# GHiML: VLM-Guided Hierarchical Multi-Label Learning with Applications to Benthic Imagery

**Anonymous authors**
Paper under double-blind review

## Abstract

Vision language models (VLMs) encode substantial expert knowledge within their weights, yet their direct application to specialised classification domains such as benthic imagery remains limited. Meanwhile, vision transformers trained on these tasks rely predominantly on low-level visual latents (colours, textures, and shapes) that lack the higher-level reasoning humans employ for identification. We propose Guided Hierarchical Multi-Label (GHiML) learning, a pipeline that extracts conceptual latent variables from a VLM conditioned on training data, generates sample-relevant captions grounded in those variables, and fuses the resulting textual embeddings with image embeddings via cross-attention and a per-node gated router. This approach provides semantic information complementary to low-level visual features, without requiring the VLM to be fine-tuned or a generative model to synthesise training images. We also introduce the inverted HML score ($d_{\text{inv}}$), a similarity score that evaluates hierarchical multi-label predictions through graph traversal distance and a cardinality penalty, providing an interpretable alternative to existing measures, which are insensitive to realistic degradations in sparse, deep hierarchies. Experiments on BenthicNet, FathomNet and iNaturalist show that the router improves macro $F_1$ over the vision-only baseline by up to 6.7 pp and $d_{\text{inv}}$ by up to 21.1 pp on in-domain benthic data, with gains persisting under distribution shift. Ablations locate the contribution of the text itself: on the multi-label biotic hierarchies, removing the per-sample text signal degrades $d_{\text{inv}}$ while leaving macro $F_1$ unchanged, consistent with text acting through prediction cardinality rather than node discrimination, whereas iNaturalist, whose annotations are single-label, loses 18.7% of macro $F_1$. Our results support the hypothesis that VLMs contain exploitable specialised knowledge that, when appropriately elicited, can augment classification in domains where direct visual learning is insufficient.

## 1 Introduction

For general tasks, vision language models (VLMs) have seen widespread adoption (Radford et al., 2021; Alayrac et al., 2022; Yang et al., 2025). However, recent research suggests that out-of-the-box VLMs remain limited in their application to specialised domains such as ocean and medical sciences (Wong et al., 2025; Nath et al., 2025). Consequently, in conjunction with more traditional convolutional neural networks (CNNs) (Fukushima, 1980; He et al., 2016), vision transformers (ViTs) (Dosovitskiy et al., 2021) remain the standard for classification tasks in these research areas (Langenkämper et al., 2017; Battach et al., 2025; Lowe et al., 2025; Aburass et al., 2025). Vision neural networks tend to work through the extraction of low-level latent variables (latents), such as colours, textures, and shapes (Zeiler & Fergus, 2014; Geirhos et al., 2019; Raghu et al., 2021; Ribeiro et al., 2016), resembling the hierarchical visual cortex (Fukushima, 1980). However, these latents lack the generalisable reasoning characteristics that humans use to identify objects. The lack of reasoning in visual tasks within specialised domains therefore remains a research gap. This motivation is illustrated in Figure 1.

Our application domain, benthic (seafloor) imagery classification, supports ecological and economic monitoring of maritime environments under increasing anthropogenic pressure (Scovil et al., 2025), at a data volume that far exceeds the capacity of manual analysis.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

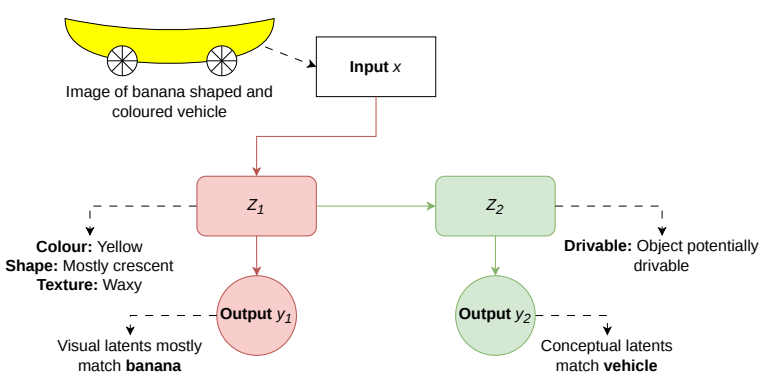



Figure 1: **Difference in Latents.** A hypothetical example. The red (left) branch depicts classification typical of contemporary vision-based neural networks, whose low-level latents learned from a limited dataset are insufficient to classify the object correctly. The green (right) branch additionally draws on general conceptual knowledge that cannot be learned from visual characteristics alone.

Benthic image classification is often a hierarchical multi-label (HML) task (Katija et al., 2022; Lowe et al., 2025), meaning a single image may contain multiple organism subjects or substrate types, each situated within a taxonomic or visually derived hierarchy that spans broad morphological categories down to fine-grained species-level distinctions. This combination of multi-label outputs and structured hierarchical dependencies poses challenges that extend beyond standard image classification (Giunchiglia & Lukasiewicz, 2020; Wehrmann et al., 2018). The problem is further compounded by substantial visual variability in seafloor imagery, arising from inconsistent lighting conditions, colour casts, and diverse substrate appearances, all of which undermine the robustness of learned representations (Lowe et al., 2025; Katija et al., 2022; Battach et al., 2025; Chen et al., 2021; Langenkämper et al., 2017). Moreover, rare classes, often those of the greatest ecological interest, remain poorly understood and difficult to detect by existing HML methods, as limited training examples prevent models from confidently learning discriminative features (Xu et al., 2024). Collectively, these factors make benthic image classification a particularly demanding domain in which current approaches struggle to achieve the robustness and reliability that ecological monitoring demands.

In an effort to improve robustness in image detection for this domain, we propose **G**uided **Hi**erarchical **M**ulti-**L**abel (GHiML) learning, an approach that seeks to unite the general knowledge contained in large VLMs with the specialised training of ViTs. GHiML consists of four steps: (1) identifying high-level latents relevant to reasoning about the domain, data points, and their annotations, (2) generating sample-relevant captions for extracted latents, (3) cross-attention fusion between image embeddings and generated textual embeddings for downstream classification, and (4) training a gated mixture of all three modalities across classes (hierarchical nodes). Our hypothesis is that the relationships between tokens encoded in a sufficiently trained VLM contain useful oceanographic connections beyond what low-level visual latents represent. Conditioning on those latents then helps to elicit this stored knowledge as a sample-specific textual signal.

Additionally, we propose an evaluation score for HML predictions, based on graph traversal distance and comparisons of multi-label cardinality (the degree of "multi-labelled-ness") (Tsoumakas & Katakis, 2009) between prediction and annotation. We refer to it as the inverted HML score ($d_{\text{inv}}$), a bounded similarity score ranging from 0 to 1, where higher values indicate better predictions. In practice, $d_{\text{inv}}$ is reported alongside the node-wise (macro-averaged) $F_1$ score (Xu et al., 2024; Lin et al., 2025). The two evaluate different objects, with $F_1$ scoring per-node detection and $d_{\text{inv}}$ scoring the predicted label set as a whole, and we do not treat either as a substitute for the other.

Our main contributions are therefore:

1. A use-inspired pipeline that elicits domain knowledge from a VLM as a sample-specific textual signal and integrates it with an existing HML framework, rather than a new HML method in itself;

2. $d_{\text{inv}}$, an evaluation score for HML predictions that accounts for both hierarchical proximity and prediction cardinality; and
3. An analysis of what the elicited text contributes, which we find acts as a coarse prior over regions of the hierarchy rather than as a discriminative signal within them.

## 2 BACKGROUND

### 2.1 HIERARCHICAL MULTI-LABEL LEARNING

In HML, annotations may be envisioned as a path starting from a root and terminating at a node. Multi-label implies that these annotations can consist of any combination of viable paths. The viability of an annotation is determined by the hierarchical constraint, that is $p_{\text{parent}} \geq p_{\text{child}}$. To enforce this condition during training, we apply Coherent Hierarchical Multi-Label Classification Neural Networks (C-HMCNN) (Giunchiglia & Lukasiewicz, 2020), which has established itself as the predominant framework for constrained HML classification and remains a robust standard (Xu et al., 2026; Liu et al., 2023). C-HMCNN converts the hierarchy of $|\mathcal{V}|$ nodes ($\mathcal{V}$ being the set of nodes) into an adjacency matrix ($\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$) that encodes ancestor-descendant relationships, then applies a maximum operation over each node's sub-hierarchy to ensure no child exceeds its parent's predicted probability. This constraint mechanism ($f_{\text{CM}}$) is insufficient alone, and a modified binary cross-entropy (BCE) loss is required to ensure gradients point in the correct direction:

$$\mathcal{L}_{\text{MC}} = \text{BCE}\Bigg(\underbrace{\overbrace{(1-\mathbf{Y})\, f_{\text{CM}}\big(\theta(\mathbf{X}),\, \mathbf{A}\big)}^{\text{Negative-node contribution}} + \overbrace{f_{\text{CM}}\big(\mathbf{Y}\theta(\mathbf{X}),\, \mathbf{A}\big)}^{\text{Positive-node contribution}}}_{\text{Full constrained output}},\ \mathbf{Y}\Bigg); \tag{1}$$

where $\mathcal{L}_{\text{MC}}$ is referred to as the *max constraint loss* (MCLoss) (Giunchiglia & Lukasiewicz, 2020; Xu et al., 2026), $\mathbf{X} \in \mathbb{R}^{n_{\text{b}}\times d}$ is the input of batch size $n_{\text{b}}$ and feature size $d$, $\theta$ is the model which maps from $d$ to $|\mathcal{V}|$, and $\mathbf{Y} \in \mathbb{R}^{n_{\text{b}}\times|\mathcal{V}|}$ is the annotation in binary representation. Finally, the $\mathbf{A}$ in Equation 1 has been expanded to be in $\mathbb{R}^{n_{\text{b}}\times|\mathcal{V}|\times|\mathcal{V}|}$. Most notably, the positive-node contribution must be filtered by the positive annotations prior to constraint to remove the effect of negatively-annotated descendants on their ancestors.

