A Taxonomy-Aware Evaluation for Open-Vocabulary Wildlife Detection

Wenqi Xue
University of Washington

Pengxi Zhang
University of Washington

William Wang
University of Toronto

Yijia Cai
Case Western Reserve University

Jieyu Zhang
University of Washington

Abstract

Open-vocabulary object detection (OVD) can recognize unseen categories via text prompts, but its ability to discriminate among semantically related concepts under long-tail conditions remains poorly understood. We introduce ***Wildlife OV-Bench****, a taxonomy-aware benchmark built from ImageNet-21K and V3Det, covering 3,586 animal categories organized by the WordNet hierarchy. Two complementary protocols are defined:* ***E1*** *evaluates node-level detection with macro-averaged F1 across the head-to-tail spectrum;* ***E2*** *augments prompts with semantically controlled hard negatives at varying WordNet distances to measure concept-level discrimination. Evaluating YOLO-World, Grounding DINO, OWLv2, and SAM3 under unified settings reveals a consistent pattern: strong baseline detection does not predict hard-negative robustness, and rare categories are systematically underserved.*

1. Introduction

Open-vocabulary object detection (OVD) extends conventional detection by recognizing categories specified at inference time through text prompts, rather than a fixed label set. Enabled by large-scale vision–language pretraining, recent OVD models transfer semantic knowledge from image–text corpora to detection and can localize categories beyond those seen during detector training. This capability is promising for real-world recognition, where category vocabularies are large and continually evolving.

Despite this progress, evaluating OVD reliability remains challenging. Most existing benchmarks inherit closed-set protocols and evaluate categories as flat labels without modelling semantic relationships [5]. Such summaries under-characterize two properties central to open-vocabulary deployment: (i) robustness under long-tailed category frequency, and (ii) discrimination among semantically related categories. In practice, detectors must not only find a queried category, but also reject plausible alternatives.

Animal recognition makes this gap concrete. When prompted with *impala*, a detector may return visually similar species such as *gazelle* or *antelope*. These predictions reflect coarse semantic alignment while exposing weak concept-level discrimination. Standard flat-label evaluation provides limited insight into this behavior because it does not explicitly model taxonomy-level relationships or isolate semantic ambiguity in a controlled setting.

To address this gap, we introduce **Wildlife OV-Bench**, a taxonomy-aware benchmark for analyzing OVD under both frequency imbalance and semantic ambiguity. Rather than measuring aggregate detection accuracy, Wildlife OV-Bench treats each taxonomic node as an independent concept and asks: *can a model reliably localise a concept in isolation, and does it remain discriminative when similar concepts are present?* Wildlife OV-Bench organizes categories using the WordNet hierarchy and combines ImageNet-21K animal categories with V3Det detection images and annotations, yielding 3,586 categories. This design enables controlled analysis across both category rarity and semantic proximity.

2. Dataset Construction

This section describes the data construction pipeline of Wildlife OV-Bench, including image collection, taxonomy-aware filtering, and annotation setup.

2.1. Image Collection

ImageNet-21K. The primary image source for Wildlife OV-Bench is ImageNet-21K, a large-scale dataset containing over 21,000 categories organized according to the WordNet lexical hierarchy. We select ImageNet-21K for two key reasons: it provides broad visual coverage suitable for long-tail analysis, and its synset-aligned labels enable direct integration with taxonomy-aware evaluation. Since

our benchmark focuses on animal recognition, we extract an animal-specific subset from ImageNet-21K. ImageNet-21K images are indexed by WordNet synsets, which correspond to semantic concepts in the WordNet taxonomy. Leveraging WordNet relations (e.g., hypernym and hyponym), we identify animal-related synsets by traversing the hierarchy and selecting all synsets belonging to the subtree rooted at `animal.n.01`. To perform this filtering at scale, we first collect metadata from dataset shards and retain only the fields required for taxonomy filtering (image identifier and class label). Synset prefixes are parsed from image identifiers, and taxonomy membership is evaluated once per unique synset and then propagated to all images belonging to that synset. We further manually remove ambiguous or non-informative synsets (e.g., overly generic terms or developmental-stage categories such as "baby"). After this filtering process, the resulting subset contains 3,586 animal categories and 17,373 images, forming the taxonomy foundation of Wildlife OV-Bench. We define long-tail structure using per-category image frequency in the ImageNet-21K animal subset. Categories are sorted by frequency, the least frequent one-third are treated as long-tail categories, and the remaining two-thirds are treated as head categories.

V3Det. While ImageNet-21K provides extensive taxonomy coverage, it is primarily designed for image classification rather than object detection. To complement the taxonomy derived from ImageNet-21K, we incorporate images and annotations from V3Det, a large-scale dataset designed for vast-vocabulary object detection. To align V3Det with the taxonomy defined by our ImageNet-21K animal subset, we identify overlapping categories between the two datasets. This process yields 1,687 overlapping animal categories (3860 images) between ImageNet-21K and V3Det. Using these categories, we filter the V3Det annotations to retain only images and bounding boxes corresponding to the selected classes. The resulting subset introduces additional detection annotations while preserving the taxonomy structure defined by ImageNet-21K.

2.2. Annotation Collection

Bounding-box annotations for ImageNet-21K-derived images are collected via a crowdsourced workflow with two-annotator review. We use the first WordNet lemma of each synset as the canonical category name, both for annotation labels and text prompts at inference time (see Fig. 1). For V3Det images, we adopt the official bounding-box annotations directly (see Fig. 1, bottom).

2.3. Dataset Statistics

Wildlife OV-Bench covers 3,586 animal categories and 21,233 images (17,373 from ImageNet-21K; 3,860 from V3Det across 1,687 overlapping categories). Including 413 WordNet-only intermediate nodes, the full taxonomy comprises 3,999 nodes; after bottom-up propagation, 3,650 of these contain at least one image and are evaluated. Categories are split into *common* (top two-thirds by ImageNet-21K frequency) and *rare* (bottom one-third).

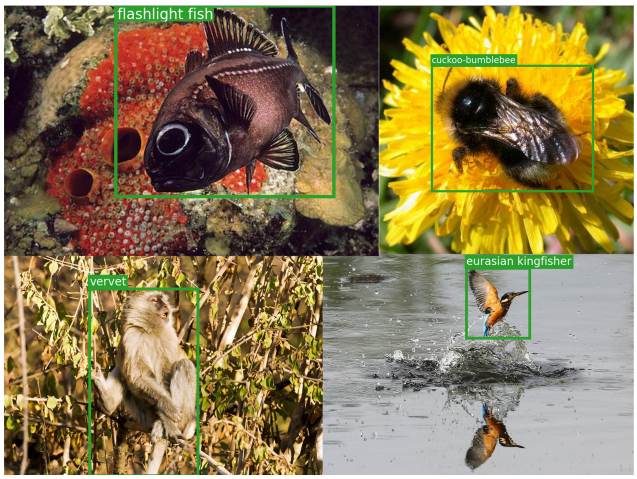

Figure 1. Example annotations in Wildlife OV-Bench. Top: ImageNet-21K images with fine-grained WordNet lemmas (*flashlight fish*, *cuckoo-bumblebee*). Bottom: V3Det images with official bounding-box annotations (*vervet*, *Eurasian kingfisher*).

2.4. Link Dataset to Taxonomy

Wildlife OV-Bench organizes animal categories into a taxonomy $\mathcal{T}$ over WordNet [11], where the taxonomy forms a directed acyclic graph (DAG) in which each edge encodes a hypernym–hyponym relation, enabling evaluation at every level of semantic granularity, from coarse (*mammal*) to fine (*Siberian husky*). A key challenge is that many internal nodes in $\mathcal{T}$ are WordNet-only concepts with no associated images in ImageNet-21K. To support evaluation at these nodes, we associate images via bottom-up propagation processed in reverse topological order (leaves first): $I_c = \bigcup_{d \in \{c\} \cup \mathrm{desc}(c)} \mathrm{imgs}(d)$, where $\mathrm{desc}(c)$ are all descendants of c in $\mathcal{T}$ and $\mathrm{imgs}(d)$ the ImageNet-21K images for node d. Duplicate images arising from multi-parent nodes are removed. This propagation ensures that every node with at least one image in its subtree participates in evaluation.

3. Evaluation

Models and Setup. We benchmark four families of open-vocabulary detectors under unified settings: **Grounding DINO** [10] (Tiny, Base), **OWLv2** [13] (Base, Base-Ensemble, Large), **YOLO-World** [3] (S, M, L, X), and **SAM3** [2]. Each model is queried with the first WordNet lemma of the target synset as the sole text prompt. Confidence thresholds are fixed per model: 0.05 for YOLO-

Table 1. Main results. E1 and E2 $F_{1,\text{macro}}$ are reported as mean $\pm$ std over three seeds (42, 123, 456) unless noted. E1 Rare and E1 Common report macro-averaged F1 over rare and common categories, respectively. Gap = Common $-$ Rare. E2/E1 measures robustness to hard negatives. †Single seed only (seed 123). ‡Two seeds (42, 123). *E2/E1 ≈ 1 reflects prompt insensitivity, not genuine robustness (see Sec. 4); excluded from bold.

