24

[41] Rui Wang, Cong Zou, Weizhong Zhang, Zixuan Zhu, and Lihua Jing. Consistency-aware feature learning for hierarchical fine-grained visual classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 1, 3, 28

[42] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. 28

[43] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010. 2, 5, 28

[44] Jianlong Wu, Haozhe Yang, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. Chmatch: contrastive hierarchical matching and robust adaptive threshold boosted semi-supervised learning. In *CVPR*, 2023. 3, 6, 26, 27, 28

[45] Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving long-tailed recognition with deep realistic taxonomic classifier. In *ECCV*, 2020. 28

[46] Seokju Yun and Youngmin Ro. Shvit: Single-head vision transformer with memory efficient macro design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5756–5767, 2024. 1

[47] Siqi Zeng, Remi Tachet des Combes, and Han Zhao. Learning structured representations by embedding class hierarchy. In *The Eleventh International Conference on Learning Representations*, 2022. 28

[48] 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*, 2023. 24

[49] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3, 28

[50] Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 28

[51] Zhaoheng Zheng, Jingmin Wei, Xuefeng Hu, Haidong Zhu, and Ram Nevatia. Large language models are good prompt learners for low-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 28

Free-Grained Hierarchical Visual Recognition

Supplementary Material

Contents

A. ImageNet-3L Dataset Construction

Our hierarchy design is inspired by cognitive studies [31], which identify the *basic level* (e.g., *dog*) as the most natural and informative category for human recognition. Motivated by this, we construct a three-level hierarchy (*basic*, *subordinate*, and *fine-grained*), starting from this most informative level and avoiding overly abstract categories (e.g., *entity*, *artifact*) that are less meaningful for visual recognition.

Importantly, the notion of the basic level is not fixed. As shown in [31], category assignments can vary depending on context (e.g., "*fish*" may appear at superordinate or basic depending on the experiment), suggesting that level boundaries are inherently flexible. We therefore adopt Rosch's taxonomy as a guideline rather than a strict definition.

1. Defining the Basic Level: We define the *basic level* as categories that are both semantically meaningful and visually informative, ensuring comparisons at a consistent granularity. We primarily adopt categories aligned with Rosch's superordinate level (e.g., *bird*, *musical instrument*), which better match the scale of ImageNet, while avoiding overly coarse concepts (e.g., *entity*, *animal*) and overly fine-grained ones. For example, we prefer *dog*, *bird*, and *snake* over broader categories such as *mammal* or *reptile*.

Classes not covered by Rosch's taxonomy are mapped to WordNet nodes at comparable semantic depths, and categories that are too coarse or too fine are adjusted to nearby levels. The resulting basic level is fixed globally, and all nodes above it are removed to enforce a uniform starting point. This avoids mismatched comparisons, such as between fine-grained label (e.g., *teddy bear*) and coarse label (e.g., *conveyance*), and ensures semantically coherent grouping across levels.

2. Enforcing a Multi-level Hierarchy: We retain only categories that form meaningful three-level structures from *basic* to *subordinate* to *fine-grained*. Categories that collapse into a single path are removed, including cases where nodes have only one child at each level or where the hierarchy terminates early. For example, if a basic-level category leads to only one subordinate class, which further has only one fine-grained class, the hierarchy provides no meaningful distinction across levels. Similarly, we exclude shallow structures where the hierarchy does not extend to all three levels after defining the basic level. These pruning steps ensure that each retained category supports non-trivial branching and meaningful differentiation across levels.

3. Selecting Categories for Diversity: When multiple valid candidates exist at a given level of the taxonomy, we select the category that leads to the richest set of descendant classes. This encourages greater intra-group diversity and supports more meaningful distinctions at finer levels. For example, under *bird*, both *parrot* and *cockatoo* are valid candidates. However, *cockatoo* leads to only a single fine-grained class (e.g., *sulphur-crested cockatoo*), whereas *parrot* covers multiple diverse species (e.g., *African grey*, *sulphur-crested cockatoo*). We therefore select *parrot* to ensure broader coverage and richer fine-grained classification.

4. Handling Ambiguities and Naming: While WordNet provides a structured hierarchy, some categories are ambiguous or inconsistently defined. In such cases, we reorganize them using semantically coherent groupings. For example, instead of ill-defined categories such as *Women's Clothing*, we restructure them into functional groups (e.g., *Underwear*) to improve clarity and consistency.

5. Quality Control: We ensure taxonomy quality through a rule-based, human-in-the-loop process that verifies parent–child consistency and sibling-level coherence throughout construction. We further validate the structure using AI-assisted review (e.g., ChatGPT [1]) to identify potential inconsistencies or violations of intuitive categorization, followed by manual verification.

B. Complete Hierarchy of ImageNet-3L

Basic	**Subordinate**	**Fine-Grained**
bird	passerine bird	brambling, indigo bunting, robin, jay, bulbul, water ouzel, house finch, chickadee, junco, magpie, goldfinch
	parrot	macaw, sulphur-crested cockatoo, African grey, lorikeet
	piciform bird	toucan, jacamar
	seabird	king penguin, pelican, albatross
	anseriform bird	drake, red-breasted merganser, black swan, goose