## 3 METHODS

Our approach is generalised to HML annotations which are not obligated to reach a leaf node (partial-depth). Technical details required for reproduction are provided in Appendix A.

### 3.1 EXTRACTING USEFUL LATENTS

The training partition of an image dataset is supplied to a VLM with an appropriate prompt, as shown in the first part of Figure 2. Formally, using terms from Figure 1, we sample a set of conceptual latents $\mathbf{Z}_2 \sim p(\mathbf{Z}_2 \mid \mathbf{X}, \mathbf{Y};\, \boldsymbol{\phi})$, where the model $\boldsymbol{\phi}$ must first leverage low-level visual latents $\mathbf{Z}_1$. The variables in $\mathbf{Z}_2$ should not resemble those we might expect from $\mathbf{Z}_1$. The goal is not to generate a description of the image and introduce redundant information for a visual encoder. Instead, the usefulness of $\mathbf{Z}_2$ is predicated on its ability to bias the model towards conceptual knowledge acquired through pretraining on large-scale language data.

With a sufficiently large and diverse dataset, this procedure can return thousands of candidate variables, many of which are redundant or weakly informative. We therefore refine the proposed set in four steps, specified in full in Appendix E. (1) We compute the empirical distribution $p(v)$ over nodes from the training annotations. (2) Since samples may be multi-labelled, an EM procedure resolves the ambiguous association between each proposed variable $z_{2,j}$ and individual nodes, yielding a distribution $q_j(v)$ over nodes. (3) Each $q_j$ is smoothed towards the prior, and candidates are scored by $D_{\text{KL}}(\tilde{q}_j \,\|\, p)$; variables with high divergence concentrate their annotation mass on rarer nodes and are hypothesised to carry the most discriminative information beyond the prior, and we retain only those scoring $\geq 1$. (4) After deduplication, a uniform per-node budget $B$ selects the top-$B$ candidates by KL score for each leaf node, as providing smaller VLMs with a large quantity of keywords can

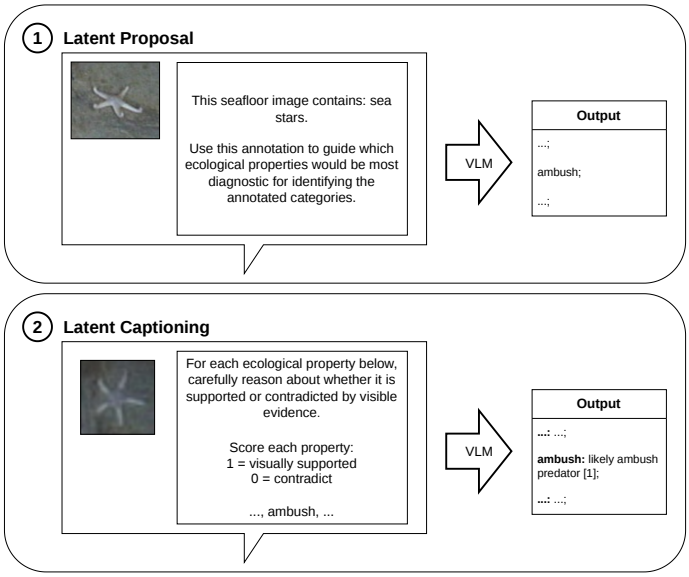



Figure 2: **VLM Variable Proposal and Captioning.** A sea star image and its annotation are used to propose the variable "ambush"; during captioning, the VLM is prompted to reason about each variable for each sample.

introduce anchoring bias (Shi et al., 2023; Wei et al., 2023). The final variable set typically comprises 15-20 variables in our runs.

### 3.2 Latent Captioning

We refer again to Figure 2. This step consists of supplying the processed set of variables to the VLM for captioning. In addition to captioning for each variable, the model is prompted to assign a binary score for the relevance of the variable to each sample. This is done for analysis purposes, and is inspired by Liu et al. (2024). A quality check is performed to ensure that all variables are accounted for in the captions of each sample.

### 3.3 Models and Architecture

We refer to Figure 3 for an overview of the models and training pipeline we use. Images are embedded with a DINOv3 ViT-H backbone (Siméoni et al., 2025), Qwen3-VL-8B-Instruct is used for variable proposal and captioning, and Qwen3-Embedding-8B for embedding captions (Yang et al., 2025). These models were chosen for their open-source availability, reliability, and ability to fit on consumer-grade GPUs accessible to domain-specialised research labs.

To summarise our models, we train self-attention-based pooling mechanisms to transform sequential image and text embeddings into dense vectors. These unimodal models serve as baselines and reveal the overlap between textual and visual performance, specifically the degree to which each modality carries redundant or complementary information. A third cross-attention mechanism is trained with image as the query modality and text as the key-value modality[1]. Finally, we train a router (Jacobs et al., 1991) that takes as input the concatenated pooled embeddings and outputs softmax weights over the three modalities for each node, so that the contribution of each modality may vary by node and by sample. We run two versions of this router: one that loads frozen pretrained pooling layers from our unimodal runs and trains the classifiers and router mechanism (Router), and one with all parameters trained from scratch (Router$^{\dagger}$).

[1] In preliminary testing, image querying text performed stronger than the reverse.

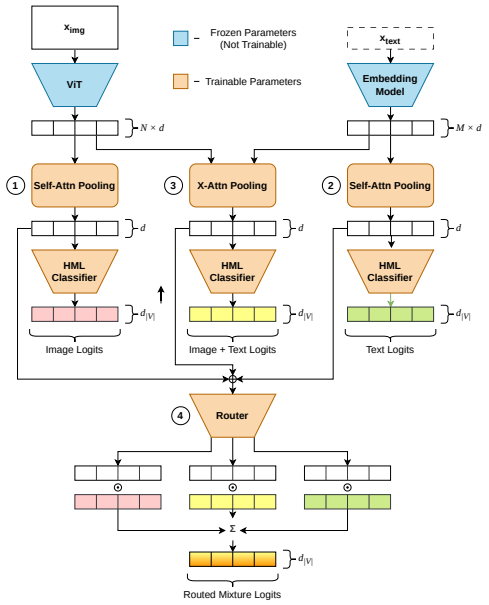



Figure 3: **Model Architectures.** The (1) image and (2) text streams apply self-attention pooling followed by an HML-constrained classifier; (3) applies cross-attention followed by pooling; (4) a router predicts per-node weights that mix the logits of (1)–(3).

Every row of Table 1 uses the same hierarchical classification head, an implementation of C-HMCNN (Giunchiglia & Lukasiewicz, 2020) with the max-based constraint $f_{\text{CM}}$ and its accompanying loss, and the same per-node thresholding procedure (subsection 3.5). The rows therefore differ only in which modalities are available and how they are combined, so C-HMCNN is held constant as a control rather than compared against.

### 3.4 Inverted HML Score $d_{\text{INV}}$

Evaluating HML predictions requires a score that captures both *closeness* between prediction and ground truth and *multi-label cardinality*, penalising predictions that are deficient or excessive, given that real-world annotations may be incomplete (Lowe et al., 2025; Katija et al., 2022). Existing scores do not fully satisfy these requirements. Node-wise $F_1$ penalises cardinality errors but forfeits hierarchical proximity, particularly at standard $0.5$ thresholding, where deeper nodes frequently score zero regardless of prediction closeness (Xu et al., 2026). Micro-averaged AP (Giunchiglia & Lukasiewicz, 2020; Wehrmann et al., 2018) pools across all nodes, so evaluation in sparse hierarchies lets easily ranked negatives dominate, producing overly optimistic scores. A comparison of these scores is presented in Appendix I.

We address these gaps with $d_{\text{inv}}$, which directly exploits hierarchical distance and explicitly accounts for multi-label cardinality (with user-stipulated flexibility). Because it maps graph traversal distance onto a normalised $[0, 1]$ scale, the score is directly interpretable: a value of one indicates a perfect prediction, while intermediate values communicate precisely how far the prediction sits in terms of the hierarchy's own structure, and how far its cardinality departs from the annotation. We define:

$$d_{\text{inv}} = s_{\max}\,\gamma, \tag{2}$$

where $s_{\max}$ encapsulates both normalisation and "identification" (determining which multi-label component in the prediction corresponds to which component in the ground truth). Both $\hat{\mathcal{Y}}$ and $\mathcal{Y}$ are ancestor-closed by construction, and $d_{\text{inv}}$ is computed over their *endpoints*: the active nodes carrying no further active descendants. Ancestors are therefore not scored twice, and a prediction is evaluated by where its most specific claims fall. For normalisation, we map the prediction distance $d^{(ij)}_{\text{pred}}$ between ground-truth endpoint $i$ and predicted endpoint $j$ onto a $[0, 1]$ scale via $[\mathbf{D}_{\text{inv}}]_{ij} = (d^{(i)}_{\max} - d^{(ij)}_{\text{pred}})/d^{(i)}_{\max}$, where $d^{(i)}_{\max}$ is the eccentricity of ground-truth endpoint $i$, i.e. the greatest tree distance from it to any other node in the hierarchy[2]. For identification, we construct a

[2]Under unit edge weights.

matrix $\mathbf{D}_{\text{inv}} \in \mathbb{R}^{n_{\text{truth}} \times n_{\text{pred}}}$ of pairwise similarity scores for all component combinations, and solve the resulting rectangular assignment problem using the Jonker–Volgenant algorithm (Crouse, 2016). Writing $f_{\text{JV}}(\hat{\mathcal{Y}}, \mathcal{Y}; \mathbf{D}_{\text{inv}})$ for the total similarity of the optimal assignment, we normalise by the number of ground-truth components to obtain

$$s_{\max} = \frac{1}{n_{\text{truth}}} f_{\text{JV}}(\hat{\mathcal{Y}}, \mathcal{Y}; \mathbf{D}_{\text{inv}}), \tag{3}$$

the optimistically matched combination score between prediction ($\hat{\mathcal{Y}}$) and ground truth ($\mathcal{Y}$). The cardinality factor $\gamma$ penalises excessive predictions, an important consideration in HML settings. Defining the number of excessive predictions as $n_{\text{E}} = n_{\text{pred}} - n_{\text{truth}}$ and the maximum number of excessive predictions as $n_{\text{ME}} = |\mathcal{V}| - n_{\text{truth}}$, we model $\gamma$ with exponential decay:

$$\gamma(n_{\text{E}};\, k_{\text{P}},\, n_{\text{ME}}) = \begin{cases} a\, e^{-b\, n_{\text{E}}} + c & \text{if } k_{\text{P}} \neq k_{\text{P}}^*, \\ 1 - \dfrac{n_{\text{E}}}{n_{\text{ME}}} & \text{otherwise,} \end{cases} \tag{4}$$

where $k_{\text{P}} \in (0, 1)$ governs the penalty incurred by a single excess prediction, $k_{\text{P}}^* = (n_{\text{ME}} - 1)/n_{\text{ME}}$ is the critical value at which the decay reduces to the linear case, and the parameters $a = (1 - e^{-b n_{\text{ME}}})^{-1}$ and $c = 1 - a$ are determined by boundary conditions $\gamma(0) = 1$ and $\gamma(n_{\text{ME}}) = 0$, with $b$ obtained as the unique non-trivial root of $(k_{\text{P}} - 1)e^{-b n_{\text{ME}}} + e^{-b} - k_{\text{P}} = 0$ (formal proofs of existence and uniqueness are provided in Appendix J). Edge cases (disjoint roots, empty and deficient predictions) are defined in Appendix J.