Model	E1 F_1	E1 Rare	E1 Common	Gap	E2 F_1	E2/E1
Grounding DINO Base	$\mathbf{0.876}_{\pm 0.000}$	**0.874**	**0.878**	**0.004**	$0.223_{\pm 0.001}$	0.25
Grounding DINO Tiny	$0.807_{\pm 0.000}$	0.797	0.812	0.015	$0.234_{\pm 0.001}$	0.29
OWLv2 Large	$0.692_{\pm 0.001}$	0.616	0.736	0.120	$\mathbf{0.538}^{\dagger}$	0.78
OWLv2 Large Ensemble	$0.682_{\pm 0.001}$	0.584	0.737	0.154	$0.536^{\dagger}$	**0.79**
OWLv2 Base Ensemble	$0.684_{\pm 0.001}$	0.579	0.743	0.165	$0.521_{\pm 0.001}$	0.76
OWLv2 Base	$0.666_{\pm 0.001}$	0.570	0.719	0.149	$0.512_{\pm 0.001}$	0.77
SAM3	$0.493_{\pm 0.001}$	0.333	0.577	0.245	$0.494^{\ddagger}_{\pm 0.001}$	1.00^{*}
YOLOv8x-World	$0.486_{\pm 0.001}$	0.374	0.545	0.171	$0.307_{\pm 0.000}$	0.63
YOLOv8l-World	$0.403_{\pm 0.000}$	0.266	0.475	0.209	$0.245_{\pm 0.001}$	0.61
YOLOv8s-World	$0.276_{\pm 0.001}$	0.159	0.337	0.178	$0.179_{\pm 0.001}$	0.65
YOLOv8m-World	$0.263_{\pm 0.000}$	0.141	0.327	0.185	$0.205_{\pm 0.001}$	0.78

World, 0.1 for Grounding DINO, 0.2 for OWLv2, and 0.25 for SAM3. To account for sampling variability, experiments are repeated over three random seeds (42, 123, 456) and reported as mean ± std unless otherwise noted in Tab. 1.

Metric. We adopt *node-level macro-averaged Precision, Recall, and F1* as primary metrics. Macro-averaging over nodes gives equal weight to every concept regardless of its image count, exposing long-tail failures that frequency-weighted metrics would mask. For each image, predictions are filtered to those whose label matches the target class. Ground-truth and predicted boxes are matched greedily GT-first: each ground-truth box is paired with the highest-IoU unmatched prediction, and the pair counts as a true positive (TP) if IoU ≥ 0.5. Precision is TP divided by the number of correct-label predictions; recall is TP divided by the number of ground-truth boxes; F1 is their harmonic mean.

E1: Baseline Across the Taxonomy E1 evaluates baseline node-level detection across the taxonomy under unified prompting and matching settings, enabling consistent head-to-tail analysis from common to rare categories. It measures each model's baseline ability to localise a concept when given its name as the sole prompt—no competing categories. For each node $c \in \mathcal{T}$, up to k=10 images are sampled from I_c to form $\tilde{I}_c$, and the model is queried using only c's category name. This includes WordNet-only intermediate nodes whose images are inherited via the bottom-up propagation described in Sec. 2.4.

Node-level Precision, Recall, and F1 are obtained by averaging the corresponding per-image metrics over $\tilde{I}_c$: $\text{F1}_c = \frac{1}{|\tilde{I}_c|} \sum_{i \in \tilde{I}_c} F_{1,c,i}$, with P_c and R_c defined analogously. We report all three per-node metrics for analysis; the scalar **E1 score** used for model comparison is the macro-averaged F1 over all evaluated nodes $\mathcal{C} = \{c \mid \tilde{I}_c \neq \emptyset\}$: $\text{E1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{F1}_c$. This two-level averaging gives equal weight to every node regardless of its image count, ensuring that long-tail nodes are not dominated by frequent ones.

E2: Hard-Negative Discrimination Across the Taxonomy E2 tests whether a model can *identify* a target concept while *rejecting* semantically similar distractors by augmenting target prompts with taxonomy-guided hard negatives controlled by semantic distance. Using the same $\tilde{I}_c$ from E1, the prompt is augmented with N hard negatives sampled at exact WordNet distance D from c, excluding ancestors, descendants, and ancestors of descendants (nodes that may be partially correct in the multi-parent DAG). We systematically vary $D \in \{2, 5, 10\}$ (near to distant semantic neighbours) and $N \in \{2, 5, 10\}$ (few to many distractors), forming a 3×3 difficulty grid. The **E2 score** is computed in three levels: per-node F1 is first averaged over images, then averaged over all valid grid cells (cells where fewer than N valid negatives exist at distance D are excluded), and finally macro-averaged over all nodes $\mathcal{C}$.

4. Results and Findings

Main Results Tab. 1 reports E1 and E2 $F_{1,\text{macro}}$ scores. All models show high stability across seeds (std < 0.002), confirming that our protocol yields reliable estimates.

The most striking result is the divergence between E1 and E2 scores. Grounding DINO Base achieves the highest E1 score (0.876), yet its E2 score drops to 0.223 (E2/E1 = 0.25)—retaining roughly a quarter of its baseline performance under hard-negative conditions. By contrast, OWLv2 models score lower on E1 but retain substantially more: OWLv2 Base (E1 = 0.666) achieves E2/E1 = 0.77, OWLv2 Large (E1 = 0.692) reaches E2/E1 = 0.78, and

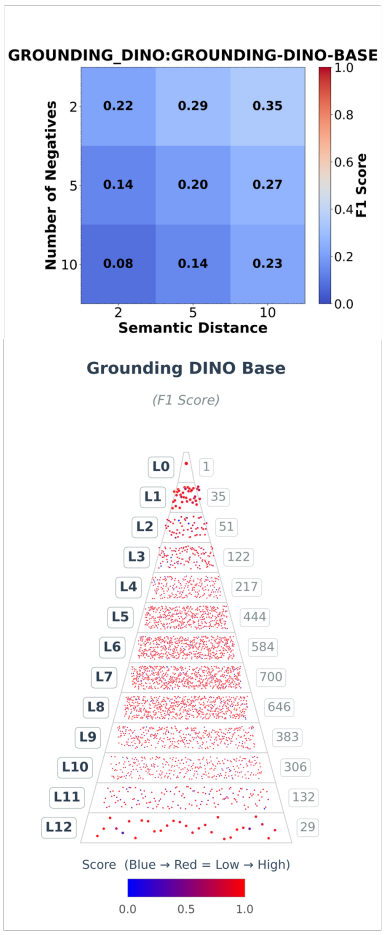

Figure 2. Grounding DINO Base: E2 F_1 heatmap across the (D, N) grid (top) and E1 per-node depth distribution across WordNet taxonomy (bottom). Colour encodes F1 (blue = low, red = high).

OWLv2 Large Ensemble attains E2/E1 = 0.79—the highest ratio among non-SAM3 models. Fig. 2 visualises the full (D, N) grid for Grounding DINO Base. Grounding DINO Base degrades 77% from $(D{=}10, N{=}2)$ to $(D{=}2, N{=}10)$ (0.35→0.08), collapsing to near-zero when distractors are semantically close and numerous. OWLv2 Large and YOLO-World X show similar trends (see Figs. 7 and 10 in the appendix). Strong baseline detection does not imply robustness to semantically confusable distractors.

Finding 2: Systematic Bias Against Rare Categories All models exhibit a consistent performance gap between common and rare categories on E1. Grounding DINO Base shows the smallest gap (0.004), indicating near-uniform coverage across the head-to-tail spectrum. OWLv2 models sit between the two extremes, with gaps ranging from 0.12 (OWLv2 Large) to 0.165 (OWLv2 Base Ensemble). YOLO-World models show larger gaps (0.17–0.21), and SAM3 exhibits the widest gap (0.245), with rare-category F1 (0.333) reaching only 58% of its common-category F1 (0.577). Fig. 2 visualises per-node F1 scores across WordNet depth levels; Grounding DINO Base shows consistently high-F1 (red) nodes throughout the hierarchy, while other models exhibit substantial variation (see Figs. 8, 11 and 17 in the appendix).

Finding 3: SAM3 Is Insensitive to Hard Negatives SAM3's E2 score (0.494) matches its E1 score (0.493), yielding E2/E1 $\approx$ 1.00—the highest ratio across all evaluated models, and the only model where E2 does not fall below E1. As shown in Fig. 16 in the appendix, its E2 heatmap is visually uniform across all (D, N) configurations (0.50–0.51)—in stark contrast to every other model. This is inconsistent with genuine fine-grained discrimination, which would produce lower scores as distractors become closer or more numerous. We hypothesize that SAM3's scoring mechanism assigns near-uniform confidence across text prompts, making it insensitive to prompt composition—consistent with its training objective of open-world segmentation rather than fine-grained category separation. The per-node E2 pyramid (see Fig. 5 in the appendix) corroborates this at the per-node level. Grounding DINO Base's E2 pyramid is dominated by blue nodes across *all* depth levels, confirming that its E2 collapse is not restricted to rare leaves but is a model-wide failure under hard negatives. OWLv2 Large retains a visible proportion of red nodes, consistent with its higher E2/E1 ratio. SAM3's pyramid retains a persistent mix of blue and red nodes similar to its E1 pyramid—consistent with a model whose scores are insensitive to prompt composition.

4.1. Summary

Grounding DINO leads on E1 with the smallest rare/common gap but collapses under hard negatives (E2/E1 = 0.25); **OWLv2** offers the best E2 robustness (E2/E1 up to 0.79), making it the most balanced family; **YOLO-World** trails on both axes; **SAM3** achieves E2/E1 $\approx$ 1.00 but not through genuine discrimination—its scores are near-uniform regardless of distractor difficulty. E1 alone is an insufficient proxy for open-vocabulary robustness, motivating E2 as a complementary diagnostic.