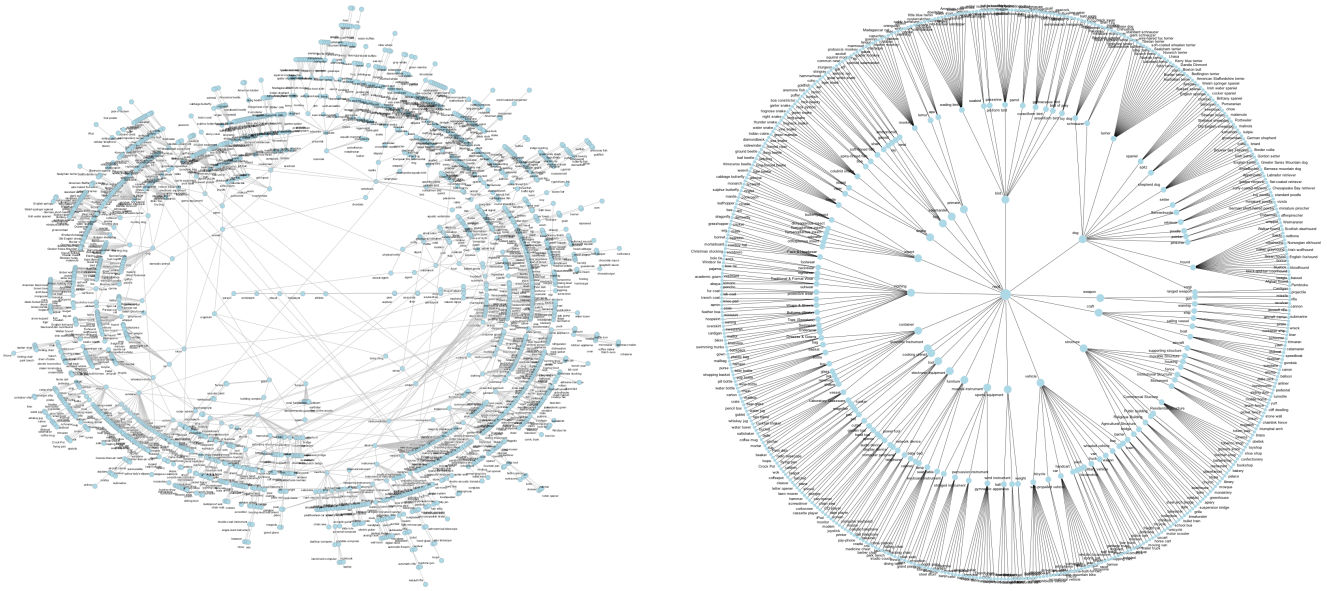
	coraciiform bird	bee eater, hornbill
	bird of prey	kite, great grey owl, vulture, bald eagle
	gallinaceous bird	partridge, prairie chicken, ruffed grouse, peacock, quail, black grouse, ptarmigan
	wading bird	flamingo, American coot, redshank, American egret, little blue heron, white stork, limpkin, spoonbill, red-backed sandpiper, dowitcher, crane, ruddy turnstone, bittern, oystercatcher, black stork, bustard
dog	spitz dog	malamute, Pomeranian, keeshond, Siberian husky, chow, Samoyed
	pointer dog	vizsla, German short-haired pointer
	spaniel dog	Brittany spaniel, clumber, English springer, Sussex spaniel, Irish water spaniel, Welsh springer spaniel, cocker spaniel
	hound dog	basset, bloodhound, Irish wolfhound, Walker hound, redbone, English foxhound, Italian greyhound, Ibizan hound, bluetick, Scottish deerhound, borzoi, Norwegian elkhound, whippet, Weimaraner, Saluki, beagle, Afghan hound, black-and-tan coonhound, otterhound
	terrier dog	Boston bull, silky terrier, Lakeland terrier, Yorkshire terrier, Tibetan terrier, American Staffordshire terrier, Irish terrier, Airedale, Norwich terrier, soft-coated wheaten terrier, wire-haired fox terrier, Staffordshire bullterrier, West Highland white terrier, Australian terrier, Dandie Dinmont, Kerry blue terrier, Lhasa, cairn, Sealyham terrier, Bedlington terrier, Scotch terrier, Border terrier, Norfolk terrier
	corgi dog	Pembroke, Cardigan
	poodle dog	miniature poodle, toy poodle, standard poodle
	setter dog	Irish setter, Gordon setter, English setter
	pinscher dog	Doberman, affenpinscher, miniature pinscher
	shepherd dog	kelpie, briard, German shepherd, Old English sheepdog, Border collie, Bouvier des Flandres, collie, Rottweiler, komondor, malinois, groenendael, Shetland sheepdog
	retriever dog	curly-coated retriever, Labrador retriever, Chesapeake Bay retriever, flat-coated retriever, golden retriever
	schnauzer dog	standard schnauzer, miniature schnauzer, giant schnauzer
	Sennenhunde dog	Bernese mountain dog, Greater Swiss Mountain dog, Appenzeller, EntleBucher
	toy dog	toy terrier, Blenheim spaniel, Maltese dog, Shih-Tzu, papillon, Pekinese, Chihuahua, Japanese spaniel
fish	soft-finned fish	coho, tench, eel, goldfish
	shark	tiger shark, great white shark, hammerhead
	spiny-finned fish	anemone fish, puffer, lionfish, rock beauty
	ray	stingray, electric ray

	ganoid fish	sturgeon, gar
primate	ape	gibbon, siamang, orangutan, chimpanzee, gorilla
	monkey	titi, langur, colobus, squirrel monkey, baboon, guenon, marmoset, macaque, spider monkey, patas, howler monkey, proboscis monkey, capuchin
	lemur	Madagascar cat, indri
snake	colubrid snake	water snake, garter snake, green snake, night snake, hognose snake, ringneck snake, king snake, thunder snake, vine snake
	elapid snake	sea snake, Indian cobra, green mamba
	viper	diamondback, horned viper, sidewinder
	boa snake	boa constrictor, rock python
salamander	newt	eft, common newt
	ambystomid salamander	spotted salamander, axolotl
insect	beetle	dung beetle, weevil, leaf beetle, tiger beetle, ladybug, rhinoceros beetle, long-horned beetle, ground beetle
	orthopterous insect	cricket, grasshopper
	dictyopterous insect	cockroach, mantis
	hymenopterous insect	bee, ant
	butterflyinsect	cabbage butterfly, lycaenid, monarch, admiral, sulphur butterfly, ringlet
	odonate insect	dragonfly, damselfly
	homopterous insect	cicada, leafhopper
furniture	table	desk, dining table
	baby bed	cradle, crib, bassinet
	seat	rocking chair, barber chair, park bench, throne, folding chair, toilet seat, studio couch
	lamp	table lamp
	cabinet	china cabinet, medicine chest
musical instrument	wind instrument	ocarina, flute, panpipe, oboe, cornet, sax, harmonica, bassoon, French horn, trombone
	stringed instrument	banjo, harp, violin, cello, acoustic guitar, electric guitar
	percussion instrument	steel drum, gong, marimba, drum, chime, maraca
	keyboard instrument	upright, grand piano, accordion, organ
scientific instrument	laboratory glassware	Petri dish
	magnifier	loupe, radio telescope
sports equipment	ball	golf ball, baseball, basketball, croquet ball
	gymnastic apparatus	parallel bars, balance beam, horizontal bar
	weight	barbell, dumbbell
electronic equipment	telephone	dial telephone, pay-phone, cellular telephone
	computer peripheral	printer, joystick, computer keyboard, mouse

	audio device	tape player, cassette player, CD player, iPod
	network device	modem
	display device	monitor, screen
clothing	bottoms (skirts)	hoopskirt, sarong, miniskirt, overskirt
	tops (sweaters)	sweatshirt, cardigan
	outwear	trench coat, poncho, fur coat
	swimwear	maillot, bikini, swimming trunks
	face & headwear	wig, sombrero, mortarboard, bonnet, mask, cowboy hat, bearskin
	nightwear	pajama
	protective wear	apron, knee pad, lab coat
	dresses & Gowns	gown
	underwear	brassiere
	footwear	sock, Christmas stocking
	neckwear	bow tie, bolo tie, Windsor tie
	traditional & formal Wear	abaya, kimono, vestment, academic gown
	wraps & shawls	stole, feather boa
container	reservoir	water tower, rain barrel
	bag	mailbag, plastic bag, backpack, purse
	jug	water jug, whiskey jug
	vessel	mortar, pitcher, tub, ladle, bucket, coffee mug
	bottle	wine bottle, beer bottle, pop bottle, water bottle, pill bottle
	basket	hamper, shopping basket
	box	mailbox, carton, pencil box, chest, crate
	glass	goblet, beer glass
	shaker	saltshaker, cocktail shaker
cooking utensil	pan	frying pan, wok
	cooker	Crock Pot
	pot	teapot, caldron, coffeepot
structure	monument	brass, megalith, triumphal arch, obelisk, totem pole
	religious building	church, mosque, boathouse, monastery, stupa
	housing	yurt, cliff dwelling, mobile home
	public building	planetarium, library
	movable structure	sliding door, turnstile
	supporting structure	plate rack, honeycomb, pedestal
	fence	stone wall, picket fence, chainlink fence, worm fence
	bridge	steel arch bridge, viaduct, suspension bridge
	residential structure	palace
	agricultural structure	greenhouse, barn, apiary