As $d_{\text{inv}}$ is defined per sample, we report it macro-averaged across nodes, as with $F_1$: each node receives the mean $d_{\text{inv}}$ of the test samples annotated with it, and these node scores are averaged with equal weight. We set $k_{\text{P}} = 0.5$, as a significant contemporary challenge in the benthic domain is the possibility of incomplete annotations (Barnard & Chrobak, 2026). Consequently, the annotation cardinalities cannot be fully trusted.

### 3.5 Optimal $F_1$ Thresholding

Deeper and rare nodes are poorly detected in HML (Pereira et al., 2021; Lipton et al., 2014), possibly in part due to poor calibration (Xu et al., 2024). Following Lipton et al. (2014), we grid-search per-node thresholds maximising $F_1$ ($0.01$ to $0.99$ in $0.01$ increments, starting from $0.50$), subject to the hierarchical constraint $t_{\text{parent}} \leq \min_{v \in \mathcal{V}_{\text{d}}} t_v$, where $\mathcal{V}_{\text{d}}$ denotes the set of descendants. Thresholds are fitted on training predictions only, and the procedure is universally applied to every model reported.

## 4 Experiments

### 4.1 Datasets

We evaluate on three datasets, described in full in Appendix C. BenthicNet (Lowe et al., 2025) provides 188,679 annotated images labelled under the CATAMI scheme (Althaus et al., 2015), from which we form two overlapping tasks, substrate (BN-Sub, 23 nodes) and biota (BN-Bio, 293 nodes). FathomNet (2025) (FN) (FathomNet 2025) is a benthic biota competition dataset of 9,305 full images, for which we extracted a 192-node taxonomic hierarchy from WoRMS (WoRMS Editorial Board, 2026), making the problem HML (full-depth). For both, the innate test partitions serve as out-of-domain (OOD) test sets, and we additionally construct in-domain (ID) test sets from a random 20% of the data. These partitions are "OOD" because BenthicNet's test set is geospatially partitioned and FathomNet's competition test set was hand-selected. Lastly, iNaturalist (iNat) (Van Horn et al., 2018) spans terrestrial fauna, flora, and fungi with single-label annotations over a hierarchy of over 16,000 nodes; we train on the 500,000 training samples of iNat (mini) and test on its 100,000 validation samples, replacing MCLoss with a softmax over the 10,000 endpoints, marginalised up the hierarchy. Since each prediction and annotation consists of exactly one endpoint, $n_{\text{E}} = 0$ and $\gamma \equiv 1$, so $d_{\text{inv}}$ reduces to normalised traversal similarity on this dataset.

### 4.2 Results and Discussion

As shown in Table 1, the router models achieve the strongest performance in nearly all settings. The ranking across methods is consistent for each metric across all random seeds. Although text

Table 1: **HML Results.** Macro-average performance across datasets, modalities, and distribution settings. Values are mean ± std (%). Superscript indicates mean rank across seeds. Ten seeds were tested for BN datasets, and thirty for FN. Coloured values differ significantly from Img alone (Wilcoxon signed-rank test; green: $p < 0.05$; **bold dark green**: $p < 0.01$).

| | | BN-Sub | | BN-Bio | | FN | |
|---|---|---|---|---|---|---|---|
| | | **F1** | $d_{\text{inv}}$ | **F1** | $d_{\text{inv}}$ | **F1** | $d_{\text{inv}}$ |
| **ID** | Txt | $48.5_{\pm 2.9}^{(5.0)}$ | $40.5_{\pm 5.3}^{(5.0)}$ | $7.8_{\pm 1.6}^{(5.0)}$ | $1.2_{\pm 1.7}^{(5.0)}$ | $15.6_{\pm 2.9}^{(5.0)}$ | $0.3_{\pm 0.2}^{(5.0)}$ |
| | Img | $77.9_{\pm 3.5}^{(3.4)}$ | $68.1_{\pm 4.5}^{(3.5)}$ | $14.3_{\pm 3.8}^{(3.2)}$ | $5.1_{\pm 5.5}^{(3.5)}$ | $44.5_{\pm 5.1}^{(3.5)}$ | $16.5_{\pm 3.2}^{(3.5)}$ |
| | Img + Txt | $78.3_{\pm 3.7}^{(3.0)}$ | $68.7_{\pm 4.3}^{(3.0)}$ | $14.0_{\pm 3.5}^{(3.8)}$ | $4.9_{\pm 5.6}^{(3.5)}$ | $44.5_{\pm 5.2}^{(3.5)}$ | $16.7_{\pm 3.3}^{(3.4)}$ |
| | Router | $\mathbf{81.4_{\pm 2.5}^{(1.1)}}$ | $\mathbf{72.7_{\pm 3.0}^{(1.1)}}$ | $\mathbf{16.6_{\pm 3.5}^{(2.0)}}$ | $\mathbf{7.6_{\pm 6.5}^{(1.5)}}$ | $\mathbf{51.2_{\pm 4.2}^{(1.1)}}$ | $\mathbf{37.6_{\pm 4.1}^{(1.0)}}$ |
| | Router† | $79.4_{\pm 2.3}^{(2.5)}$ | $69.4_{\pm 3.9}^{(2.4)}$ | $\mathbf{17.8_{\pm 4.0}^{(1.0)}}$ | $\mathbf{7.7_{\pm 6.9}^{(1.5)}}$ | $\mathbf{49.7_{\pm 4.6}^{(1.9)}}$ | $\mathbf{23.0_{\pm 4.6}^{(2.0)}}$ |
| **OOD** | Txt | $38.6_{\pm 3.2}^{(5.0)}$ | $36.6_{\pm 5.7}^{(5.0)}$ | $7.4_{\pm 1.5}^{(5.0)}$ | $2.5_{\pm 3.4}^{(4.8)}$ | $12.4_{\pm 4.8}^{(5.0)}$ | $0.4_{\pm 0.3}^{(5.0)}$ |
| | Img | $64.8_{\pm 4.5}^{(2.5)}$ | $60.2_{\pm 5.3}^{(3.2)}$ | $11.4_{\pm 2.7}^{(3.4)}$ | $4.4_{\pm 5.1}^{(3.8)}$ | $37.4_{\pm 10.0}^{(3.4)}$ | $18.3_{\pm 4.8}^{(3.5)}$ |
| | Img + Txt | $64.2_{\pm 4.6}^{(3.1)}$ | $60.4_{\pm 5.1}^{(2.7)}$ | $11.3_{\pm 2.8}^{(3.6)}$ | $4.7_{\pm 5.8}^{(3.0)}$ | $37.4_{\pm 10.0}^{(3.4)}$ | $18.4_{\pm 5.1}^{(3.3)}$ |
| | Router | $\mathbf{67.7_{\pm 3.1}^{(1.1)}}$ | $\mathbf{64.5_{\pm 3.6}^{(1.0)}}$ | $\mathbf{13.2_{\pm 3.0}^{(1.9)}}$ | $\mathbf{5.9_{\pm 5.8}^{(2.0)}}$ | $\mathbf{40.9_{\pm 9.2}^{(1.7)}}$ | $\mathbf{32.5_{\pm 6.4}^{(1.0)}}$ |
| | Router† | $64.0_{\pm 4.8}^{(3.3)}$ | $60.4_{\pm 5.3}^{(3.1)}$ | $\mathbf{14.3_{\pm 3.7}^{(1.1)}}$ | $\mathbf{6.4_{\pm 6.4}^{(1.4)}}$ | $\mathbf{41.7_{\pm 9.2}^{(1.5)}}$ | $\mathbf{23.1_{\pm 6.5}^{(2.2)}}$ |

Table 2: **iNaturalist Results.** Macro-average performance over the 100,000 held-out images. Values are mean ± std (%) across five seeds; superscript indicates mean rank across seeds. Coloured values are significantly higher than Img (one-sided Wilcoxon signed-rank test; green: $p < 0.05$).

| | **F1** | $d_{\text{inv}}$ |
|---|---|---|
| Txt | $0.9_{\pm 0.1}^{(5.0)}$ | $28.4_{\pm 2.9}^{(5.0)}$ |
| Img | $22.8_{\pm 4.2}^{(4.0)}$ | $56.3_{\pm 2.7}^{(4.0)}$ |
| Img + Txt | $29.8_{\pm 6.0}^{(3.0)}$ | $61.2_{\pm 4.2}^{(3.0)}$ |
| Router | $55.6_{\pm 2.1}^{(2.0)}$ | $77.9_{\pm 1.7}^{(1.0)}$ |
| Router† | $59.0_{\pm 2.9}^{(1.0)}$ | $75.9_{\pm 2.6}^{(2.0)}$ |

underperforms vision on $F_1$ at nearly every node, the two metrics disagree in a specific way: at 75 of the 171 BN-Bio nodes with sufficient support, text scores higher on $d_{\text{inv}}$ while failing to identify the node on $F_1$, against 6 nodes where the reverse holds. This indicates that text contributes generalisable semantic information about where predictions belong in the hierarchy, rather than finer discrimination between nodes.