5. Conclusion

Wildlife OV-Bench evaluates open-vocabulary detectors across 3,586 animal categories via E1 (baseline detection) and E2 (hard-negative discrimination). Strong E1 does not predict E2 robustness, rare categories are systematically underserved, and failure modes such as SAM3's prompt insensitivity remain invisible to flat-label benchmarks. We hope this spurs future work on robust, taxonomy-aware detection.

References

[1] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks, 2020. 1

[2] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. 2, 1

[3] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection, 2024. 2, 1

[4] Simon Ging, María A. Bravo, and Thomas Brox. Open-ended vqa benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy, 2024. 1

[5] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 1

[6] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset, 2018. 1

[7] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities, 2023. 1

[8] Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3): 820–865, 2014. 1

[9] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2022. 1

[10] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 2, 1

[11] George A. Miller. WORDNET: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993. 2

[12] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. 1

[13] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2024. 2, 1

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[15] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 1

[16] 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, 2015. 1

[17] Vésteinn Snæbjarnarson, Kevin Du, Niklas Stoehr, Serge Belongie, Ryan Cotterell, Nico Lang, and Stella Frank. Taxonomy-aware evaluation of vision-language models, 2025. 1

[18] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset, 2023. 1

A Taxonomy-Aware Evaluation for Open-Vocabulary Wildlife Detection

Supplementary Material

6. Related Work

Open-Vocabulary Object Detection: Models and Benchmarks. CLIP [14] demonstrated that contrastive vision–language pretraining enables zero-shot transfer to open label sets, but it is an image-level model rather than a detector. Subsequent models extended language-conditioned visual recognition to detection: OWL-ViT [12] and OWLv2 [13] transfer image–text models to open-vocabulary detection; GLIP [9] and Grounding DINO [10] combine grounding with detection pretraining; YOLO-World [3] emphasises real-time inference. SAM3 [2] extends the paradigm to promptable concept segmentation, accepting text, image exemplars, or combinations as queries. Despite this diversity of architectures, evaluation practice has changed little: AP/mAP computed against fixed label sets—as on COCO, then scaled to LVIS [5] and V3Det [18]—remains dominant. LVIS addressed frequency imbalance through federated annotation; V3Det extended vocabulary to 13,204 categories and provides a hierarchical category tree; yet both inherit the flat-label assumption from their COCO lineage. Under this assumption, flat category matching does not distinguish semantically adjacent mistakes from unrelated ones—detecting *gazelle* when the target is *impala* costs as much as detecting *chair*—a design that obscures concept-level discrimination in open-vocabulary settings.

Hierarchy and Taxonomy in Visual Recognition. Hierarchically organised label spaces have a long history in visual recognition. ImageNet [16] and its 21K extension [15] organise categories around WordNet synsets, while iNaturalist [6] provides fine-grained species recognition under a biological taxonomy; however, neither is designed as a taxonomy-aware evaluation benchmark for open-vocabulary detection. Bertinetto *et al*. [1] propose hierarchy-aware classification losses that distinguish near mistakes from far ones, and Kosmopoulos *et al*. [8] develop evaluation measures for hierarchical classification; neither models multiple box predictions or spatial localisation and therefore neither transfers directly to detection. OVEN [7] scales visual recognition to a shared Wikipedia entity space spanning millions of entities, but it does not evaluate controlled semantic confounders of the kind studied here. The most closely related evaluation frameworks are taxonomy-aware VLM assessments: Ging *et al*. [4] use semantic hierarchies to generate follow-up questions for fine-grained VQA assessment, and Snæbjarnarson *et al*. [17] map free-form VLM predictions onto a taxonomy before applying hierarchical precision and recall. Both operate over open-ended textual outputs rather than bounding-box predictions; in particular, Snæbjarnarson *et al*. show that standard similarity measures do not reliably capture taxonomic structure. Wildlife OV-Bench fills this gap by bringing taxonomy-aware evaluation into the detection setting, where a prediction is only correct if it names the right concept *and* localises it spatially.

Fine-Grained Discrimination and Hard Negatives. Hard-negative mining is a well-established *training* technique: contrastive objectives rely on carefully selected negative pairs to shape the embedding space. As an *evaluation* dimension, however, controlled semantic interference has been studied less systematically. Related work in image–text matching has also explored hard negatives constructed by perturbing objects or attributes in captions, but these benchmarks target cross-modal matching and do not require spatial localisation. In VQA, Ging *et al*. [4] use semantic hierarchies to generate follow-up questions for fine-grained assessment—conceptually the closest precedent to our design—but scoring relies on open-ended language-model outputs rather than on spatial predictions. Snæbjarnarson *et al*. [17] similarly show that without taxonomy-aware scoring, subtle errors in specificity go undetected; yet their setting again presupposes free-form text generation. To our knowledge, existing object detection benchmarks do not systematically vary both the *semantic proximity* and the *count* of distractors as independent evaluation axes. Our E2 protocol addresses this directly: hard-negative distractors are sampled from the WordNet graph at varying shortest-path distances D and counts N, making difficulty fully reproducible, interpretable, and decoupled from any language-model oracle. Wildlife OV-Bench is, to our knowledge, the first benchmark to jointly address long-tail category frequency, bounding-box localisation, and parameterised taxonomy-aware discrimination in open-vocabulary detection.

7. Discussion and Limitations

Evaluating open-vocabulary detectors under long-tail and semantically structured conditions requires going beyond the flat-label protocols that dominate current benchmarks. In this paper, we introduce structure into OVD evaluation by grounding both protocols in the WordNet animal taxonomy: E1 measures baseline concept-level coverage, and E2 probes discriminability by varying the proximity and count of hard-negative distractors. Our results reveal that E1 and E2 capture largely orthogonal aspects of model capability—models with the highest baseline detection (Grounding DINO) collapse under hard negatives,

while models with contrastive pretraining (OWLv2) retain far more discriminative robustness. SAM3's near-uniform E2 scores further suggest that E2 can diagnose whether a model meaningfully adjusts its confidence under changes in prompt composition, a failure mode that is not visible under E1 alone. Despite this, several limitations remain.

First, the E2 grid is not fully populated for all nodes. Filling a cell (D, N) requires at least N taxonomy neighbours at exact WordNet distance D from the target; taxonomically isolated categories with few semantic neighbours cannot satisfy the hardest configurations (small D, large N). In our benchmark, 74.3% of nodes are evaluated under fewer than the full nine (D, N) cells, introducing a slight upward bias in E2 scores for isolated categories. Since coverage is determined by taxonomy structure rather than model behaviour, relative rankings are unaffected. Future work could address this through approximate-distance sampling or by restricting evaluation to the subset of nodes with full grid coverage.

Second, Wildlife OV-Bench is scoped to animal categories. This provides a coherent taxonomy with well-defined hierarchical distances, but limits the generalisability of findings to other fine-grained domains such as plant species, vehicle sub-types, or medical imaging. Whether the E1–E2 divergence and depth-dependent degradation patterns observed here transfer to domains with different taxonomic structures and training-data distributions remains an open question.

Third, each category is queried using only the first WordNet lemma of its synset as the text prompt, without templates or synonym augmentation. This ensures reproducibility and avoids prompt-engineering confounds, but may disadvantage models whose pretraining associated a concept primarily with an alternative name or a descriptive phrase. Extending E1 and E2 to multi-synonym or template-augmented prompting is a natural direction for future work.

Finally, confidence thresholds are fixed per model family and held constant across E1 and E2 conditions; introducing hard-negative prompts may shift a model's score distribution such that a threshold calibrated for E1 is no longer optimal. Similarly, all evaluations use a single IoU threshold of 0.5, which may penalise models that produce well-centred but slightly loose boxes for rare, small-scale species. A threshold-free or multi-threshold evaluation protocol would provide a more complete picture of model behaviour and is left to future work.

8. Taxonomy Visualisation

Fig. 3 shows the full WordNet animal taxonomy used in Wildlife OV-Bench as a radial tree, with node colour indicating dataset membership (common, rare, or WordNet-only intermediate). Fig. 4 illustrates the semantic distance metric D used in E2 on a representative subtree, showing how shortest-path distance separates near siblings (D=2) from more distant cousins (D=4).

9. Per-Node E2 Pyramid (Large Variants)

Fig. 5 visualises per-node E2 F_1 across WordNet depth for the largest variant of each model family, complementing the E1 pyramid shown in the main paper.

10. Additional Per-Variant Results

The main paper reports results for the largest variant of each model family. Here we present the full per-variant breakdown for all four families: OWLv2 (Sec. 10.2), YOLO-World (Sec. 10.3), Grounding DINO (Sec. 10.4), and SAM3 (Sec. 10.5). Each family is shown with E2 F_1 heatmaps and per-node F_1 scatter pyramids for both E1 and E2, enabling direct within-family scaling comparisons.

10.1. E2 Performance Trends (Large Variants)

Fig. 6 shows how E2 F_1 varies as a function of semantic distance D and distractor count N for the largest variant of each family. OWLv2 Large benefits substantially from larger D, while Grounding DINO Base and YOLO-World X decline monotonically on both axes; SAM3 remains flat throughout.

10.2. OWLv2 Variants

Figs. 7–9 show E2 heatmaps and per-node pyramids for all four OWLv2 variants. Ensemble variants consistently outperform their non-ensemble counterparts, and Large variants retain more red nodes under E2 than Base variants.