	commercial stucture	toyshop, restaurant, cinema, confectionery, bookshop, grocery store, tobacco shop, bakery, butcher shop, barbershop, shoe shop
	barrier	grille, bannister, breakwater, dam
	institutional structure	prison
tool	hand tool	hammer, plunger, screwdriver
	garden tool	lawn mower, shovel
	cutter	cleaver, plane, letter opener, hatchet
	power tool	chain saw
	opener	corkscrew, can opener
craft	sailing vessel	trimaran, schooner, catamaran
	boat	fireboat, canoe, yawl, gondola, speedboat, lifeboat
	ship	wreck, pirate, container ship, liner
	warship	aircraft carrier, submarine
	aircraft	airliner, warplane, airship, balloon
vehicle	bicycle	bicycle-built-for-two, mountain bike
	bus	minibus, school bus, trolleybus
	car	ambulance, beach wagon, cab, convertible, jeep, limousine, Model T, racer, sports car
	truck	fire engine, garbage truck, pickup, tow truck, trailer truck
	van	minivan, moving van, police van
	locomotive	electric locomotive, steam locomotive
	military vehicle	half track
	self-propelled vehicle	forklift, recreational vehicle, snowmobile, tank, tractor, golfcart, snowplow, go-kart, moped, streetcar, amphibious vehicle
	handcart	barrow, shopping cart
	sled	bobsled, dogsled
	train	bullet train
	wagon	horse cart, jinrikisha, oxcart
	wheeled vehicle	freight car, motor scooter, tricycle, unicycle
weapon	gun	rifle, assault rifle, revolver, cannon
	ranged weapon	missile, projectile

Table 3. **Complete hierarchy tree for our proposed ImageNet-3L dataset.**



(a) Original ImageNet's WordNet hierarchy

(b) Our 3-level hierarchy

Figure 10. **We restructure the original WordNet hierarchy into a clean, consistent three-level hierarchy for hierarchical recognition.**

C. Experiments on Messy Hierarchy

Setup. To evaluate robustness under the original WordNet structure, we construct a subset of ImageNet restricted to the *mammal* subtree. All ancestor nodes above *mammal* (e.g., *vertebrate*, *animal*, *entity*) are removed, and only classes under *mammal* are retained. This results in 120 fine-grained classes, with 152,548 training and 6,050 validation samples. The resulting hierarchy remains highly irregular (Fig. 11), with path lengths ranging from 5 to 10 levels. The distribution of chain lengths is skewed (e.g., most samples concentrate at depths 8–9, while deeper levels are sparse), leading to significant imbalance across levels.

Limitations of the Raw WordNet Hierarchy for Full-Path Prediction. Our goal follows the standard objective of hierarchical recognition [5, 7, 26], where a model predicts the full path across all levels. However, the raw WordNet hierarchy is not well-suited for this setting due to its irregular structure: classes have varying depths, making consistent level alignment impossible. To illustrate this, we construct a subset under the *mammal* hierarchy and perform full-path prediction on the raw structure (Fig. 11). We anchor all samples at the fine-grained level (Level 10) and propagate upward from *mammal*, leaving intermediate levels empty when necessary. While this enables leaf-level comparison, it introduces semantic misalignment across higher levels. For example, *hound* is compared with *sporting dog* rather than *spaniel*, and both *spaniel* and *English toy spaniel* appear at Level 8 despite their hierarchical inclusion. These issues lead to (1) *semantic inconsistency*, where nodes at the same level represent different levels of abstraction, and (2) *structural sparsity*, where certain levels contain very few classes (e.g., Level 5–7), resulting in weak supervision. While such hierarchies can still support flat classification with hierarchical penalties, they are **not suitable for full-path prediction**, which requires consistent semantic alignment across levels. This motivates the need to restructure and align the hierarchy, as in our ImageNet-3L, so that each level corresponds to a coherent semantic granularity.

Results and Analysis. Despite these challenges, Text-Attr (H-CAST) consistently improves performance over H-CAST across most levels (Level 3–Level 10), demonstrating robustness to structural noise. Levels 1 and 2 are excluded, as each contains only one class, resulting in trivial (100%) accuracy. The only exception is Level 9, where both methods perform poorly due to extreme sparsity and inconsistent supervision (some samples do not naturally have this level but are forced to predict it). These results highlight that the gains from *Text-Attr* are not tied to a well-structured hierarchy, but persist even in this ill-posed setting. At the same time, they motivate the need for a structured benchmark such as ImageNet-3L, which resolves these inconsistencies and provides a more meaningful evaluation of hierarchical recognition.

Table 4. **Performance under the original WordNet hierarchy**, which exhibits irregular depth and highly imbalanced levels across classes, making it less suitable for hierarchical evaluation. Despite this, we train and evaluate on the given structure as a robustness test, where Text-Attr provides effective semantic supervision and improves performance.

Depth (# class)	Level 10 (120)	Level 9 (3)	Level 8 (15)	Level 7 (8)	Level 6 (8)	Level 5 (4)	Level 4 (4)	Level 3 (2)
H-CAST	63.5	0.0	78.0	82.2	86.7	95.4	95.2	96.9
Text-Attr (H-CAST)	**67.7**	0.0	**80.0**	**83.8**	**87.5**	**96.6**	**96.0**	**98.0**