One contributing factor is that several of the selected variables encode functional traits, such as mobility, attachment strategy, and predation mode, that are particularly informative for sessile organisms. Algae and sponges, for instance, share many such traits, and the resulting captions place them in close proximity within the hierarchy even when exact classification fails.

Table 2 reports the same comparison on iNat, where the hierarchy is roughly 55 times larger and annotations are single-label. The ordering is unchanged: both routers exceed the image branch on both metrics at every seed, and text alone remains the weakest by a wide margin on $F_1$ while retaining half of the image branch's $d_{\text{inv}}$. Because $d_{\text{inv}}$ has no cardinality term on iNat, this is a direct measure of proximity. A text stream with poor $F_1$ on its own still places predictions within reach of the target.

We next ask whether text and vision contribute distinct information, and whether the router successfully exploits both. Among the top-performing nodes for text, there is low overlap between the two modalities, and the router frequently surpasses vision while retaining a margin of advantage comparable to that of text alone (Appendix H). There, we also study the relevant variables and their correlations with text advantage, as well as with individual nodes. We observe that the proposed variables provide strong contextual information that benefits classification, even when they do not directly improve prediction of the original nodes with which they were associated via KL divergence.

This last point warrants emphasis. Although variables were assigned to modal nodes, the captions they generate encode a far richer set of combinations. The variable `ambush`, for example, was initially selected for its association with sea stars. In practice, it also describes rocky substrates

Table 3: **Mean Replacement.** Relative change in scores across five seeds when each sample's text embedding is replaced by the mean. Negative values indicate degradation. Bolded indicate $p \leq 0.05$ on paired one-sided Wilcoxon signed-rank test.

| dataset | annotation | $\Delta F_1$ | $\Delta d_{\text{inv}}$ |
|---|---|---|---|
| BN-Sub (23 nodes) | multi-label | $+1.5\%$ | $-0.1\%$ |
| BN-Bio (293 nodes) | multi-label | $+0.6\%$ | $\mathbf{-21.7\%}$ |
| iNaturalist (10,000 endpoints) | single-label | $\mathbf{-18.7\%}$ | $\mathbf{-10.2\%}$ |

(habitats in which ambush predators tend to occur), detects lobsters (themselves ambush predators), and identifies organisms or environments facilitating camouflage consistent with ambush strategy. A single functional concept thus provides signal across multiple nodes. This is consistent with considering ecological function and context rather than relying solely on low-level visual latents such as colour and texture.

### 4.3 Text as a Sub-Tree Prior

Beyond the trained branches, we also examine the embeddings themselves. We fit two $k$-NN classifiers over the frozen image and text embeddings, mean-pooled across the sequence dimension, and evaluate them on the same samples by cross-validation (Appendix B). No trained component of our pipeline is used, so any complementarity observed is a property of the representations rather than of our architecture. Were the text embedding correct only on samples that the image embedding already correctly classifies, it would be strictly redundant and the independence ratio $P(\text{text correct} \mid \text{image wrong}) \,/\, P(\text{text correct})$ would be zero. Across the BN and iNat datasets, $3.5$–$10.8\%$ of samples across hierarchy levels are classified correctly by the text embedding but not by the image embedding. A more detailed $k$-NN comparison against generic captioning (image description) embeddings in BN-Bio (Appendix B) demonstrates that our extracted keywords help reduce redundancy with images precisely at the clade levels which benefit classification the most.

In Table 3, a mean embedding vector is supplied to the models in lieu of the original text embedding, such that no per-sample textual information survives. On BN-Bio the degradation occurs only for $d_{\text{inv}}$, while $F_1$ is unaffected. On iNat both metrics degrade. BN-Sub, the smallest hierarchy, is unaffected in either metric. The text signal therefore appears to matter more for larger and more complex hierarchies. The two biotic datasets differ in which metrics respond, and we attribute this to their label structure. Under multi-label annotations, the clade signal appears to suppress spurious subtree activations, which is expressed through prediction cardinality rather than hierarchical placement (subsection I.2). However, iNat's single-label annotations foreclose that channel. Replacing the text with scale-matched Gaussian noise degrades performance below the image-only baseline on three of the four datasets, indicating that the gains do not follow from the added dimensionality alone (subsection G.2).

We also report on zero-shot VLM classification, which produced unscorable results. A zero-shot VLM cannot produce scheme-conformant predictions for this task without conditioning, and even with scheme-adherent conditioning, it over-predicts substantially (subsection G.1). Additionally, we test a descriptor-matching baseline, which recovers roughly half the node $F_1$ of a trained image branch and under a tenth of its $d_{\text{inv}}$, with descriptors reused heavily between sibling nodes (subsection G.3). Many benthic images are homogeneous in colour, texture and tone despite differing content, so generic appearance-based descriptions tend to be interchangeable across nodes. Finally, we analyse the router's gate weights on BN-Bio and confirm that they remain open to text (subsection H.2).

## 5 Related Works

Our work is influenced by the causal discovery literature, most notably COAT (Liu et al., 2024). COAT employs LLMs to discover a phenomenon's Markov Blanket by proposing candidate variables and annotating each sample with a trinary encoding per variable (relevant, irrelevant, or inversely relevant). COAT demonstrates that LLMs can actively mediate between unstructured observations and structured representations. Our pipeline differs in that we target hierarchical multi-label classification,

select variables per hierarchy node by KL divergence, and have a VLM score each variable with a brief visual justification. The resulting captions are embedded as dense text and fused with the image representation through a learned per-node router, so the variables serve as contextual signal for a vision model and are never used as discrete features for causal analysis.

Closer in line with our motivation are diffusion-based augmentation strategies (Yerramilli et al., 2025; Li et al., 2024; Trabucco et al., 2024; Samuel et al., 2024), which leverage the visual priors of large-scale generative models to augment downstream classification. SGID (Li et al., 2024) uses BLIP-generated captions (Li et al., 2023) and CLIP-based selection (Radford et al., 2021) to guide Stable Diffusion toward class-consistent samples, DA-Fusion (Trabucco et al., 2024) semantically augments real images via Stable Diffusion, and SeedSelect (Samuel et al., 2024) optimises the initial noise seed to improve rare class generation. However, these approaches are largely inapplicable to benthic habitat classification, where many taxa remain underrepresented in the web-scale data on which diffusion models are trained (Zhang et al., 2025). Instead, much of our knowledge of benthic taxa and habitat is encoded textually, motivating our alternative strategy of extracting higher-level semantic information from a VLM rather than synthesising novel training samples.

Within the HML literature, research into improving detection and addressing class imbalance remains sparse, with existing methods broadly employing either resampling techniques (Pereira et al., 2021) or weighted objectives (Xu et al., 2026). Our work enhances node detection broadly by leveraging the general knowledge of an external foundation model.

Lastly, we compare $d_{\text{inv}}$ with the hierarchical $F_1$ of Lin et al. (2025) and with CoPHE (Falis et al., 2021) in subsection I.1. We find $hF_1$ comparatively lenient toward excessive prediction cardinality, since excess predictions that share correct ancestors inflate its ancestor-expanded sets only marginally, whereas CoPHE, which addresses this leniency, does not measure how far a prediction sits from its target and requires a well-defined depth for every node.

## 6 Conclusion

We have presented GHiML, a pipeline for extracting and leveraging higher-level semantic knowledge from VLMs to augment hierarchical multi-label classification in specialised visual domains. Through a structured variable proposal, EM-based filtering, and sample-conditioned captioning procedure, we demonstrated that a VLM can be positioned to generate complementary contextual information to standard visual features. The gated router architecture exploits this information on a per-node basis, consistently outperforming both unimodal baselines and the cross-attention fusion model across two benthic datasets under both in-domain and out-of-domain considerations, and on iNat.

We additionally introduced $d_{\text{inv}}$, a metric that grounds HML evaluation in the structure of the hierarchy itself. By combining graph traversal distance with an explicit cardinality penalty governed by a single interpretable parameter, $d_{\text{inv}}$ addresses gaps left by existing methods, and offers an intuitive understanding of model performance. Limitations of the work are discussed in greater detail in Appendix F.

### AI Use Statement

Vision language models are a core component of the method studied in this work, as described in section 3, and our work required exposing VLMs to image data. Separately, we used generative AI tools to implement code for auxiliary testing and for some of the experiments reported in the appendices; the main pipeline and the main experiments were not implemented with generative AI. We also used generative AI tools to provide critical feedback on our research methodology and experiments by searching for flaws in the work, which required exposing the models to results as well. Additionally, we used generative AI tools for coding assistance, including the implementation of common functions, and for editing and formatting the manuscript. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation, or clean or reformat datasets. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### Ethics Statement

Our research does not involve human subjects. The data we use are open source imagery (benthic and iNaturalist) with appropriate licenses, and do not have privacy concerns. We believe that our research field is beneficial to environmental preservation and necessary for making ecologically-informed decisions. We discuss the necessity of our work for ecological preservation in section 1. We do not believe the models we train, the research goals we target, or the data we use carries with them high risks of misuse. The majority of code we use was produced by us. Where applicable, we have referenced other repositories and the licenses they shared their code under. The data used are open source, as are any pretrained models we use.

### Reproducibility Statement

Our Methods section describes the conceptual overview of our approach (section 3). We also provide detailed technical information for reproducibility (Appendix A, Appendix D, and Appendix E). Additional details are specified in section 4 and Appendix A. An anonymised version of our pipeline is provided with supplementary materials. The datasets we use are publicly available with the minor exception of a privately retained part of FathomNet-2025's test annotations, which requires approval from the FathomNet team. Similarly, the pretrained models we use are open-source, although DINOv3 access requires approval from the Meta Research team. Our paper contains descriptions of the mathematical behaviour of an introduced metric. Proofs regarding this subject are in Appendix J.