10.3. YOLO-World Variants

Figs. 10–12 present results for all four YOLO-World variants (S, M, L, X). Performance scales generally with model size on both E1 and E2, and all variants show a global shift toward blue nodes under hard-negative conditions.

10.4. Grounding DINO Variants

Figs. 13–15 show results for Grounding DINO Tiny and Base. Both variants achieve near-uniform red coverage under E1 but collapse to blue under E2, confirming that their strong baseline detection does not confer robustness to hard-negative distractors.

10.5. SAM3

Figs. 16–18 show the E2 heatmap and per-node pyramids for SAM3. The near-uniform heatmap and the near-identical E1 and E2 pyramids together confirm the prompt-insensitive behavior discussed in Finding 3.

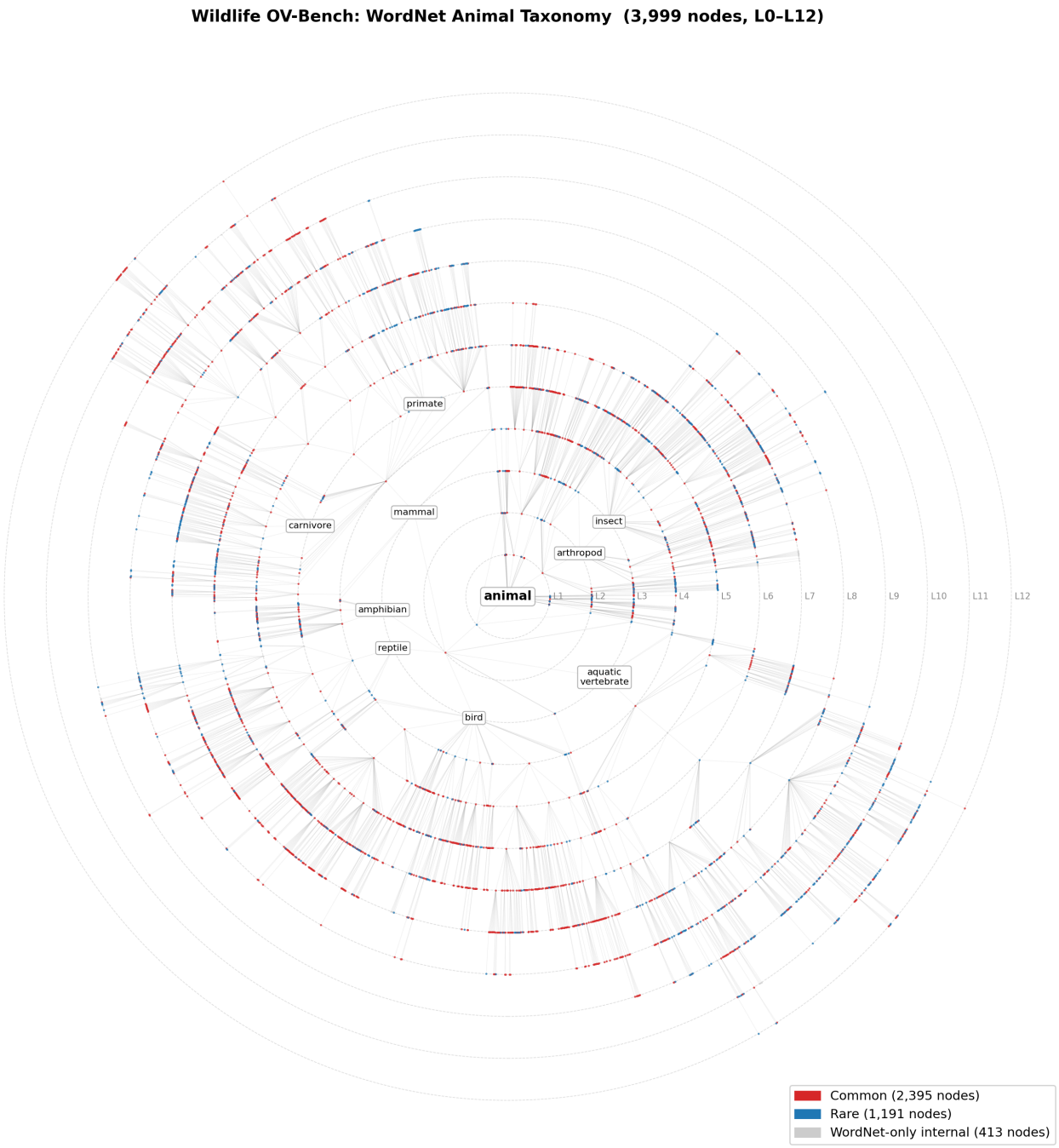

Figure 3. Full WordNet animal taxonomy used in Wildlife OV-Bench, rendered as a radial tree (3,999 nodes, max depth L12). Root (`animal.n.01`) is at the centre; radial distance encodes WordNet depth. Node colour indicates dataset membership: red = common categories (2,395), blue = rare categories (1,191), grey = WordNet-only nodes with no direct ImageNet images (413). Major branches are annotated. The long-tail structure is visible in the dense fine-grained leaves at deeper levels.

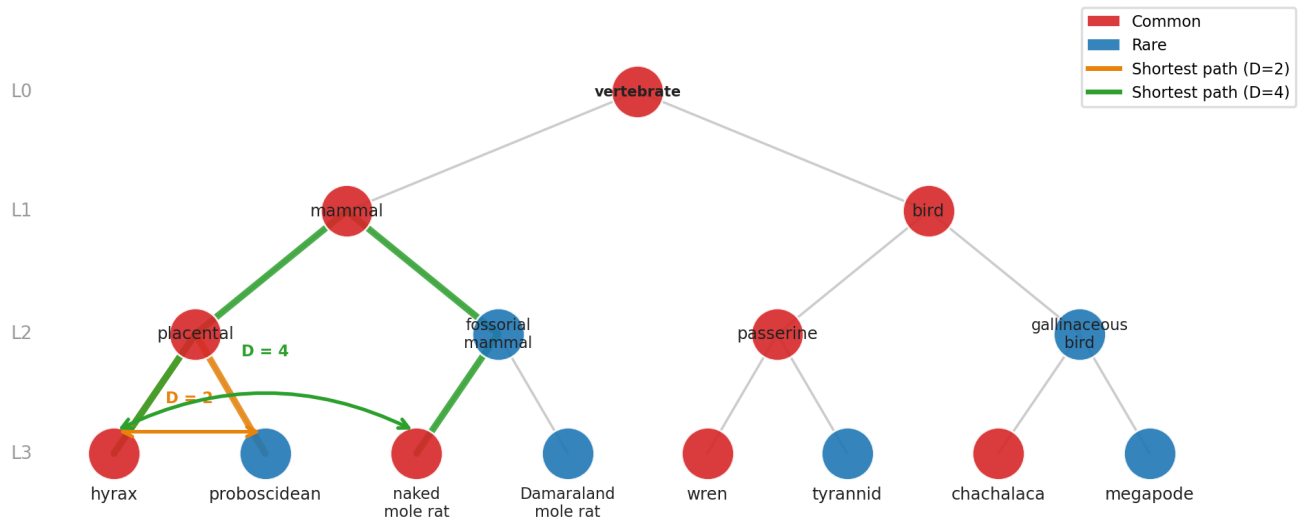

Figure 4. Representative 4-level subtree of the WordNet animal taxonomy illustrating the semantic distance metric used in E2. Each node is coloured by dataset membership (red = common, blue = rare). D denotes the shortest undirected path length between two nodes. Orange path (D=2): siblings *hyrax* and *proboscidean* share the same parent (*placental*). Green path (D=4): cousins *hyrax* and *naked mole rat* share a common grandparent (*mammal*). E2 hard negatives are sampled at exact distances $D \in \{2, 5, 10\}$, providing a controlled difficulty axis from semantically close (D=2) to distant (D=10) distractors. All edges shown are strict direct parent–child relations in the WordNet graph; D is the unweighted shortest-path length in the undirected version of the graph.