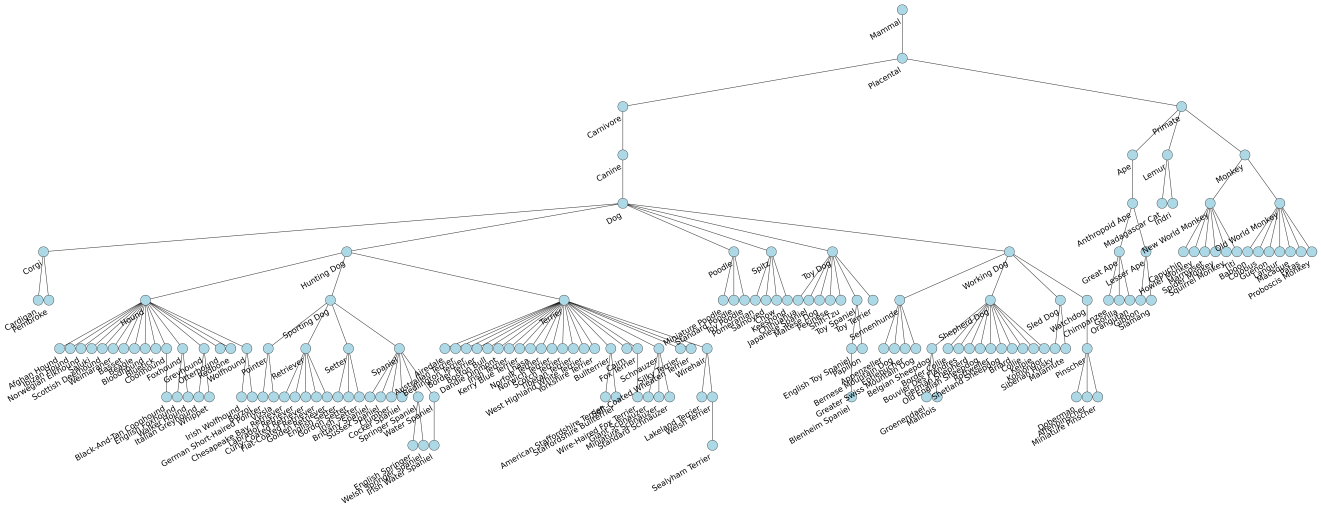


Figure 11. **Messy hierarchy of the ImageNet mammal subset.** The taxonomy exhibits irregular depths and inconsistent granularity across branches, making full-path prediction ill-posed. Despite these structural issues, our method consistently improves over the baseline, demonstrating robustness to noisy and misaligned hierarchies. Best viewed when zoomed in.

D. More Experimantal Results

D.1. Evaluaion on CUB-F

On the small-scale, single-domain dataset CUB-F (Table 5), Taxon-SSL achieves the best performance (63.96% FPA), showing the advantage of structured label propagation when per-class samples are scarce. Text-Attr methods perform moderately well (53.99–57.59% FPA) but are less effective here, as the bird-only domain limits textual diversity and reduces the benefit of language-based supervision. Still, they clearly outperform conventional hierarchical baselines (44.30% for HRN, 45.10% for H-CAST), underscoring the overall effectiveness of our approach. Unlike the trend on large-scale, diverse datasets such as ImageNet-F, where Text-Attr provides richer cues and stronger gains, these results confirm that there is no single recipe for free-grain learning: performance is tightly coupled with dataset characteristics, making the problem inherently challenging.

Table 5. **Taxon-SSL shows strong effectiveness on the small-scale dataset CUB-F, where label propagation provides reliable supervision.** Text-Attr methods are assumed to offer limited benefit due to the restricted textual diversity of this bird-only dataset.

CUB-F (13-38-200)	FPA (↑)	Species (↑)	family (↑)	Order (↑)	TICE (↓)
HRN [7]	44.30	46.72	81.20	96.36	27.15
H-CAST [26]	45.10	47.52	87.78	97.50	25.89
Taxon-SSL	**63.96**	**65.50**	**92.84**	<u>98.40</u>	**7.39**
Taxon-SSL + Text-Attr	<u>63.05</u>	<u>64.86</u>	<u>92.54</u>	98.38	<u>7.61</u>
Text-Attr (H-ViT)	57.59	59.10	91.60	98.05	10.72
Text-Attr (H-CAST)	53.99	55.58	91.72	**98.41**	18.95

D.2. Evaluation under Varying and Severe Label Sparsity Conditions

To evaluate model performance under diverse free-grain conditions, we experiment with various label availability ratios by randomly removing multi-level labels, e.g., (100%-60%-30%), (100%-50%-10%), and (100%-20%-10%), which represent the available proportions of basic, subordinate, and fine-grained labels, respectively. Each experiment is repeated with three different random seeds, and we report the average performance. The variance across runs was minor (0.1–1.8).

Consistent with our main results, these experiments (Table 6 & 7 & 8) also show that **there is *no* single method that performs best across all settings**. Instead, the most effective method varies depending on the dataset and the specific ratio of available labels, highlighting the importance of adaptable free-grain learning strategies.

For consistency, we refer to the three levels in CUB-Rand (order-family-species) and Aircraft-Rand (maker-family-model) as basic, subordinate, and fine-grained levels. We summarize the key findings below:

(1) Conventional hierarchical classification methods struggle under the free-grain setting, where label supervision is sparse and uneven across levels. For example, when labels are highly missing (e.g., only 10% available at the fine-grained level), HRN [7] and H-CAST [26] suffer more than a 50% drop in accuracy across all levels compared to the fully labeled (100%-100%-100%) setting on CUB-Rand (Fig. 6 & Table 8). This highlights the difficulty of the free-grain setting and the need for methods that can robustly handle incomplete supervision at multiple semantic levels.

(2) The performance of different methods varies with the amount of available supervision per class: Text-Attr methods perform better when more labeled samples are available, while Taxon-SSL is more effective under extreme label sparsity. For example, in Table 6, the average number of available fine-grained labels per class is approximately 9 for CUB-Rand and about 20 for Aircraft-Rand. Consistent with this difference, Taxon-SSL outperforms other methods on CUB-Rand, whereas Text-Attr (H-CAST) performs best on Aircraft-Rand. This trend persists across settings. In the most sparse setting, CUB-Rand (100-20-10, Table 8), where only about 3 fine-grained labels are available per class, Taxon-SSL shows a clear advantage. We attribute this to how supervision is utilized. Text-Attr relies on available labels and indirect semantic guidance via text features. In contrast, Taxon-SSL actively leverages unlabeled data through pseudo-labeling and strong augmentations, making it more effective when labeled examples are extremely limited.

(3) Sometimes, Taxon-SSL's high fine-grained accuracy comes at the cost of lower accuracy at higher levels in the taxonomy. For example, in Table 7, Taxon-SSL achieves the highest fine-grained accuracy (65.01%), but its subordinate and basic-level accuracies (85.53% and 92.81%) are lower than those of Text-Attr (H-CAST), which achieves 86.30% and 94.17%, respectively. This highlights a key challenge in free-grain learning: improving accuracy across all levels simultaneously is non-trivial, and optimizing for fine-grained performance alone may degrade consistency at coarser levels.

Table 6. **No single method performs best across all conditions—performance depends strongly on the amount of available supervision per class.** Text-Attr methods tend to perform better when more labeled samples are available, while Taxon-SSL is more effective under extreme label sparsity. For example, Taxon-SSL performs best on CUB-Rand with around 9 fine-grained labels per class, while Text-Attr (H-CAST) performs best on Aircraft-Rand with around 20, reflecting the impact of supervision density. These results highlight that method effectiveness is highly sensitive to label sparsity, emphasizing the need for adaptable approaches in free-grain learning.