## References

Sanad Aburass, Osama Dorgham, Jamil Al Shaqsi, Maha Abu Rumman, and Omar Al-Kadi. Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases. *Journal of imaging informatics in medicine*, 38(6):3928—3971, December 2025. ISSN 2948-2933. doi: 10.1007/s10278-025-01481-y. URL https://europepmc.org/articles/PMC12701147.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Franziska Althaus, Nicole Hill, Renata Ferrari, Luke Edwards, Rachel Przeslawski, Christine H. L. Schönberg, Rick Stuart-Smith, Neville Barrett, Graham Edgar, Jamie Colquhoun, Maggie Tran, Alan Jordan, Tony Rees, and Karen Gowlett-Holmes. A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: The CATAMI classification scheme. *PLOS ONE*, 10(10):1–18, 10 2015. doi: 10.1371/journal.pone.0141039. URL https://doi.org/10.1371/journal.pone.0141039.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Kevin Barnard and Laura Chrobak. FathomNetCLEF2026 @ LifeCLEF & CVPR-FGVC. https://kaggle.com/competitions/fathomnet-2026, 2026. Kaggle.

Yahia Battach, Abdulwahab Felemban, Faizan Farooq Khan, Yousef A. Radwan, Xiang Li, Fabio Marchese, Sara Beery, Burton H. Jones, Francesca Benzoni, and Mohamed Elhoseiny. ReefNet: A large scale, taxonomically enriched dataset and benchmark for hard coral classification. *arXiv*, 2025. URL https://arxiv.org/abs/2510.16822.

Qimin Chen, Oscar Beijbom, Stephen Chan, Jess Bouwmeester, and David Kriegman. A new deep learning engine for coralnet. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3686–3695, 10 2021. doi: 10.1109/ICCVW54120.2021.00412.

David F. Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. doi: 10.1109/TAES.2016.140952.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

eor123, Kevin Barnard, Laura Chrobak, and picekl. FathomNet 2025 @ CVPR-FGVC. https://kaggle.com/competitions/fathomnet-2025, 2025. Kaggle.

Matús Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. Cophe: A count-preserving hierarchical evaluation metric in large-scale multi-label text classification. *CoRR*, abs/2109.04853, 2021. URL https://arxiv.org/abs/2109.04853.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 4 1980. ISSN 1432-0770. doi: 10.1007/BF00344251. URL https://doi.org/10.1007/BF00344251.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9662–9673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6dd4e10e3296fa63738371ec0d5df818-Paper.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL http://arxiv.org/abs/1606.08415.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, and Katherine L. C. Bell. FathomNet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports*, 12(1):15914, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-19939-2. URL https://doi.org/10.1038/s41598-022-19939-2.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c32319f4868da7613d78af9993100e42-Paper-Conference.pdf.

Daniel Langenkämper, Martin Zurowietz, Timm Schoening, and Tim W. Nattkemper. Biigle 2.0 - browsing and annotating large marine image collections. *Frontiers in Marine Science*, 4, 2017. ISSN 2296-7745. doi: 10.3389/fmars.2017.00083. URL https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2017.00083.

Bohan Li, Xiao Xu, Xinghao Wang, Yutai Hou, Yunlong Feng, Feng Wang, Xuanliang Zhang, Qingfu Zhu, and Wanxiang Che. Semantic-guided generative image augmentation method with diffusion models for image classification. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i4.28084. URL https://doi.org/10.1609/aaai.v38i4.28084.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Sanne Lin, Flavius Frasincar, and Jasmijn Klinkhamer. Hierarchical deep learning for multi-label imbalanced text classification of economic literature. *Applied Soft Computing*, 176:113189, 2025. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2025.113189. URL https://www.sciencedirect.com/science/article/pii/S1568494625005009.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize F1 measure. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2014)*, volume 8725 of *Lecture Notes in Computer Science*, pp. 225–239, Germany, 2014. Springer. doi: 10.1007/978-3-662-44851-9_15.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 102307–102365. Curran Associates, Inc., 2024. doi: 10.52202/079017-3249. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b99a07486702417d3b1bd64ec2cf74ad-Paper-Conference.pdf.

Rundong Liu, Wenhan Liang, Weijun Luo, Yuxiang Song, He Zhang, Ruohua Xu, Yunfeng Li, and Ming Liu. Recent advances in hierarchical multi-label text classification: A survey. *CoRR*, abs/2307.16265, 2023. doi: 10.48550/ARXIV.2307.16265. URL https://doi.org/10.48550/arXiv.2307.16265.

Scott C. Lowe, Benjamin Misiuk, Isaac Xu, Shakhboz Abdulazizov, Amit R. Baroi, Alex C. Bastos, Merlin Best, Vicki Ferrini, Ariell Friedman, Deborah Hart, Ove Hoegh-Guldberg, Daniel Ierodiaconou, Julia Mackin-McLaughlin, Kathryn Markey, Pedro S. Menandro, Jacquomo Monk, Shreya Nemani, John O'Brien, Elizabeth Oh, Luba Y. Reshitnyk, Katleen Robert, Chris M. Roelfsema, Jessica A. Sameoto, Alexandre C. G. Schimel, Jordan A. Thomson, Brittany R. Wilson, Melisa C. Wong, Craig J. Brown, and Thomas Trappenberg. BenthicNet: A global compilation of seafloor images for deep learning applications. *Scientific Data*, 12(1):230, 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-04491-1. URL https://doi.org/10.1038/s41597-025-04491-1.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jlAjNL8z5cs.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, Can Zhao, Ziyue Xu, Yufan He, Greg Heinrich, Yee Man Law, Benjamin Simon, Stephanie Harmon, Stephen Aylward, Marc Edgar, Michael Zephyr, Song Han, Pavlo Molchanov, Baris Turkbey, Holger Roth, and Daguang Xu. Vila-m3: Enhancing vision-language models with medical expert knowledge, 2025. URL https://arxiv.org/abs/2411.12915.

Rodolfo M. Pereira, Yandre M.G. Costa, and Carlos N. Silla Jr. Toward hierarchical classification of imbalanced data using random resampling algorithms. *Information Sciences*, 578:344–363, 2021. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2021.07.033. URL https://www.sciencedirect.com/science/article/pii/S0020025521007234.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28270. URL https://doi.org/10.1609/aaai.v38i5.28270.

Allie M. Scovil, Heather Breeze, Jessica Wingfield, Sarah Marteinson, Alexa Goodman, and Noreen E. Kelly. Effects of marine debris on Atlantic Canadian species and ecosystems. Canadian Technical Report of Fisheries and Aquatic Sciences 3662, Fisheries and Oceans Canada, Maritimes Region, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, 2025.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv*, 2025. doi: 10.48550/arXiv.2508.10104. URL https://arxiv.org/abs/2508.10104.

Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZWzUA9zeAg.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3:1–13, 09 2009. doi: 10.4018/jdwm.2007070101.

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. URL https://arxiv.org/abs/1707.06642.

Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5075–5084. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/wehrmann18a.html.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, 2023. URL https://arxiv.org/abs/2303.03846.

YuK-Kwan Wong, Tuan-An To, Jipeng Zhang, Ziqiang Zheng, and Sai-Kit Yeung. Marineeval: Assessing the marine intelligence of vision-language models, 2025. URL https://arxiv.org/abs/2512.21126.

WoRMS Editorial Board. World Register of Marine Species. Available from https://www.marinespecies.org at VLIZ, 2026. Accessed 2026-08-25. https://doi.org/10.14284/170.

Isaac Xu, Benjamin Misiuk, Scott C. Lowe, Martin Gillis, Thomas Trappenberg, and Craig J. Brown. Hierarchical multi-label classification with missing information for benthic habitat imagery. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2024. doi: 10.1109/IJCNN60899.2024.10650176.

Isaac Xu, Martin Gillis, Ayushi Sharma, Benjamin Misiuk, Craig J. Brown, and Thomas Trappenberg. Improving detection of rare nodes in hierarchical multi-label learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=hf4zEWWIvE.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv*, 2025. URL https://arxiv.org/abs/2505.09388.

Sahiti Yerramilli, Jayant Sravan Tamarapalli, Tanmay Girish Kulkarni, Jonathan Francis, and Eric Nyberg. Semantic augmentation in images using language, 2025. URL https://arxiv.org/abs/2404.02353.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

Tianyi Zhang, Weiming Zhi, Joshua Mangelson, and Matthew Johnson-Roberson. Infinite leagues under the sea: Photorealistic 3d underwater terrain generation by latent fractal diffusion models, 2025. URL https://arxiv.org/abs/2503.06784.

# A Technical Details

## A.1 Attention-Based Pooling

We consider two architectures for producing a fixed-dimensional representation from variable-length token sequences, both sharing a common structure consisting of a contextual encoding stage followed by learned query attention pooling. In the self-attention pooling variant, given an input sequence $\mathbf{X} \in \mathbb{R}^{n_\mathrm{b} \times T \times d}$ of batch size $n_\mathrm{b}$, sequence length $T$, and embedding dimension $d$, we first apply multi-headed (four heads) self-attention with residual connections and layer normalisation (Ba et al., 2016), followed by a position-wise feed-forward network, yielding a contextually enriched sequence

Table 4: **VLM Inference Diagnostics.** Throughput, GPU memory allocation, and KV cache overhead for the variable proposal and captioning stages. Throughput is reported as mean ± std across batches. Dataset sizes differ between tasks due to filtering (variable proposal operates on the training partition and captioning operates on the full dataset).

| Task | Dataset | Dataset Size [samples] | Throughput [samples/s] | Peak Alloc [GB] (mean/max) | Peak Reserved [GB] (mean/max) | KV Overhead [GB] |
|---|---|---|---|---|---|---|
| Var. Proposal | BN-Sub | 102,118 | $5.73 \pm 0.22$ | 19.09 / 19.23 | 21.61 / 22.75 | 2.75 |
| | BN-Bio | 105,133 | $5.61 \pm 0.40$ | 19.02 / 19.21 | 20.64 / 21.15 | 2.68 |
| | FN | 8,980 | $1.65 \pm 0.10$ | 21.57 / 21.94 | 22.78 / 22.88 | 5.23 |
| Captioning | BN-Sub | 132,639 | $0.38 \pm 0.04$ | 17.06 / 17.13 | 21.15 / 21.15 | 0.72 |
| | BN-Bio | 142,637 | $0.29 \pm 0.01$ | 17.04 / 17.12 | 21.15 / 21.15 | 0.70 |
| | FN | 9,305 | $2.09 \pm 0.17$ | 18.90 / 19.17 | 22.80 / 22.80 | 2.56 |