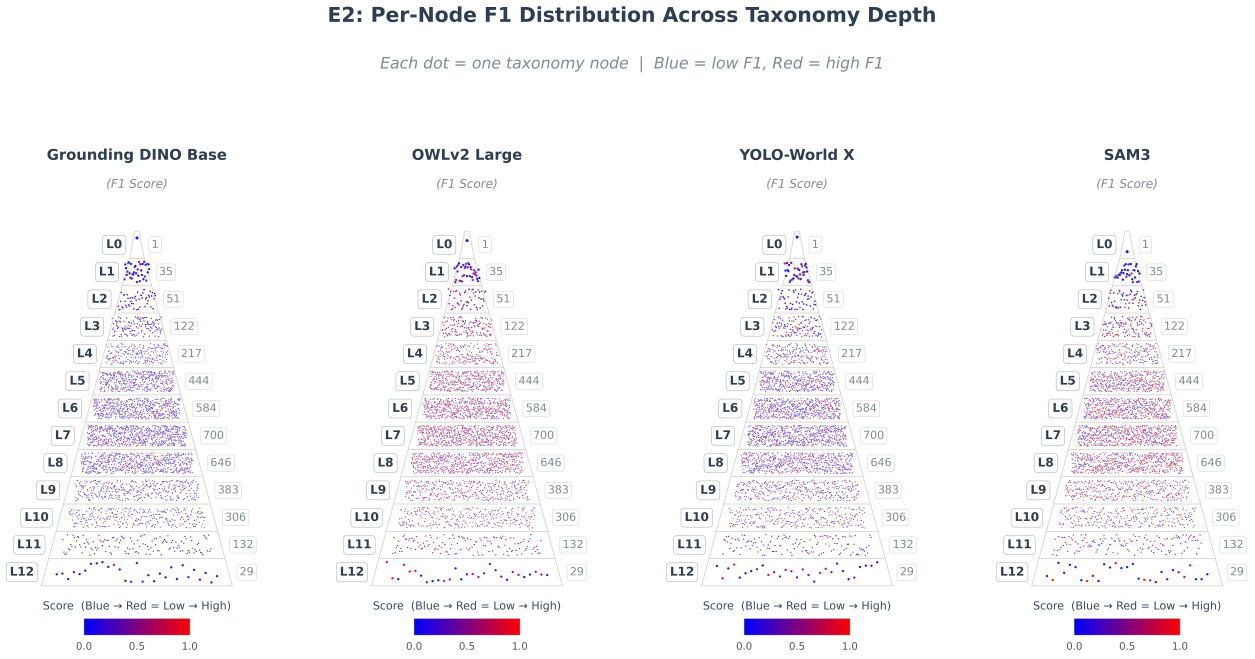

Figure 5. Per-node E2 F_1 distribution across WordNet taxonomy depth for the largest variant of each model family. Each dot represents one evaluated taxonomy node; colour encodes node-level E2 F1 (blue = low, red = high). Grounding DINO Base shows near-uniform blue nodes at all depths, revealing a model-wide E2 collapse rather than a depth-specific one. OWLv2 Large retains a proportion of red nodes, reflecting its superior hard-negative robustness. YOLO-World X is predominantly blue across the hierarchy. SAM3 exhibits a persistent mix of blue and red nodes similar to its E1 distribution, consistent with near-uniform confidence assignment across text prompts.

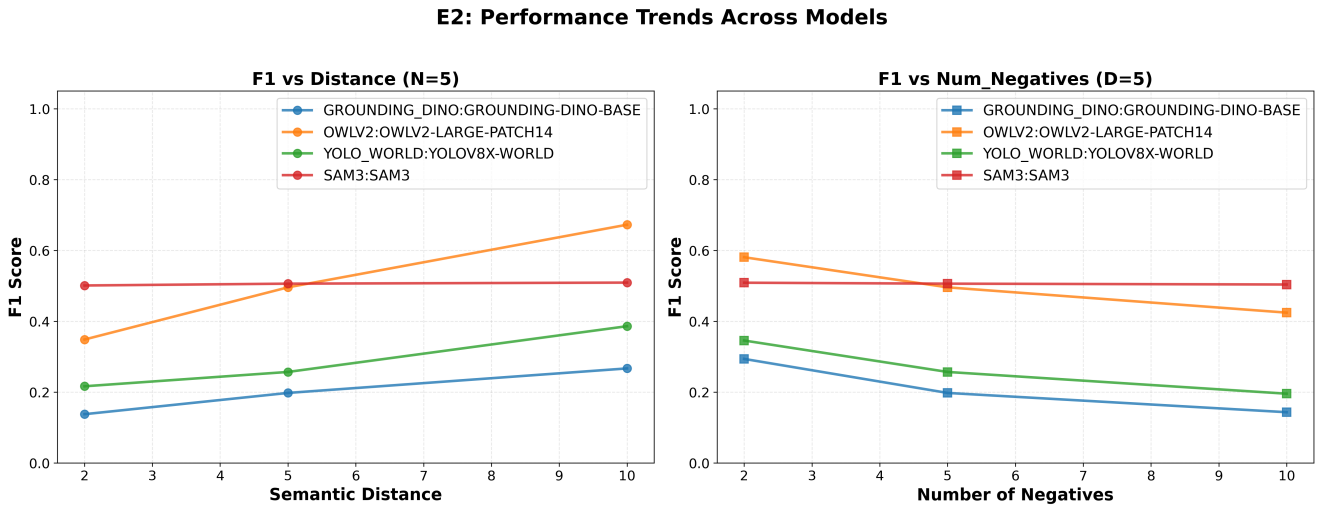

Figure 6. E2 F_1 trend lines for the largest variant of each model family, shown as a function of semantic distance D (left, with N=5 fixed) and number of distractors N (right, with D=5 fixed). OWLv2 Large improves markedly as D increases, confirming that it can discriminate concepts when distractors are semantically distant but degrades under close-proximity interference. Grounding DINO Base and YOLO-World X decline monotonically with both harder axes. SAM3 remains flat across all configurations, consistent with its prompt-insensitive scoring (see Sec. 4).

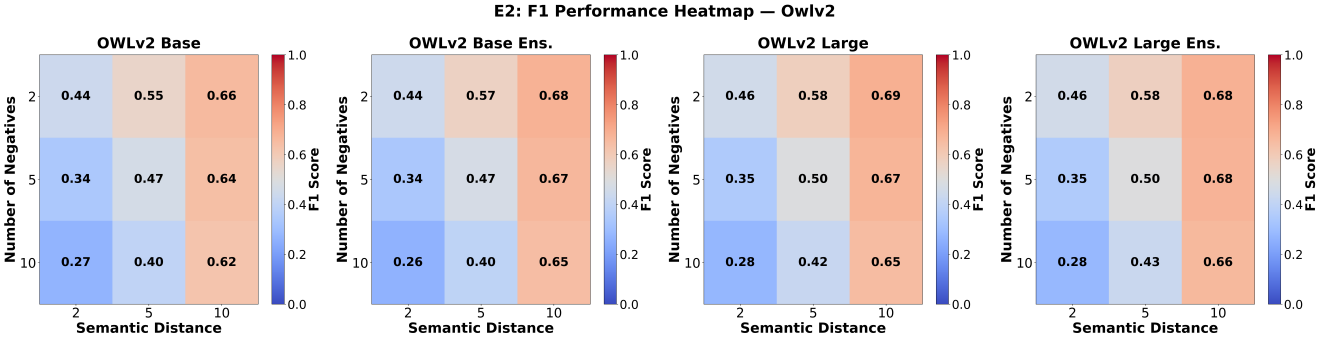

Figure 7. E2 F_1 heatmaps across the (D, N) grid for all OWLv2 variants. Columns correspond to semantic distance $D \in \{2, 5, 10\}$; rows to number of distractors $N \in \{2, 5, 10\}$. Smaller D and larger N denote harder conditions. Colour encodes F_1 score (blue = low, red = high). Ensemble variants consistently outperform their non-ensemble counterparts, and the Large variants show stronger discrimination at close semantic distances (D=2) than the Base variants.

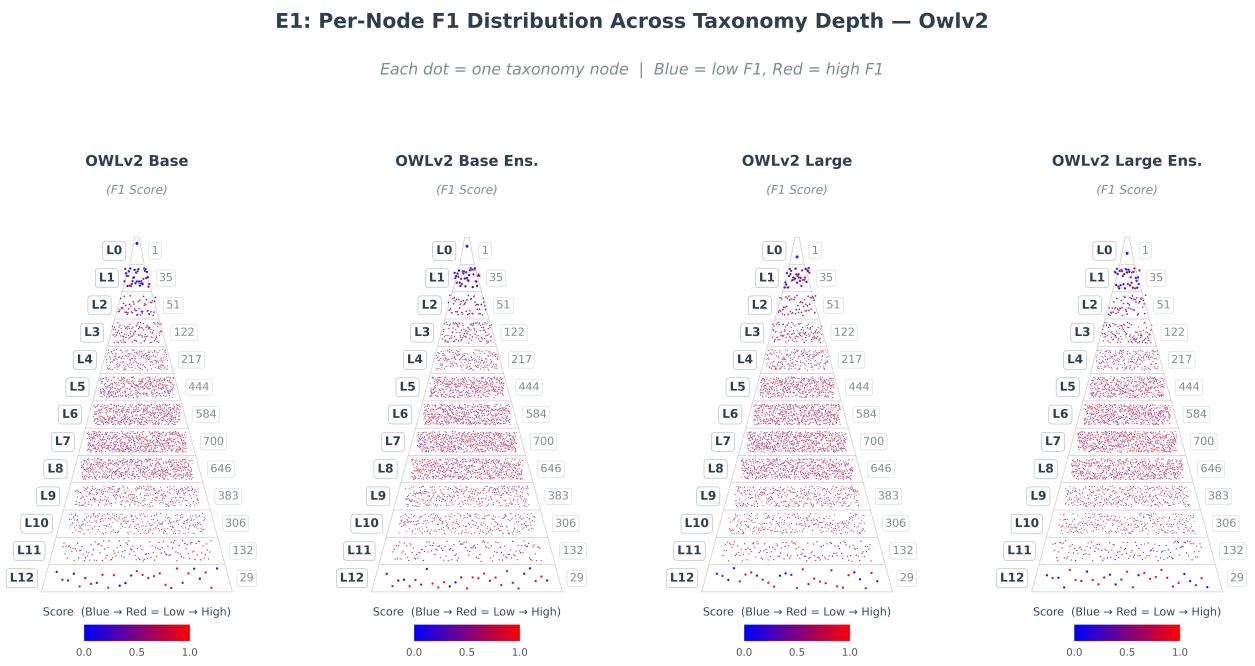

Figure 8. Per-node E1 F_1 distribution across WordNet taxonomy depth for all OWLv2 variants. Each dot represents one evaluated taxonomy node; colour encodes node-level F_1 (blue = low, red = high). Numbers on the right of each layer give the node count at that depth. Larger and ensemble variants show a generally higher proportion of red nodes throughout the hierarchy, indicating better overall category coverage.