Label Ratio	CUB-Rand (100%-60%-30%)					Aircraft-Rand (100%-60%-30%)				
	FPA(↑)	spec.(↑)	fam.(↑)	order(↑)	TICE(↓)	FPA(↑)	maker(↑)	fam.(↑)	model(↑)	TICE(↓)
HRN [7]	57.87	62.73	85.53	96.45	13.77	57.33	64.42	76.95	86.38	23.30
H-CAST [26]	61.88	67.36	90.05	94.32	13.04	64.67	68.88	85.58	91.43	13.76
Taxon-SSL	74.82	76.92	93.38	98.33	5.06	70.33	72.22	87.06	93.50	**7.18**
Taxon-SSL + Text-Attr	**74.90**	**76.95**	**93.41**	98.38	**4.91**	69.89	72.24	86.92	93.29	7.77
Text-Attr (H-ViT)	67.89	72.48	90.63	95.37	10.39	64.15	68.92	85.88	89.87	15.80
Text-Attr (H-CAST)	69.65	71.31	92.88	**98.48**	8.35	**71.43**	**73.56**	**89.66**	**95.31**	9.71

Table 7. **Maintaining accuracy across all hierarchy levels remains more challenging under sparse supervision.** For example, in 100%-50%-10% case, Taxon-SSL achieves the highest fine-grained accuracy (65.01%), but its subordinate and basic-level accuracies (85.53%, 92.81%) are lower than those of Text-Attr (H-CAST) (86.30%, 94.17%), which better preserves consistency across levels. This result illustrates the inherent difficulty of improving accuracy across all levels simultaneously, as objectives at different levels can be conflicting.

Label Ratio	Aircraft-Rand (100%-50%-10%)					Aircraft-Rand (100%-20%-10%)				
	FPA(↑)	maker(↑)	fam.(↑)	model(↑)	TICE(↓)	FPA(↑)	maker(↑)	fam.(↑)	model(↑)	TICE(↓)
HRN [7]	40.35	47.85	70.76	85.68	37.56	32.06	46.73	55.43	85.58	48.43
H-CAST [26]	47.57	51.93	78.31	87.11	28.42	40.33	45.44	67.28	84.12	35.61
Taxon-SSL	62.61	65.01	85.53	92.81	**10.22**	**58.73**	**61.10**	80.90	92.24	**11.77**
Taxon-SSL + Text-Attr	**62.95**	**65.49**	86.01	92.64	10.25	58.55	60.88	**80.97**	92.04	11.89
Text-Attr (H-ViT)	47.83	52.25	81.13	87.82	30.57	38.73	43.89	66.13	84.81	38.69
Text-Attr (H-CAST)	53.31	55.32	**86.30**	**94.17**	24.43	48.85	51.37	77.11	**93.01**	27.25

Table 8. **Taxon-SSL is more robust under extreme label sparsity.** In CUB-Rand (100%-20%-10%), where each class has only 3 fine-grained and 3 subordinate labels, Taxon-SSL achieves the best performance, while other methods struggle. HRN and H-CAST suffer over 50% drop in fine-grained accuracy compared to the fully-supervised (100%-100%-100%) setting. Text-Attr methods perform more robustly (10%+ higher than HRN/H-CAST), but still struggle under sparse supervision. We attribute this to how each method leverages supervision: Text-Attr depends on the provided labels and text-derived semantics, whereas Taxon-SSL can better exploit unlabeled data through pseudo-labeling and augmentations, leading to stronger performance when label sparsity is severe.

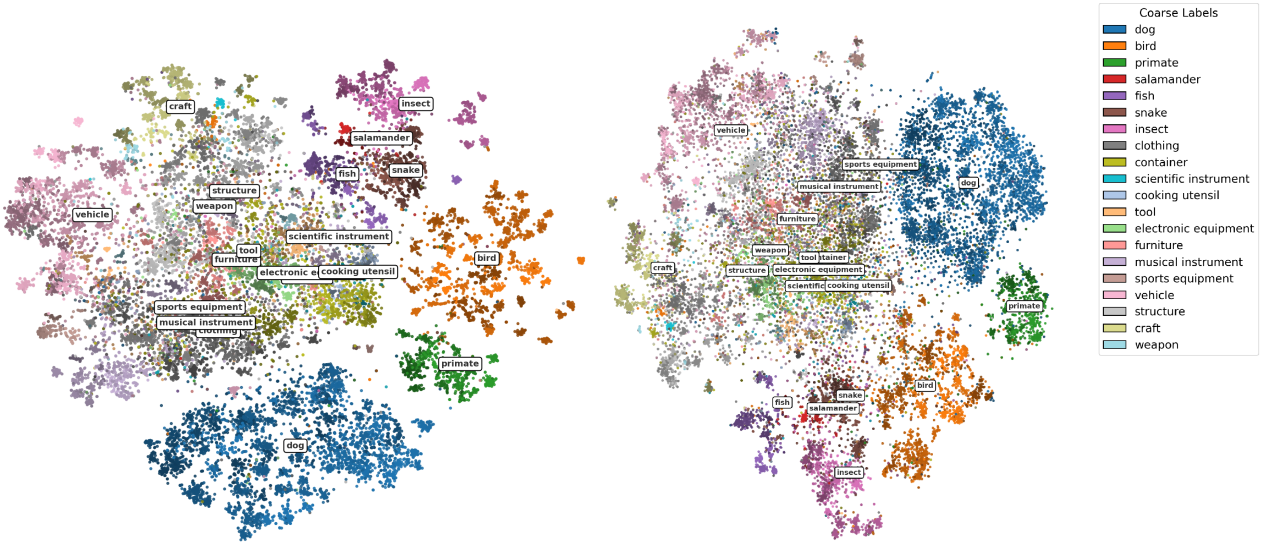
Label Ratio	CUB-Rand (100%-50%-10%)					CUB-Rand (100%-20%-10%)				
	FPA(↑)	spec.(↑)	fam.(↑)	order(↑)	TICE(↓)	FPA(↑)	spec.(↑)	fam.(↑)	order(↑)	TICE(↓)
HRN [7]	40.23	43.70	82.75	95.94	22.34	33.53	41.18	72.56	95.79	30.50
H-CAST [26]	39.03	43.41	85.74	93.23	24.60	32.97	38.66	76.89	92.50	29.43
Taxon-SSL	62.40	64.14	**92.33**	**98.26**	**6.01**	**59.18**	**61.44**	**89.79**	**98.20**	**7.65**
Taxon-SSL + Text-Attr	**62.52**	**64.87**	87.94	94.45	8.98	57.98	60.59	89.42	98.12	8.39
Text-Attr (H-ViT)	47.42	50.74	88.22	94.67	18.09	42.46	46.99	80.92	94.43	20.27
Text-Attr (H-CAST)	44.63	45.89	91.06	98.19	22.72	40.41	42.76	84.24	97.97	24.05

E. t-SNE Visualization

We visualize ImageNet-F embeddings of Text-Attr (H-CAST) and Taxon-SSL using t-SNE [21] to assess whether the learned representations capture semantic and hierarchical structure. Each point denotes an image embedding, colored by its basic-level class (20 categories), with brightness variations indicating fine-grained subclasses (505 total).