Table 5: **Training Model Diagnostics.** Trainable parameters, throughput (mean ± std across batches), and GPU memory usage for each model variant.

| Dataset | Model | Num. Classes | Batch Size | Trainable Params. | Throughput [samples/s] | Model Mem [GB] | Peak Alloc. [GB] (mean/max) | Peak Reserved [GB] (mean/max) |
|---|---|---|---|---|---|---|---|---|
| BN-Sub | Txt | 23 | 64 | 19M | $985.96 \pm 50.77$ | 0.09 | 2.57 / 2.62 | 3.81 / 3.88 |
| | Img | 23 | 64 | 19M | $1{,}575.74 \pm 64.14$ | 0.07 | 1.64 / 1.64 | 1.74 / 1.74 |
| | Img + Txt | 23 | 64 | 19M | $1{,}281.58 \pm 59.64$ | 0.09 | 1.86 / 1.87 | 2.11 / 2.11 |
| | Router | 23 | 64 | 10M | $1{,}046.14 \pm 65.88$ | 0.26 | 1.28 / 1.30 | 2.96 / 3.01 |
| | Router$^\dagger$ | 23 | 64 | 66M | $443.79 \pm 24.70$ | 0.26 | 5.44 / 5.49 | 7.09 / 7.16 |
| BN-Bio | Txt | 293 | 64 | 19M | $980.61 \pm 121.98$ | 0.09 | 2.52 / 2.58 | 3.71 / 3.78 |
| | Img | 293 | 64 | 19M | $1{,}596.01 \pm 134.24$ | 0.09 | 1.65 / 1.65 | 1.76 / 1.76 |
| | Img + Txt | 293 | 64 | 19M | $1{,}306.87 \pm 50.50$ | 0.09 | 1.86 / 1.87 | 2.18 / 2.20 |
| | Router | 293 | 64 | 11M | $1{,}053.14 \pm 44.21$ | 0.28 | 1.28 / 1.30 | 2.91 / 2.96 |
| | Router$^\dagger$ | 293 | 64 | 69M | $446.30 \pm 9.74$ | 0.28 | 5.42 / 5.48 | 7.06 / 7.15 |
| FN | Txt | 192 | 64 | 19M | $996.81 \pm 82.17$ | 0.09 | 2.57 / 2.97 | 4.53 / 4.66 |
| | Img | 192 | 64 | 19M | $1{,}634.52 \pm 51.52$ | 0.07 | 1.65 / 1.65 | 1.74 / 1.74 |
| | Img + Txt | 192 | 64 | 19M | $1{,}310.38 \pm 39.97$ | 0.09 | 1.87 / 1.94 | 2.15 / 2.16 |
| | Router | 192 | 64 | 11M | $1{,}002.58 \pm 102.59$ | 0.27 | 1.29 / 1.43 | 3.56 / 3.65 |
| | Router$^\dagger$ | 192 | 64 | 68M | $450.18 \pm 17.97$ | 0.27 | 5.45 / 5.86 | 7.22 / 7.35 |

Table 6: **Training Hyperparameters.** Learning rates and training epochs across models.

| Dataset | Model | Learning Rate | Num. Epochs |
|---|---|---|---|
| BN$^*$ | Txt | 3e-5 | 35 |
| | Img | 1e-4 | 35 |
| | Img + Txt | 1e-4 | 35 |
| | Router | 1e-4 | 35 |
| | Router$^\dagger$ | 1e-4 | 35 |
| FN | Txt | 3e-5 | 35 |
| | Img | 1e-4 | 15 |
| | Img + Txt | 1e-4 | 15 |
| | Router | 1e-4 | 8 |
| | Router$^\dagger$ | 1e-4 | 8 |

$^*$Applies to both BN-Sub and BN-Bio.

Table 7: **Complementarity of Frozen Embeddings.** Independence ratio $P(\text{text correct} \mid \text{image wrong}) \,/\, P(\text{text correct})$ by hierarchy level, from $k$-NN classifiers on the frozen image and text embeddings. iNaturalist level 1 is excluded (image accuracy $99.6\%$).

| | L1 | L2 | L3 | L4 | L5 |
|---|---|---|---|---|---|
| BN-Sub | 0.43 | 0.39 | 0.50 | — | — |
| BN-Bio | 0.68 | 0.27 | 0.33 | 0.29 | — |
| iNaturalist | *excl.* | 0.61 | 0.63 | 0.64 | 0.58 |

$\tilde{\mathbf{X}} \in \mathbb{R}^{n_\text{b} \times T \times d}$. In the cross-attention fusion variant, two distinct token sequences are available, a query sequence $\mathbf{X}_q \in \mathbb{R}^{n_\text{b} \times N \times d}$ (image embeddings) and a key-value sequence $\mathbf{X}_{kv} \in \mathbb{R}^{n_\text{b} \times M \times d}$ (text embeddings), and we replace the self-attention encoder with an analogous cross-attention layer in which $\mathbf{X}_q$ attends over $\mathbf{X}_{kv}$, followed by the same residual connections, layer normalisation, and feed-forward network, producing a fused sequence $\tilde{\mathbf{X}}_q \in \mathbb{R}^{n_\text{b} \times N \times d}$ that integrates information from $\mathbf{X}_{kv}$ into the positions of $\mathbf{X}_q$. In both cases, the final pooling stage is identical: a single learnable query parameter $\mathbf{q} \in \mathbb{R}^d$ attends over the encoded sequence via multi-head cross-attention, followed by layer normalisation, yielding a fixed-dimensional representation $\mathbf{z} \in \mathbb{R}^d$. Padding masks are propagated through all attention stages to ensure that padded positions contribute neither to contextual encoding nor to the final pooled representation. The resulting vector $\mathbf{z}$ is subsequently passed to an HML constrained feed-forward classifier.

### A.2 Gated Router

Given image, text, and cross-modal streams, we extract pooled feature vectors $\mathbf{z}_\text{img}, \mathbf{z}_\text{txt}, \mathbf{z}_\text{cross} \in \mathbb{R}^d$ and pass each through a dedicated classification head for each stream. This yields per-stream logit matrices stacked into a tensor of shape $(n_\text{b}, 3, |\mathcal{V}|)$. A gating network ingests the concatenated features $[\mathbf{z}_\text{img};\, \mathbf{z}_\text{txt};\, \mathbf{z}_\text{cross}] \in \mathbb{R}^{3d}$ and produces per-sample, per-node softmax weights over the three modes. The gate applies layer normalisation (Ba et al., 2016), a linear projection from $3d$ to $d$, GELU activation with dropout, and a final linear layer from $d$ to $3|\mathcal{V}|$, reshaped to $(n_\text{b}, |\mathcal{V}|, 3)$ and softmax-normalised along the last axis. The fused logit for node $v$ is then

$$l_v = \sum_{s \in \{\text{img, txt, cross}\}} g_v^{(s)} \cdot l_v^{(s)}, \tag{5}$$

where $g_v^{(s)}$ is the gate weight and $l_v^{(s)}$ is the stream logit. Sigmoid activation and the hierarchical constraint from C-HMCNN are applied after blending.

### A.3 Diagnostics

All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU. Computational costs for VLM inference are reported in Table 4, and training diagnostics are provided in Table 5. We set the embedding dimension $d$ to 1024 for Qwen3-Embedding-8B (which uses Matryoshka representations (Kusupati et al., 2022)), matching the output dimensionality of DINOv3's ViT-H backbone. Each HML classifier head consists of a 3-layer MLP with hidden dimension 1024, GELU activations (Hendrycks & Gimpel, 2016), and dropout of 0.7, mapping from the 1024-dimensional embedding space to the label space $\mathbb{R}^{|\mathcal{V}|}$. We optimise with Adam (Kingma & Ba, 2015) using a weight decay of $1\text{e}{-4}$ across all experiments. Learning rates and training epochs are reported in Table 6. Training epochs were selected based on validation loss curves. BenthicNet's larger dataset was less sensitive to overfitting, permitting 35 epochs to convergence, whereas FathomNet required more careful tuning to mitigate overfitting.

## B Captioning Embeddings

Table 7 reports the frozen-embedding probe summarised in subsection 4.3 by hierarchy level. iNaturalist level 1 is excluded because image accuracy there reaches $99.6\%$, leaving a conditioning set of

Table 8: **Caption Sources.** Independence ratio (as in Table 7) on BN-Bio (OOD test, 20,000 samples) for five caption sources, each embedded with Qwen3-Embedding-8B. Lower values indicate greater redundancy with the image embedding.

| Captions | L1 | L2 | L3 | L4 |
|---|---|---|---|---|
| Elicited (ours) | 0.68 | 0.27 | 0.33 | 0.29 |
| Keyword prose (ours) | 0.64 | 0.27 | 0.33 | 0.32 |
| Description | 0.48 | 0.27 | 0.34 | 0.33 |
| Visual | 0.41 | 0.27 | 0.33 | 0.28 |
| Plain | 0.36 | 0.26 | 0.32 | 0.27 |

only $\sim$89 samples. We then ask whether this complementarity depends on the captions being built on the extracted variables.

We compare the elicited captions against four alternatives generated by the same VLM for the same images: a plain caption, a visual description restricted to shapes, colours and textures, a description of the organisms and seafloor, and the same description additionally asked to relate the image to the extracted variables (Appendix D). At the coarsest level of the hierarchy, the captions built on the extracted variables are markedly less redundant with the image than the generic captions (Table 8). The description and keyword prose prompts differ only by the reference to the extracted variables, and this addition alone accounts for most of the gap. At deeper levels the five sources are comparable; however, we note that these deeper levels are where we expect text to be less useful in general.