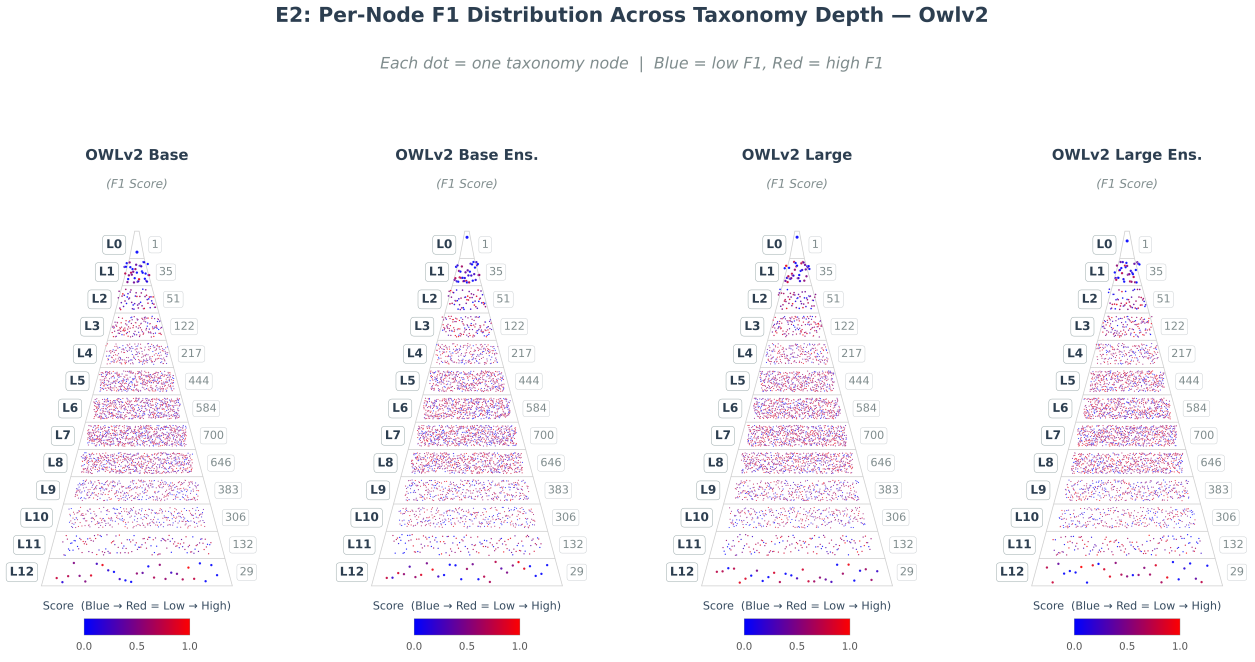

Figure 9. Per-node E2 F_1 distribution across WordNet taxonomy depth for all OWLv2 variants. The shift from red to blue nodes relative to E1 (Fig. 8) reflects degradation under hard-negative distractors. OWLv2 Large and its ensemble variant exhibit the most resilient distributions, retaining a visible proportion of red nodes across the hierarchy, while Base variants degrade more uniformly.

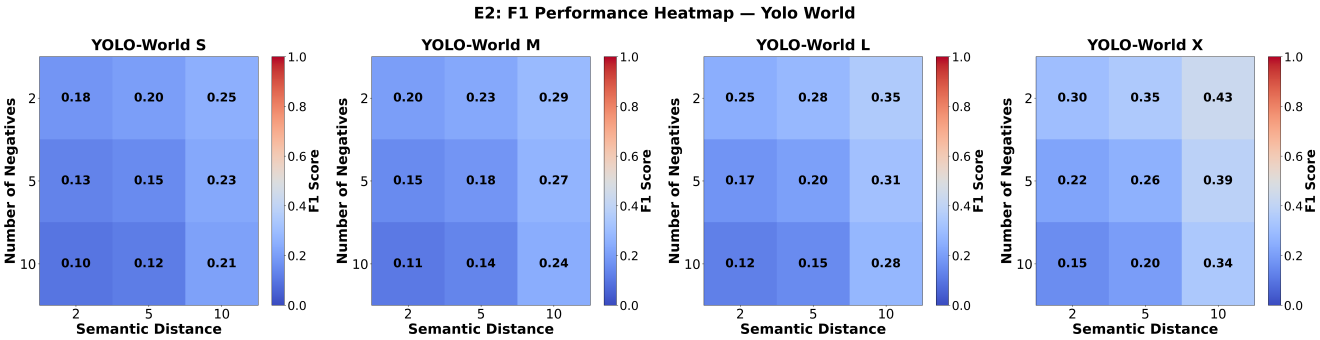

Figure 10. E2 F_1 heatmaps across the (D, N) grid for all YOLO-World variants (S, M, L, X). Colour encodes F_1 score (blue = low, red = high). Performance scales generally with model size, with YOLO-World X showing the most favourable gradient across the (D, N) grid. All variants exhibit notable degradation at $(D{=}2, N{=}10)$, confirming that close-proximity distractors remain the dominant challenge regardless of scale.

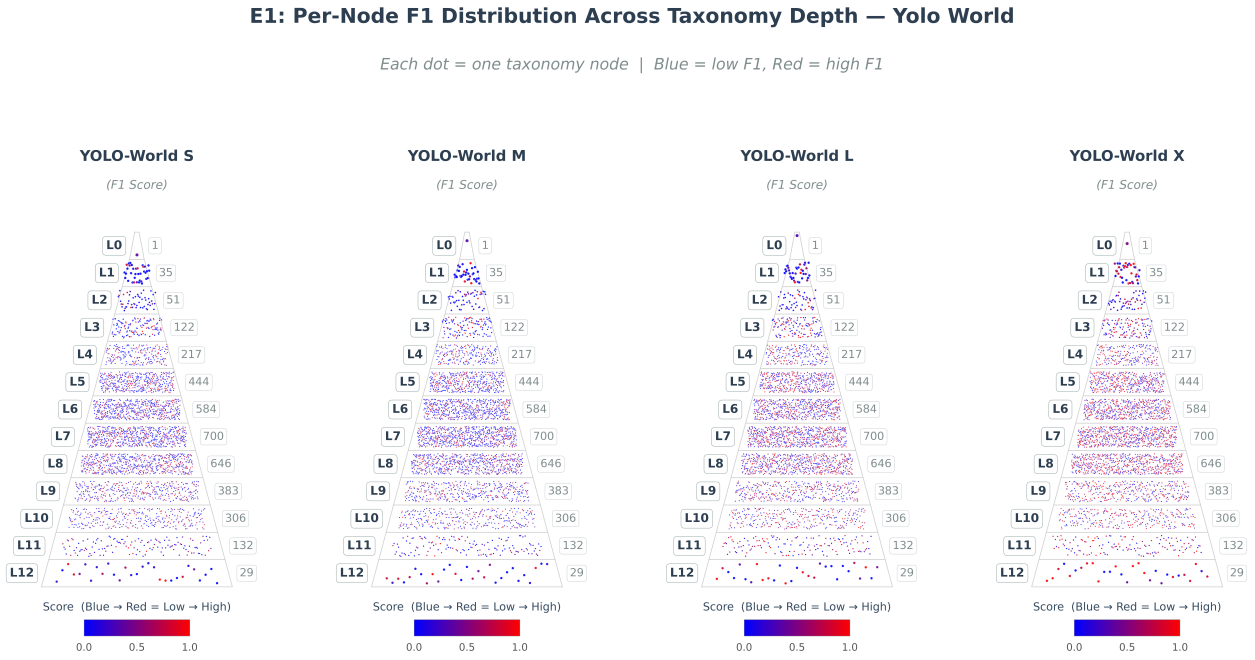

Figure 11. Per-node E1 F_1 distribution across WordNet taxonomy depth for all YOLO-World variants. Larger models show a higher proportion of red nodes across the hierarchy, reflecting the performance scaling visible in the main results table.