Both Text-Attr (H-CAST) and Taxon-SSL produce well-separated clusters consistent with the basic-level taxonomy, showing that coarse groupings are reliably captured. The key difference lies within coarse categories: **Text-Attr (H-CAST) reveals more distinct fine-grained subclusters** (e.g., breeds within *dog*, species within *bird*), whereas **Taxon-SSL yields tighter coarse clusters with less apparent fine-level separation**.

This contrast reflects their supervision signals. Text-Attr leverages diverse textual cues (attributes, parts, appearance terms), which promote discriminative, attribute-aligned features and sharpen within-class distinctions. Taxon-SSL, by propagating labels along the taxonomy and enforcing consistency under mixed-granularity supervision, regularizes embeddings within each coarse class and reduces intra-class variance—emphasizing coarse alignment over fine-level separability.



(a) Text-Attr (H-CAST) (b) Taxon-SSL

Figure 12. **t-sne Visualization on ImageNet-F.** Both methods separate coarse-level taxonomy well, but Text-Attr (H-CAST) yields clearer fine-grained subclusters (e.g., distinct groups within *dog* and *bird*) with more compact grouping, whereas Taxon-SSL shows some overlap of embeddings near cluster boundaries. This is likely due to ImageNet-F 's diverse large-scale categories, where text supervision provides rich attribute cues that sharpen fine-level distinctions.

F. Ablation Study

F.1. Importance of Text-guided Pseudo Attributes

Text-guided Pseudo Attributes jointly optimizes hierarchical label supervision ($\mathcal{L}_{\text{hier}}$) and text-guided pseudo attributes ($\mathcal{L}_{\text{text}}$) to learn semantically rich features: $\mathcal{L} = \mathcal{L}_{\text{hier}} + \alpha\mathcal{L}_{\text{text}}$ Fig. 13 quantifies $\mathcal{L}_{\text{text}}$'s impact by varying its weight α on CUB-Rand. Ablating $\mathcal{L}_{\text{text}}$ ($\alpha = 0$) causes a 5% absolute decline in both fine-grained accuracy and FPA compared to the optimal configuration ($\alpha = 0$). This gap underscores two key roles of text guidance: (1) it injects complementary visual semantics absent in class labels alone, and (2) it enforces attribute consistency across hierarchy levels. The performance recovery at ($\alpha = 1$) confirms that textual pseudo-attributes mitigate annotation sparsity while preserving taxonomic coherence.

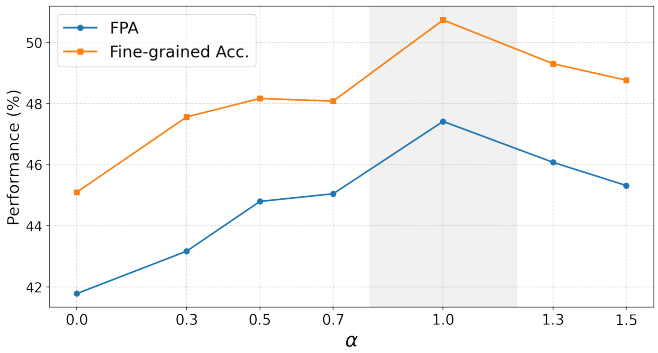


Figure 13. **Tuning α balances accuracy and taxonomic consistency.** At $\alpha = 1$ (optimal), Text-Attr (H-ViT) achieves peak fine-grained accuracy (blue) while maintaining hierarchical consistency (orange). Ablating $\mathcal{L}_{\text{text}}$ ($\alpha = 0$) causes a 5% accuracy drop and increased inconsistency, as class embeddings lose text-guided attribute alignment. Higher $\alpha > 1.0$ over-regularizes features, marginally degrading both metrics. This trade-off underscores the need to weight text supervision to resolve sparse annotations without distorting the hierarchy.

F.2. Combining Text-Attr and Taxon-SSL

We compare different training schedules for combining Text-Attr and Taxon-SSL on CUB-F. In the **joint setting**, both objectives are optimized simultaneously for 100 epochs. In the **two-stage setting**, we first train with one objective for 50 epochs and then add the other for the remaining 50 epochs, considering both orders: (1) Taxon-SSL $\rightarrow$ Text-Attr, and (2) Text-Attr $\rightarrow$ Taxon-SSL.

Table 9 show that starting with Text-Attr and then adding Taxon-SSL yields slightly higher full-path accuracy, likely because textual supervision promotes diverse feature learning before label propagation. In contrast, beginning with Taxon-SSL provides no advantage, and both two-stage variants perform similarly to joint training overall. Interestingly, joint training achieves higher consistency as measured by TICE. Given its simplicity and competitive performance, we adopt the joint strategy as our default.

Table 9. **Comparison of joint vs. two-stage training schedules for Text-Attr and Taxon-SSL on CUB-F.** While two-stage training (Text-Attr $\rightarrow$ Taxon-SSL) yields slightly higher accuracy, joint learning is simpler and provides better consistency (TICE).

CUB-F (13-38-200)	FPA (↑)	Species (↑)	family (↑)	Order (↑)	TICE (↓)
Taxon-SSL + Text-Attr (100 epochs)	<u>63.04</u>	<u>64.86</u>	<u>92.54</u>	**98.37**	**7.61**
Taxon-SSL (50 epochs) $\rightarrow$ +Text-Attr (50 epochs)	62.84	64.42	92.47	98.20	8.19
Text-Attr (50 epochs) $\rightarrow$ +Taxon-SSL (50 epochs)	**63.63**	**65.34**	**92.56**	<u>98.27</u>	<u>8.06</u>

F.3. Ablation on the Text Encoder in Text-Attr

Text-Attr relies on a text encoder to embed visual descriptions. To assess whether performance depends on the choice of text encoder, we replace CLIP with alternative encoders, including SigLIP [48] (*siglip-base-patch16-224*) and E5 [40] (*e5-base-v2*). All models use base configurations with a 768-dimensional embedding, and are integrated into our framework with minimal modification. As shown in Table 10, all encoders achieve comparable performance, with only minor differences across metrics. The text-only model (E5) performs slightly worse than CLIP and SigLIP, but the overall trend remains consistent. These results suggest that the gains of Text-Attr are not tied to a specific text encoder, but rather stem from the proposed training formulation for learning visual attributes.

Table 10. **Ablation on the text encoder used in Text-Attr on ImageNet-F**. Replacing CLIP with SigLIP and E5 results in comparable performance, with a slight drop for the text-only encoder (E5), indicating that gains mainly come from the proposed training formulation rather than the choice of encoder.