# C Dataset Details

## C.1 BenthicNet

BenthicNet (Lowe et al., 2025) is one of the largest publicly available ocean science datasets, comprising over 11 million benthic images aggregated from 2,058 constituent datasets spanning a wide variety of geographic regions, depth ranges, and imaging platforms, sourced from research groups and government agencies around the world. Its label taxonomy follows the Collaborative and Annotation Tools for Analysis of Marine Imagery (CATAMI) visual classification scheme (Althaus et al., 2015), and is unique in that annotations span both substrate and biota categories, facilitating both habitat and organism classification. We draw from it its full set of annotated imagery (188,679 images), forming two overlapping subsets for substrate and biota (BN-Sub and BN-Bio, respectively), which we treat as independent classification tasks. BN-Sub contains 23 nodes while BN-Bio contains 293. BenthicNet's test partition is geospatially subsampled, meaning test images originate from regions unseen during training. We refer to this innate partition as the out-of-domain (OOD) test set, and additionally construct an in-domain (ID) test set through random stratified sampling (of 20% of the data) for evaluation under more conventional conditions. The combination of global geographic coverage and diversity across both biota and substrate categories makes BenthicNet a suitable testbed for HML methods that must generalise across heterogeneous visual conditions.

## C.2 FathomNet (2025)

FathomNet (2025) (FN) (FathomNet 2025) is a benthic biota classification competition dataset consisting of 9,305 full images and 24,487 regions-of-interest (ROIs). Full images may contain multiple ROIs. The dataset is not inherently hierarchical, despite using taxonomic annotations. Furthermore, it has been curated to be balanced across classes. We extracted a hierarchy of 192 nodes spanning phylum, class, order, family, genus, and species[3] from the World Register of Marine Species (WoRMS) (WoRMS Editorial Board, 2026). Subsequently, the classes in the dataset become leaf nodes, and the problem becomes HML (full-depth). While earlier imagery from FathomNet was contained in BenthicNet, the 2025 competition dataset is novel with a carefully curated OOD test set (Lowe et al., 2025; FathomNet 2025). As with BenthicNet, we use both the innate test partition and a random partition (of 20%) for OOD and ID evaluation respectively.

[3]For some annotations, these levels were replaced with a sub-level alternative where necessary.

### C.3 iNATURALIST

To demonstrate the generalisability of the proposed approach, experiments are also performed on the iNaturalist (iNat) (Van Horn et al., 2018) dataset. The iNat dataset encompasses a broader domain than BenthicNet, capturing terrestrial fauna, flora, and fungi. Additionally, it presents a distinct structural classification challenge. While the dataset inherently contains 10,000 distinct species classes, its biological taxonomic hierarchy allows for the construction of a hierarchical classification problem spanning over 16,000 nodes, characterised by mutually exclusive, single-label assignments. In adapting our framework for iNat, we replace MCLoss with a softmax over the 10,000 endpoints, marginalised up the hierarchy, so that each node receives the summed probability of its descendants and the hierarchical structure is preserved. Training occurs on the 500,000 samples from iNat (mini), and the 100,000 validation samples act as a test set (since the true test partition is not publicly available).

## D PROMPT TEMPLATES

We use two prompt templates for variable proposal and captioning, and four further templates for the caption comparison in Appendix B. Where applicable, templates are parameterised by task-specific variables denoted with braces.

### D.1 VARIABLE PROPOSAL PROMPT

Variable Proposal Prompt

```
You are a marine benthic ecologist.
This seafloor image has the following annotations: {annotations}.
Use these annotations to guide which ecological properties would be
most diagnostic for identifying the annotated categories.

Task: Infer functional ecological properties that help distinguish
between different {task}.

Go beyond what a simple visual description would capture -- use
ecological reasoning to infer properties from what you observe.
Focus especially on diagnostic properties where ecological
and world knowledge would help resolve highly specific or rare
categories.

Do NOT return simple visual descriptions (e.g., color, shape,
texture). Do not restate the annotation categories themselves.
Return up to 10 of the most diagnostic properties as 1-3 lowercase
words each, separated by commas. Return nothing else.
```

### D.2 CAPTIONING PROMPT

Captioning Prompt

```
You are a marine benthic ecologist analyzing a seafloor photograph.

Task: For each ecological property below, carefully reason about
whether it is supported or contradicted by visible evidence of the
{task} in this image. Connect what you SEE to what you KNOW about
benthic ecology. Score each property:
1 = visually supported
0 = not applicable or visually contradicted

Properties:
{variables}

Format:
[property_term]: [brief reasoning and justification (one sentence,
~5 words)] [score value]
Adhere strictly to the format. Output one property per line.
```

Here, `{annotations}` denotes the ground-truth categories expressed in natural language, `{task}` specifies the ecological domain (e.g., "biota" or "substrate characteristics"), and `{variables}` is the alphabetically sorted set of all processed variables for the domain.

The following prompts generate the alternative captions compared in Appendix B. All four use the same VLM and decoding settings as the captioning prompt.

**Plain Caption Prompt**

```
Describe the image in a single paragraph of about 150 words.
```

**Visual Caption Prompt**

```
Describe in detail what is directly visible in the image: shapes,
colours, textures. Write a single paragraph of about 150 words.
```

**Description Prompt**

```
You are looking at a seafloor photograph.
Task: Describe in detail what is visible in the image: the
organisms present and the seafloor.
Write a single paragraph of about 150 words. Return nothing else.
```

**Keyword Prose Prompt**

```
You are looking at a seafloor photograph.
Task: Describe in detail what is visible in the image: the
organisms present and the seafloor. Where relevant, relate your
description to the following ecological properties: {variables}.
Write a single paragraph of about 150 words. Return nothing else.
```

# E Latent Extraction Details

This section provides the complete mathematical specification of the variable extraction pipeline summarised in subsection 3.1.

## E.1 Step 1: Node Frequency Prior

We compute the empirical distribution over nodes from the training annotations:

$$p(v) = \frac{n_v}{\sum_{v' \in \mathcal{V}} n_{v'}}, \quad \forall\, v \in \mathcal{V}. \tag{6}$$

## E.2 Step 2: EM Decomposition

We define the observed annotations for a candidate $z_{2,j} \in \mathbf{Z}_2$ as a set of label-count pairs $\{(\mathcal{A}_k, m_k)\}$, where $\mathcal{A}_k \subseteq \mathcal{V}$ is a set of co-occurring partial-depth paths within a single sample and $m_k$ is the number of samples associated with $z_{2,j}$ that carry annotation set $\mathcal{A}_k$. We partition these into single-path pairs $\mathcal{S}_j = \{(\mathcal{A}_k, m_k) : |\mathcal{A}_k| = 1\}$ and multi-path pairs $\mathcal{M}_j = \{(\mathcal{A}_k, m_k) : |\mathcal{A}_k| > 1\}$. Let $\mathcal{S}_j(v) = \{k : (\mathcal{A}_k, m_k) \in \mathcal{S}_j,\ v \in \mathcal{A}_k\}$ and $\mathcal{M}_j(v) = \{k : (\mathcal{A}_k, m_k) \in \mathcal{M}_j,\ v \in \mathcal{A}_k\}$.

Counts are initialised from unambiguous single-path instances:

$$n_j^{(0)}(v) = \sum_{k \in \mathcal{S}_j(v)} m_k. \tag{7}$$

Multi-path counts are then iteratively distributed in proportion to current estimates:

$$n_j^{(t+1)}(v) = n_j^{(0)}(v) \; + \sum_{k \in \mathcal{M}_j(v)} m_k \cdot \frac{n_j^{(t)}(v)}{\sum_{v' \in \mathcal{A}_k} n_j^{(t)}(v')}, \tag{8}$$

iterating until convergence:

$$\max_{v \in \mathcal{V}} \left| n_j^{(t+1)}(v) - n_j^{(t)}(v) \right| < \tau. \tag{9}$$

The resulting distribution over nodes for each $z_{2,j}$ is:

$$q_j(v) = \frac{n_j(v)}{\sum_{v' \in \mathcal{V}} n_j(v')}. \tag{10}$$

### E.3 Step 3: Smoothing and KL Scoring

Before computing divergence, we smooth each $q_j$ towards the prior $p$ by adding a pseudo-count volume equivalent to Laplace smoothing, distributed according to $p$:

$$\tilde{q}_j(v) = \frac{n_j(v) + |\mathcal{V}| \cdot p(v)}{N_j + |\mathcal{V}|}, \tag{11}$$

where $N_j = \sum_{v'} n_j(v')$ is the total count mass for $z_{2,j}$. This smoothing serves as a sample-size-aware regulariser. A variable supported by many annotated instances will have large $N_j$, rendering the pseudo-counts negligible and preserving the empirical distribution, whereas a variable supported by only a handful of instances will be pulled substantially towards the prior.

A key assumption underlies this scoring procedure. The empirical posterior $q_j(v)$ is constructed from the same annotation source from which the candidate variables were originally proposed, so high $q_j(v)$ reflects co-occurrence within the annotation pipeline rather than an independently observed predictive relationship. We assume that this co-occurrence is indicative of genuine discriminative value. In other words, a candidate $z_{2,j}$ whose annotation mass concentrates on a node $v$ does indeed capture a visual or semantic property informative for distinguishing $v$ from other nodes. This assumption may not hold universally, but provides the working basis for using the KL divergence as a proxy for discriminative value. Candidates are then scored by their divergence from the prior,

$$D_{\mathrm{KL}}\big(\tilde{q}_j \parallel p\big) = \sum_{v \in \mathcal{V}} \tilde{q}_j(v) \log \frac{\tilde{q}_j(v)}{p(v)}, \tag{12}$$

and only candidates scoring $\geq 1$ are retained.

### E.4 Step 4: Deduplication and Selection

The candidate set emerging from $D_{\mathrm{KL}}$ filtering undergoes three stages of refinement.

**Distribution-based deduplication.** Each surviving candidate is represented by its normalised distribution $\tilde{q}_j$. Processing in descending KL order, a candidate is discarded if its cosine similarity with any previously retained distribution meets or exceeds a threshold $\tau_{\mathrm{dist}}$:

$$\frac{\tilde{q}_j \cdot \tilde{q}_k}{\|\tilde{q}_j\| \, \|\tilde{q}_k\|} \;\geq\; \tau_{\mathrm{dist}}. \tag{13}$$

We use $\tau_{\mathrm{dist}} = 0.99$.