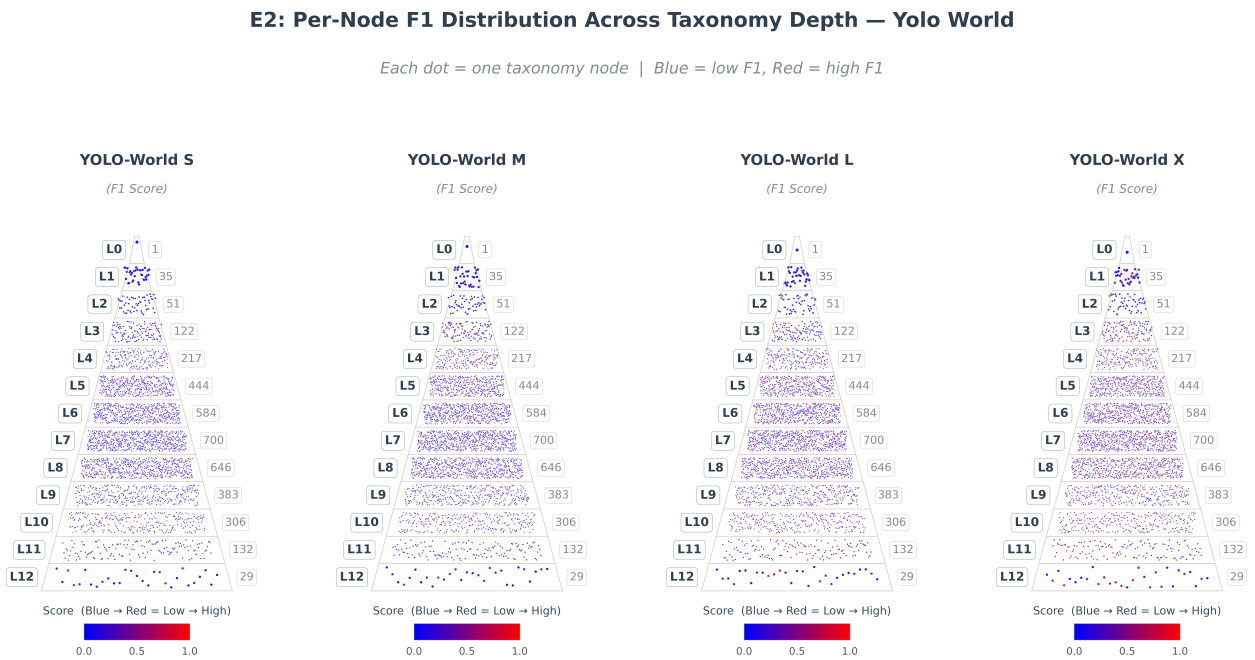

Figure 12. Per-node E2 F_1 distribution across WordNet taxonomy depth for all YOLO-World variants. Compared to E1 (Fig. 11), all variants show a global shift toward blue nodes under hard-negative conditions. YOLO-World X retains the highest overall proportion of red nodes among the four variants, consistent with its higher E2 F_1 in the main results table.

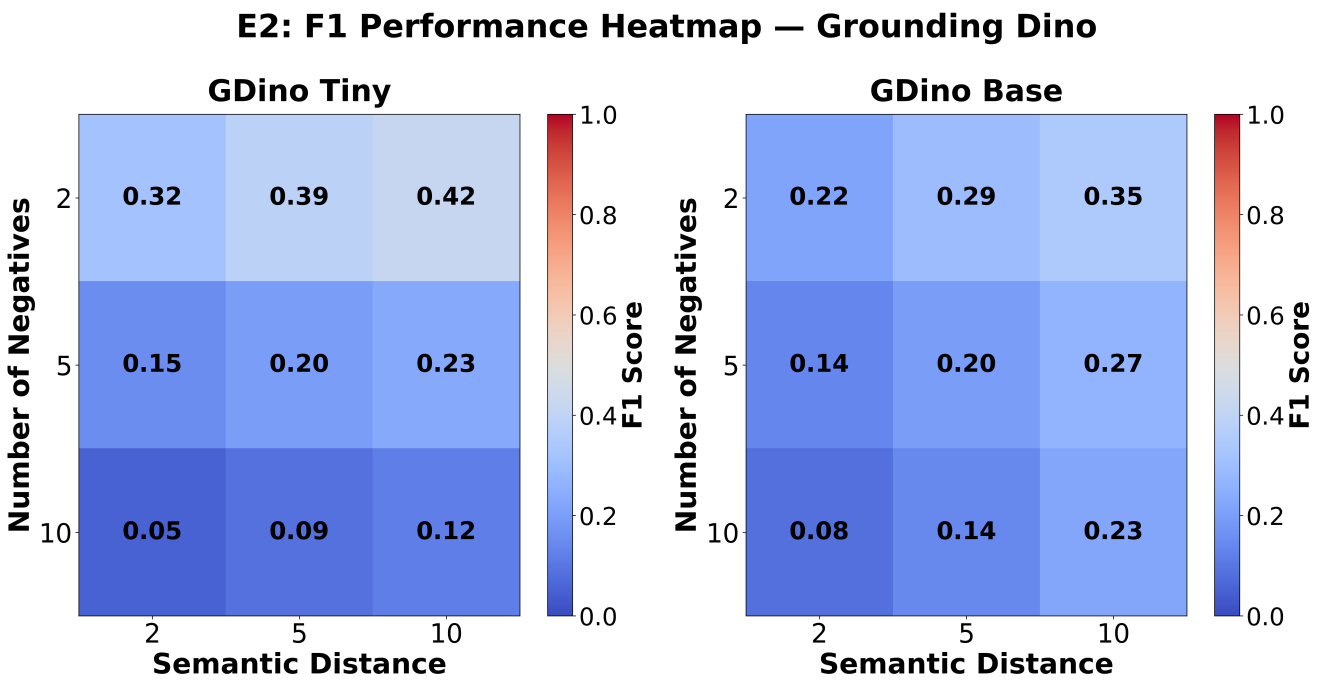

Figure 13. E2 F_1 heatmaps across the (D, N) grid for Grounding DINO Tiny and Base. Colour encodes F_1 score (blue = low, red = high). Both variants collapse toward near-zero at the hardest configuration (D=2, N=10); Grounding DINO Base maintains a slightly higher absolute F_1 across the grid, particularly at larger semantic distances (D=10).

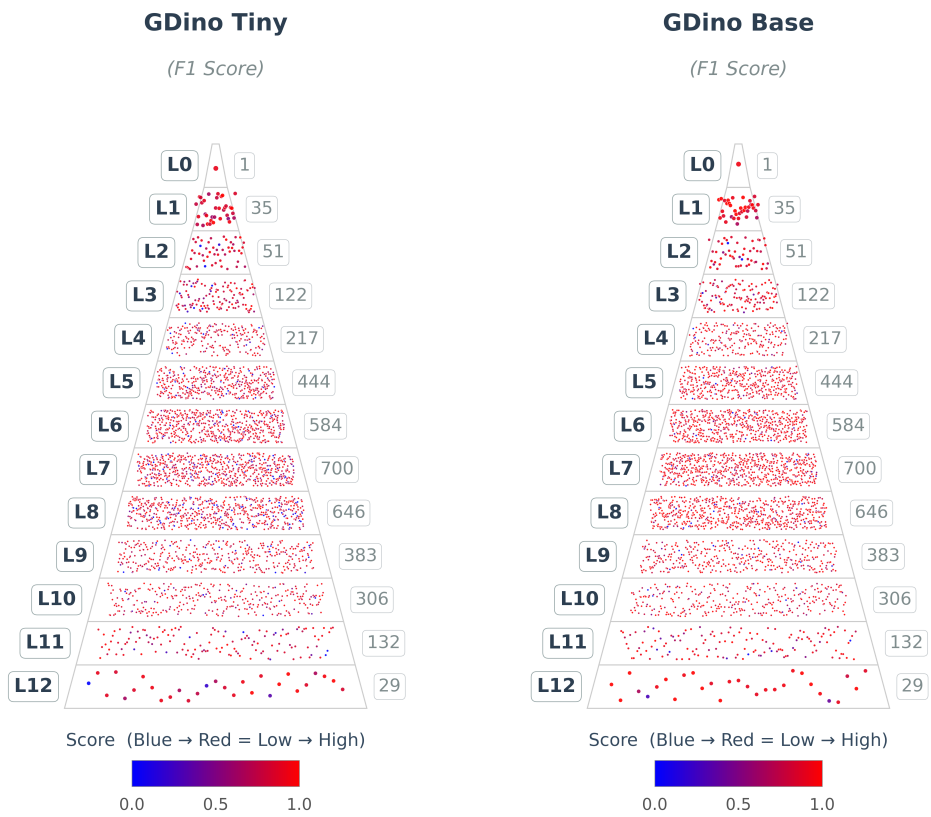

Figure 14. Per-node E1 F_1 distribution across WordNet taxonomy depth for Grounding DINO Tiny and Base. Both variants maintain a high density of red nodes across all depths, consistent with the near-zero rare/common gap reported in the main paper. Grounding DINO Base shows marginally broader red coverage at the deepest levels.

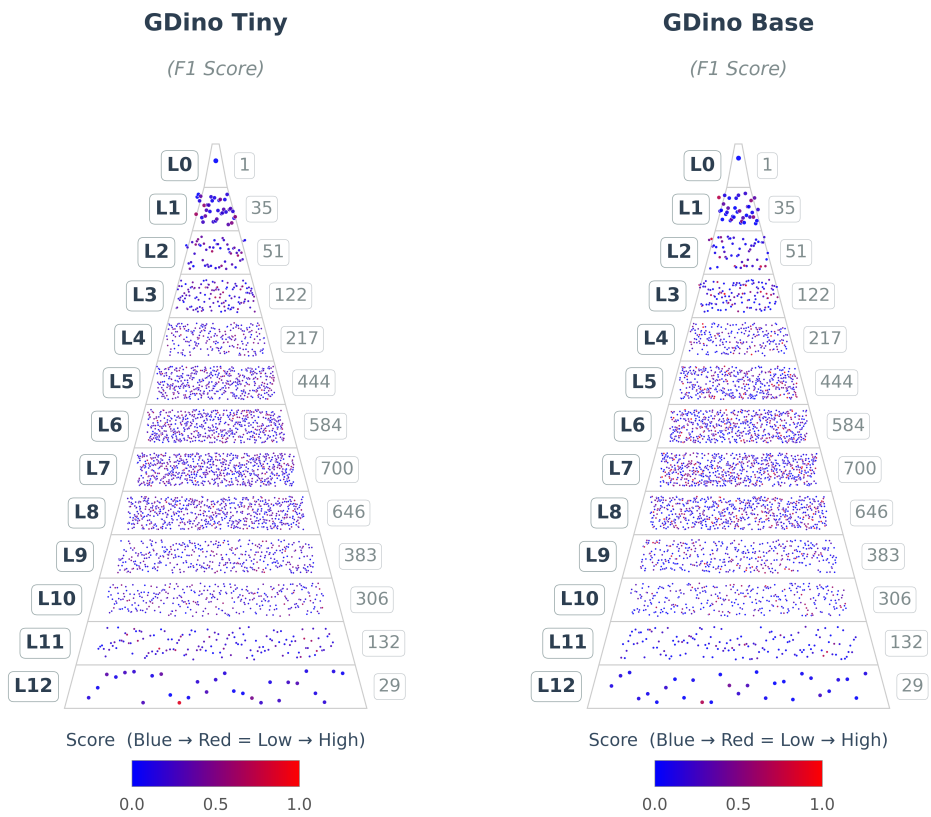

Figure 15. Per-node E2 F_1 distribution across WordNet taxonomy depth for Grounding DINO Tiny and Base. Both variants show a near-total collapse to blue nodes under E2 conditions, indicating that their strong E1 coverage does not translate to robustness against hard-negative distractors. This E2 collapse is consistent across all depth levels, rather than being concentrated at fine-grained leaves.