Text encoder	FPA (↑)	fine (↑)	sub. (↑)	basic (↑)	TICE (↓)
CLIP	**63.2**	**64.9**	**84.5**	93.6	**18.6**
SigLIP	62.9	**64.9**	84.3	**93.7**	19.5
E5	61.8	63.9	84.1	93.4	20.6

F.4. Ablation on Hierarchical Supervision in ViT

We further examine the architectural design choice of where to inject hierarchical supervision in the Vision Transformer (ViT) in Table 11. On CUB-F, we map the three taxonomy levels (Order–Family–Species) to different layers and compare multiple configurations: (6th, 9th, 12th), (8th, 10th, 12th), and (10th, 11th, 12th).

Among these, supervision at the 8th, 10th, and 12th layers yields the best performance. We interpret this as a balance between early and late representation learning: assigning hierarchy too early (e.g., 6–9–12) forces the model to align coarse categories before sufficient visual features are developed, while placing all supervision too late (e.g., 10–11–12) limits the model's capacity to gradually refine class granularity. The 8–10–12 configuration provides an appropriate middle ground, where lower-level categories benefit from moderately abstract features, and finer distinctions are introduced after the backbone has matured.

Table 11. **Performance comparison of different layer assignments for hierarchical supervision in ViT on CUB-F.** The 8th–10th–12th configuration achieves the best results, balancing early and late feature abstraction.

CUB-F (13-38-200)	FPA (↑)	Species (↑)	family (↑)	Order (↑)	TICE (↓)
6-9-12th layer	54.80	58.16	88.97	95.01	16.79
8-10-12th layer	**57.59**	**59.10**	**91.60**	**98.05**	**10.72**
10-11-12th layer	56.40	58.56	90.80	97.08	13.48

F.5. Ablation on Captioning Strategy and Caption Cost

Text-Attr relies on VLM-generated descriptions to capture visual attributes. To evaluate the benefit and cost of this step, we compare our approach with a simpler alternative that uses class-level text prompts (e.g., "a photo of [deepest available label]") instead of image-level descriptions. As shown in Table 12, image-level captions consistently outperform the simple text baseline across all metrics (+1.2 FPA), demonstrating the advantage of incorporating instance-specific visual descriptions. Nevertheless, the class-level text remains a strong baseline, indicating that even minimal textual supervision is effective.

Caption Cost. Generating image-level captions takes approximately 2.1 seconds per image on a single A40 GPU. This cost is incurred only once during preprocessing and can be further reduced through parallelization across multiple GPUs.

Table 12. **Comparison between image-level captions and simple class-level text**. Image-level descriptions improve performance across all metrics, while the simple text baseline remains competitive.

Caption	FPA (↑)	fine (↑)	sub. (↑)	basic (↑)	TICE (↓)
Simple Text	62.0	64.1	84.2	93.4	20.4
Ours (Image-level)	**63.2**	**64.9**	**84.5**	**93.6**	**18.6**

G. Implementation Details

For ViT [8] models, we use ViT-Small for Text-Attr (H-ViT) and Taxon-SSL and H-CAST-Small [26] for Text-Attr (H-CAST) to match parameter sizes.

For Text-Attr (H-ViT), we insert fully-connected layers to the class token at the 8th, 10th, and 12th layers for basic, subordinate, and fine-grained supervision. The 12th-layer patch features are projected to match the text embedding dimension via an FC layer. For Text-Attr (H-CAST), hierarchical supervision is applied to the last three blocks, following [26]. Due to low dimensionality in the final block, we align text features with the features of the second block. For Text-Attr methods, CLIP-ViT-B/32 is used to extract text embeddings, which remain frozen during training.

In Taxon-SSL, we apply a shared MLP to the class token from the final (12th) layer, followed by three separate linear classifiers for basic, subordinate, and fine-grained supervision. When combined with Text-Attr, we additionally project the class token through a linear layer and align it with the corresponding text feature.

For hierarchical classification baselines, HRN [7] and H-CAST [26], we follow their original training protocols and retrain them under our free-grain setting. We extend HRN to handle missing labels at two levels instead of one. For H-CAST, we provide supervision using the available labels at each corresponding level. Full hyperparameter configurations are provided in Table 13.

We train all models for 100 epochs, except for ImageNet-F, which are trained for 200 epochs due to the larger scale. All experiments were conducted on an NVIDIA A40 GPU with 48GB memory. We used a single GPU for all experiments, except for ImageNet-F, which was trained using 4 GPUs.

Table 13. **Hyperparameters for training Text-Attr (H-ViT), Text-Attr (H-CAST), and Taxon-SSL.** We follow the training setup of H-CAST [26] for Text-Attr methods (Text-Attr (H-ViT) and Text-Attr (H-CAST)), and adopt the settings of CHMatch [44] for Taxon-SSL.

Parameter	Text-Attr (H-ViT)	Text-Attr (H-CAST)	Taxon-SSL
batch_size	256	256	128
crop_size	224	224	224
learning_rate	$5e-4$	$5e-4$	$1e-3$
weight_decay	0.05	0.05	0.05
momentum	0.9	0.9	0.9
warmup_epochs	5	5	0
warmup_learning_rate	$1e{-}6$	$1e{-}6$	N/A
optimizer	Adam	Adam	SGD
learning_rate_policy	Cosine decay	Cosine decay	Cosine decay
α (weight for $\mathcal{L}_{\text{text}}$)	1	1	1 (for +Text-Attr)

H. Full Losses for Taxonomy-guided Semi-Supervised Learning (Taxon-SSL)

In this we provide full details of Taxonomy-guided Semi-Supervised Learning (Taxon-SSL). As described in Sec. 4.3, following standard practice [44], our classifier consists of a shared feature extractor f_{feat} and level-specific heads $\{h_l\}_{l \in \mathcal{S}_x}$. For supervised samples with known labels $y_1, \ldots, y_L$ across L taxonomy levels, we apply a per-level hierarchical supervision loss $\mathcal{L}_{\text{sup}}$:

$$\mathcal{L}_{\text{sup}} = \sum_{l=1}^{L} \mathbb{1}_{\{y_l \text{ exists}\}} \cdot \mathcal{L}(h_l(f(x)), y_l). \tag{5}$$

For unlabeled levels, we generate a weakly augmented image ($\mathcal{W}(x)$) and a strongly augmented version ($\mathcal{S}(x)$). Confident predictions from $\mathcal{W}(x)$ at each levels become pseudo-labels for $\mathcal{S}(x)$, denoted as $\mathcal{PL}_l(x)$. Concretely, our pseudo-labeling loss $\mathcal{L}_{\text{pl}}$ is

$$\mathcal{L}_{\text{pl}} = \sum_{l=1}^{L} \mathbb{1}_{\{\mathcal{PL}_l(x) \text{ is over confidence threshold}\}} \cdot \mathcal{L}(h_l(f(\mathcal{S}(x))), \mathcal{PL}_l(x)). \tag{6}$$

Confidence thresholds follow the percentile-based schedule introduced in CHMatch [44].