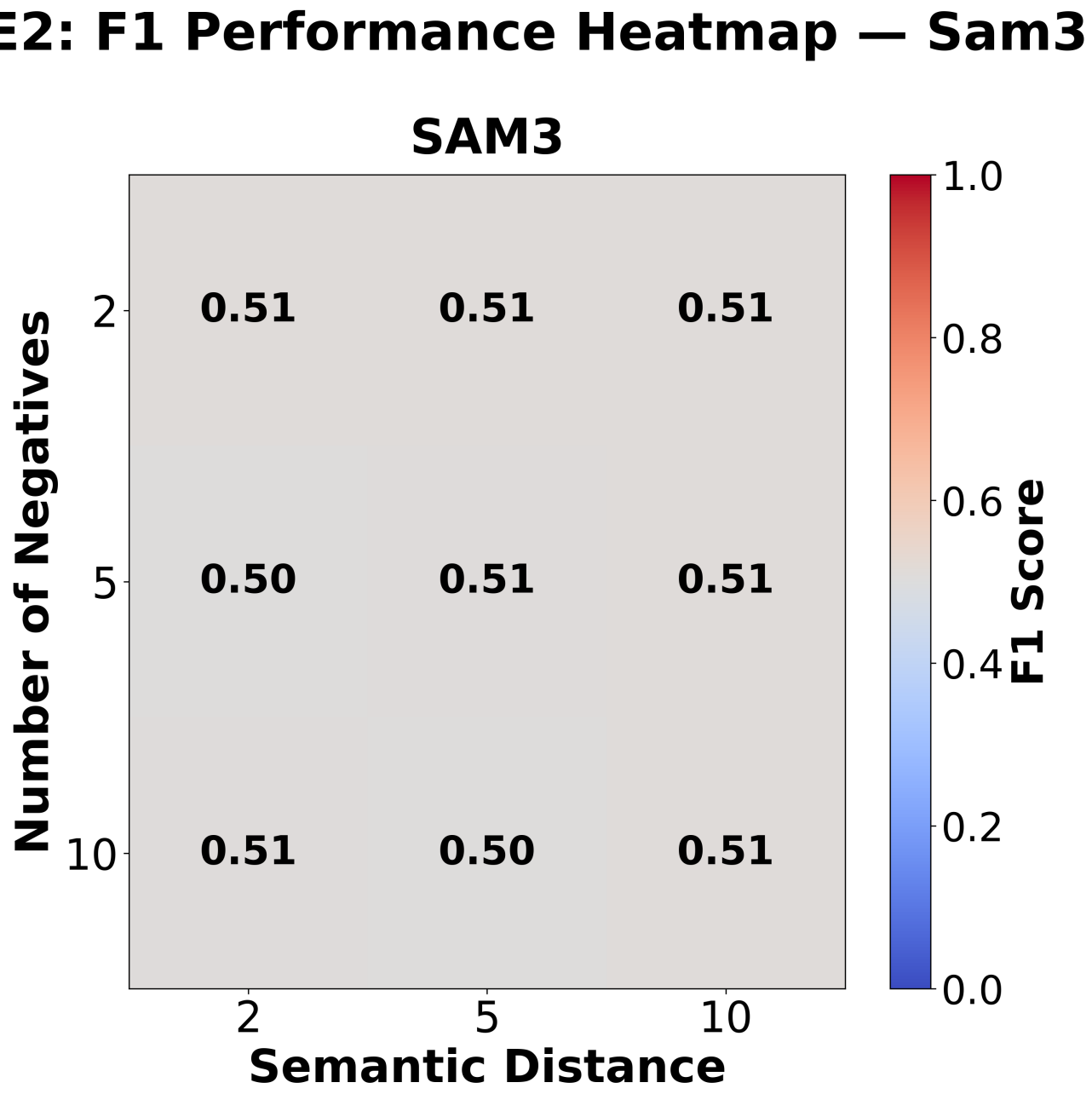

Figure 16. E2 F_1 heatmap across the (D, N) grid for SAM3. Colour encodes F_1 score (blue = low, red = high). The heatmap is near-uniform across all configurations, consistent with the prompt-insensitive behaviour discussed in Sec. 4 of the main paper: SAM3 assigns near-constant scores regardless of distractor proximity or count.

E1: Per-Node F1 Distribution Across Taxonomy Depth — Sam3

Each dot = one taxonomy node | Blue = low F1, Red = high F1

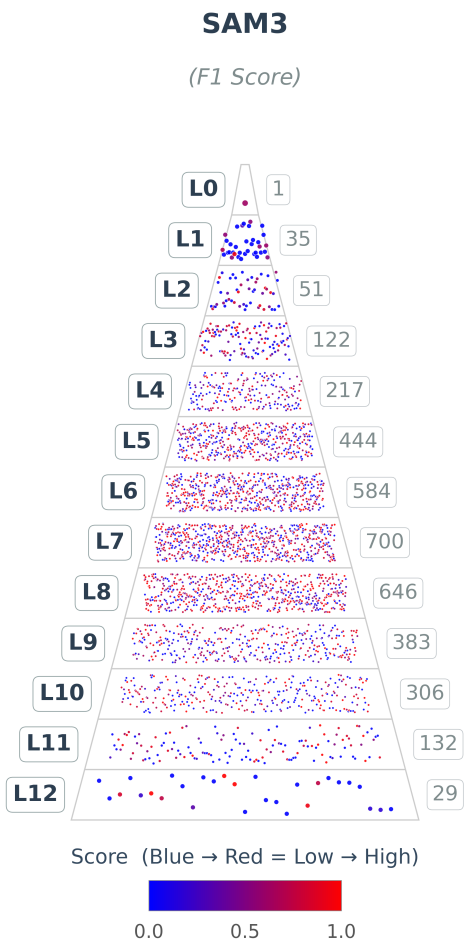

Figure 17. Per-node E1 F_1 distribution across WordNet taxonomy depth for SAM3. Blue nodes concentrate at shallow depths (L1–L2) with a more mixed distribution at mid-depths, reflecting SAM3's large rare/common performance gap and its limited sensitivity to category specificity.

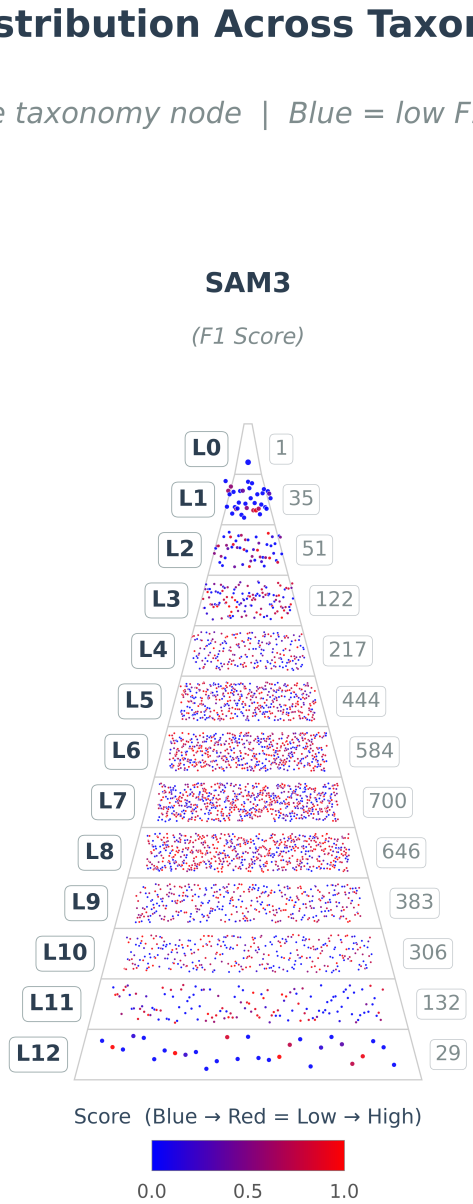

Figure 18. Per-node E2 F_1 distribution across WordNet taxonomy depth for SAM3. The distribution is virtually unchanged relative to E1 (Fig. 17), confirming that SAM3 is insensitive to the introduction of hard-negative distractors at any semantic distance or count.