Lastly, to reinforce hierarchical structure in the learned representation, we construct a taxonomy-aligned affinity graph. Within each mini-batch $\mathcal{B}$, we construct affinity graphs W^l for each hierarchy, where $W^l_{ij} = 1$ if the ith and jth image have the same pseudo-labels and $W^l_{ij} = 0$ otherwise. Then the taxonomy-aligned affinity graph W is defined by

$$W_{ij} = \begin{cases} 1 & \text{if } W^1_{ij} = \ldots = W^L_{ij} = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

We pull together positive pairs where $W_{ij} = 1$ and push apart negative pairs where $W_{ij} = 0$. Formally, $\mathcal{L}_{\text{tacl}}$ is defined by

$$\mathcal{L}_{\text{tacl}} = -\frac{1}{\sum_j W_{ij}} \cdot \log \frac{\sum_j W_{ij} \exp((g(f(x_i)) \cdot g(f(x_j))')/t)}{\sum_j (1 - W_{ij}) \exp((g(f(x_i)) \cdot g(f(x_j))')/t)}, \tag{8}$$

where i is the index of current image, g is the projection head for $\mathcal{L}_{\text{sup}}$, and t is the temperature hyperparameter.

Our final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hier}} + \beta \cdot \mathcal{L}_{\text{pl}} + \gamma \cdot \mathcal{L}_{\text{tacl}}, \tag{9}$$

where β and γ control the relative contributions of the loss terms, and we set both to 1 in all experiments.

I. Related Work

Hierarchical recognition has been used to refer to different tasks. A large body of work leverages taxonomies only to improve **leaf-node prediction** [12, 17, 47, 49]. Thus, evaluation in these works typically relies on top-1 accuracy or mistake severity at the leaf level. These methods predict a *single fine-grained label* and assume the full hierarchy can be recovered from it. However, fine-grained prediction is often impossible in real-world scenarios due to visual ambiguity, leading these models to produce uninformative outputs.

In this paper, we focus on **full-taxonomy prediction** [5, 16, 26, 41], where the model must output labels at all levels of the taxonomy. This setting introduces cross-level inconsistency, as predictions across levels must align with the hierarchical structure. Recent work [36] shows that such inconsistencies arise even in large vision–language models like GPT-4o [15] and Qwen2.5-VL-72B [3], underscoring the difficulty of the problem.

Existing full-taxonomy prediction methods enforce consistency constraints and demonstrate strong performance [5, 16, 26, 41]. However, they assume that all hierarchical levels are fully annotated for every training sample, which is unrealistic. In practice, annotation granularity naturally varies due to visual ambiguity or annotator expertise—some images may only receive a coarse label (e.g., *bird*), while others have fine-grained labels (e.g., *bank swallow*).

To address this gap, we study a more realistic setting, free-grain learning, where supervision may appear at any level of the taxonomy and the model must infer the complete label path from partially observed labels. Existing approaches related to partial hierarchical supervision do not fully capture this setting. HRN [7] considers partial labels but only in a limited two-level scenario created by randomly removing fine-grained labels, which does not reflect the structured ambiguity found in real taxonomies. Kim et al. [18] also incorporates mixed-granularity labels, but treats them in a flat manner, overlooking the hierarchical relationships essential to our formulation. These studies were also restricted to small datasets such as CUB [43], whereas our work establishes a general free-grain formulation and provides a large-scale benchmark to study it systematically.

Hierarchical recognition datasets. Full-taxonomy prediction requires datasets where each hierarchical level is meaningful to predict. However, large-scale datasets like ImageNet [32] are not designed for this purpose. Its WordNet [10] taxonomy has uneven depths (e.g., *minivan* appears around the 15th level while *teddy bear* appears around the 7th), and contains many abstract nodes such as *entity*, *object*, or *whole* that are **not useful prediction targets** (Fig. 3). Such a hierarchy supports only leaf-node prediction with mistake-severity penalties [12], where the model still predicts a single leaf label and the hierarchy is used merely to score how far an error is from the ground truth.

Because of this limitation, prior full-taxonomy prediction work has relied on small, clean datasets like CUB [43] and Aircraft [22], where the hierarchy is well-defined but the scale is limited. iNaturalist [39] provides a deeper taxonomy, but its scope is restricted to biological species and does not generalize to broad visual domains.

To enable large-scale and general hierarchical recognition, we introduce ImageNet-3L, which provides three semantically meaningful levels without the abstract superordinate nodes (e.g., *entity*, *object*) present in WordNet. We center the hierarchy on Rosch's basic-level categories [31], the level at which humans naturally identify objects (e.g., *bird*, *vehicle*), and organize categories downward into subordinate and fine-grained levels. This produces a clean and meaningful three-level taxonomy that focuses on distinctions worth predicting and is well suited for full-taxonomy recognition. Using this, we create free-grain variants to study hierarchical prediction under varying label granularity.

Long-tailed recognition has been extensively studied [2, 14, 20, 24, 25, 29, 38, 42, 50], mostly focusing on imbalance at a single fine-grained level (*inter-class* imbalance). In contrast, we address both *inter-class imbalance* (across classes) and *intra-hierarchy imbalance* (across semantic levels) in a hierarchical setting, where classes themselves may be balanced but label granularity varies across them. DeepRTC [45] considers taxonomies, but aims to improve inference reliability via early stopping rather than predicting the full taxonomy.

Semi-supervised learning typically combines labeled and unlabeled data at a single fine-grained level [4, 34, 37]. Recent work incorporates coarse labels [11, 44], but still targets fine-grained accuracy. In contrast, our setting demands consistent prediction across the full taxonomy with heterogeneous supervision, making existing methods not directly applicable.

Weakly-supervised recognition typically aims to predict fine-grained labels when only coarse labels are available during training [13, 30]. These methods assume fully observed labels at a coarse level and focus on improving predictions at a fine-grained level. In contrast, our setting requires handling multi-granularity labels and inferring the full taxonomy.

Large-language models for recognition. Recent approaches [19, 27, 33, 51] leverage vision–language models (VLMs) (e.g., CLIP [28]) or large language models (LLMs) (e.g., GPT-4 [1]) by generating textual descriptions from label names and feeding them into an LLM to improve flat classification. Their primary goal is to perform label-driven reasoning without training a new visual model. In contrast, we do not use labels to expand label descriptions. Instead, we use VLMs to extract

textual cues directly from the image, without referencing labels, so that the image encoder can learn visual attributes shared across hierarchical levels when supervision is incomplete. At inference, our model is image-only and does not rely on VLMs or LLMs, since our goal is hierarchical prediction rather than label-driven reasoning.