**Coverage-based deduplication.** Let $\mathcal{I}(z_{2,j}) \subseteq \mathbf{X}$ denote the set of training images whose proposed-variable field contains $z_{2,j}$. Candidates are processed in descending KL order. A candidate is discarded if its image set has Jaccard similarity at or above $\tau_{\mathrm{cov}}$ with any previously retained candidate:

$$J\big(\mathcal{I}(z_{2,j}), \mathcal{I}(z_{2,k})\big) = \frac{|\mathcal{I}(z_{2,j}) \cap \mathcal{I}(z_{2,k})|}{|\mathcal{I}(z_{2,j}) \cup \mathcal{I}(z_{2,k})|} \;\geq\; \tau_{\mathrm{cov}}. \tag{14}$$

We use $\tau_{\mathrm{cov}} = 0.8$.

**Per-node top-$B$ selection.** Surviving candidates are assigned to their modal node $v_j^* = \arg\max_{v \in \mathcal{V}} n_j(v)$, inducing per-node groups $\mathcal{Z}(v) = \{z_{2,j} : v_j^* = v\}$. A uniform budget $B$ is applied to leaf nodes $\mathcal{V}_L \subseteq \mathcal{V}$, prioritising rarity. For each leaf node $v$ with $\mathcal{Z}(v) \neq \emptyset$, the selected subset is:

$$S_v = \text{top}_B\Big(\mathcal{Z}(v),\ \text{KL}\big(\tilde{q}_j \,\|\, p\big)\Big). \tag{15}$$

The final variable set is $\bigcup_{v \in \mathcal{V}_L} S_v$.

# F Limitations

## F.1 Variable Proposal and Captioning

The quality of proposed variables and captions is bounded by the VLM's pretraining corpus. Domains or taxa that are poorly represented in training data may yield superficial or inaccurate variables, and the model's familiarity is not uniform across the label space. Well-studied organisms (e.g., hard corals) are likely to be better detected than understudied groups (e.g., soft corals), introducing a systematic bias in caption informativeness.

The uniform per-node budget $B$ and conservative variable selection reflect pragmatic constraints imposed by the limited token capacity of smaller VLMs. When provided with too many variables, models at this scale tend to anchor onto variable names without meaningful contextual processing (Shi et al., 2023; Wei et al., 2023). A larger VLM or a multi-pass captioning scheme could improve caption quality. Lastly, while it may only need to be performed once, the captioning cost may be quite high, as demonstrated in Table 4. Computation cost in this step can also vary significantly depending on the phrasing used in the prompt.

## F.2 Architecture and Training

The text and image embedding models (Qwen3-Embedding-8B and DINOv3 ViT-H respectively) are frozen throughout training. Their representation spaces are therefore not adapted to the target domain, and any distributional mismatch between pretraining data and benthic imagery is inherited by the downstream classifier. Fine-tuning these backbones could improve performance but would substantially increase computational requirements and reduce accessibility for domain-specialised labs with limited hardware.

The router models introduce additional parameters relative to the unimodal baselines (by a factor of 3, given each stream, and the router mechanism), and one variant (Router) loads pretrained pooling layers from unimodal runs. We therefore assess the text contribution with ablations that hold the architecture and parameter count fixed while removing the per-sample text signal (subsection 4.3; subsection G.2).

## F.3 Evaluation Scope

Our benthic experiments, our primary application focus, are restricted to two datasets (BenthicNet (Lowe et al., 2025) and FathomNet (Katija et al., 2022; FathomNet 2025)), which represent the few large-scale, globally diverse, and taxonomically broad benchmarks available in this field. Beyond the benthic domain, we evaluate only on iNaturalist. While the core methodology does not rely on domain-specific assumptions, its extension to other specialised domains, such as medical imaging, remains empirically undemonstrated and is left to future work.

## F.4 Evaluation Metric

The inverted HML score $d_{\text{inv}}$ relies on a user-specified parameter $k_{\text{P}}$ that governs the severity of the cardinality penalty. Our choice of $k_{\text{P}} = 0.5$ represents a balanced default. Where the expected unannotated multi-label cardinality of a dataset is suspected, $k_{\text{P}}$ can be adjusted to reflect it.

Additionally, $d_{\text{inv}}$ treats all edges in the hierarchy as having unit weight, which may not reflect the true semantic distance between nodes. For example, the distinction between two congeneric species is typically far smaller than that between two phyla, yet both correspond to two edge traversals in a

Table 9: **Zero-Shot VLM Classification on BenthicNet.** Scheme conformance and prediction cardinality for free-form prompting and for prompting conditioned on the classification scheme.

| | BN-Sub | BN-Bio |
|---|---|---|
| images yielding no scheme-conformant term (free-form) | 92.0% | 77.6% |
| conformance when conditioned on the scheme | 99.9% | 98.0% |
| predicted / annotated labels per image (conditioned) | 2.08 | 6.98 |

Table 10: **Scale-Matched Noise Ablation.** Median scores across seeds for the Router with clean text embeddings and retrained with scale-matched Gaussian noise in their place; $p$ is a two-sided Wilcoxon signed-rank test over paired seeds. Since iNat has only five seeds, the lowest attainable $p$-value is 0.063.

| | | Macro $F_1$ | | | Macro $d_{\text{inv}}$ | | |
|---|---|---|---|---|---|---|---|
| dataset | seeds | clean | noise | $p$ | clean | noise | $p$ |
| BN-Sub | 10 | 0.678 | 0.241 | 0.002 | 0.654 | 0.006 | 0.002 |
| BN-Bio | 10 | 0.129 | 0.047 | 0.002 | 0.037 | 0.001 | 0.002 |
| FathomNet | 30 | 0.410 | 0.414 | 0.670 | 0.323 | 0.316 | 0.008 |
| iNaturalist | 5 | 0.549 | 0.190 | 0.063 | 0.770 | 0.388 | 0.063 |

taxonomic tree. Depth-specific edge weights could be incorporated where expert knowledge provides meaningful inter-node distances, but we do not explore this idea in this work.

# G Additional Ablations

## G.1 Zero-Shot VLM Classification

A natural question is whether the VLM can perform the classification directly, without the elicitation and training pipeline. We evaluate this on BenthicNet in two prompting regimes under greedy decoding, and report structural information rather than performance (Table 9).

Free-form output cannot be scored. Both macro $F_1$ and $d_{\text{inv}}$ require an ancestor-closed subset of $\mathcal{V}$, whereas the VLM emits unconstrained text. Matching that text back to the scheme recovers 0.10 nodes per image on BN-Sub and 0.87 on BN-Bio, against ground truth cardinalities of 1.42 and 3.09. Only 0.5% and 3.3% of emitted terms correspond to any CATAMI category, and 92.0% and 77.6% of images yield no scheme-conformant prediction at all. Any score reported under this regime would be a property of the matching rule rather than of the model.

Supplying the vocabulary in the prompt helps with conformance but worsens cardinality. Conformance rises to 99.9% on BN-Sub and 98.0% on BN-Bio, while predicted cardinality inflates to 2.08 and 6.98 times ground truth. On BN-Bio the failure is bimodal: the median image returns 6 categories, 6.3% return more than 54, and the worst case returns 148 against a single annotated endpoint. This is the regime $d_{\text{inv}}$ penalises and micro-averaged AP tolerates (Appendix I): conformant predictions that share correct ancestors while carrying almost no discriminative information. It is also the reason a per-node budget $B$ is necessary (Appendix E), since instability scales with the size of the vocabulary placed in the prompt.

Neither regime yields a usable classifier. Knowledge held in the weights does not make a specialised label space available as a prediction target, which motivates using the VLM to elicit properties rather than to predict nodes.

## G.2 Scale-Matched Noise Ablation

To separate the contribution of the text content from that of the additional input dimensions, we replace the text embedding with Gaussian noise of matched scale at a mixing coefficient of one, so that no original signal remains, and retrain under the identical Router protocol. Architecture,

Table 11: **Descriptor Matching on BN-Bio.** Macro $F_1$ and macro $d_{\text{inv}}$ (%) of a DCLIP-style baseline (Menon & Vondrick, 2023), with and without the hierarchical constraint $f_{\text{CM}}$, alongside the trained models. Descriptors are embedded with the CLIP ViT-L-14-quickgelu text encoder. DCLIP rows are single deterministic runs; trained rows are the ten-seed OOD values of Table 1.

| | Macro $F_1$ | Macro $d_{\text{inv}}$ |
|---|---|---|
| DCLIP, no constraint | 5.8 | 0.12 |
| DCLIP + $f_{\text{CM}}$ | 5.8 | 0.17 |
| Text-only (trained) | 7.4 | 2.5 |
| Image-only (trained) | 11.4 | 4.4 |
| Router (trained) | 13.2 | 5.9 |
| Router$^{\dagger}$ (trained) | 14.3 | 6.4 |

embedding dimensionality and trainable parameter count are held constant, and the model trains from the start on an uninformative stream, free to disregard it in favour of the image branch.

Results are given in Table 10. On both BenthicNet domains and on iNat the corrupted model degrades on both metrics at every seed, and falls below the image-only baseline of Table 1 and Table 2. Additional feature dimensions should not make a model worse than not having them, so a gain arising from dimensionality alone would not produce this. This ablation separates content from dimensionality. A comparison against other semantically valid captions is given in Appendix B.

On FathomNet the effect is significant but small, and macro $F_1$ is unchanged. Two factors plausibly limit what the VLM can supply for this dataset: annotations given as Latin binomials rather than descriptive terms, and the limited representation of deep-sea organisms in general-purpose corpora relative to terrestrial fauna. We do not attempt to separate these.

### G.3 Descriptor-Matching Baseline

We compare against a descriptor-matching baseline following Menon & Vondrick (2023), which classifies images by their similarity to LLM-generated visual descriptors rather than to class names alone. Because the method produces flat outputs, we evaluate it both with and without the hierarchical constraint of section 3.

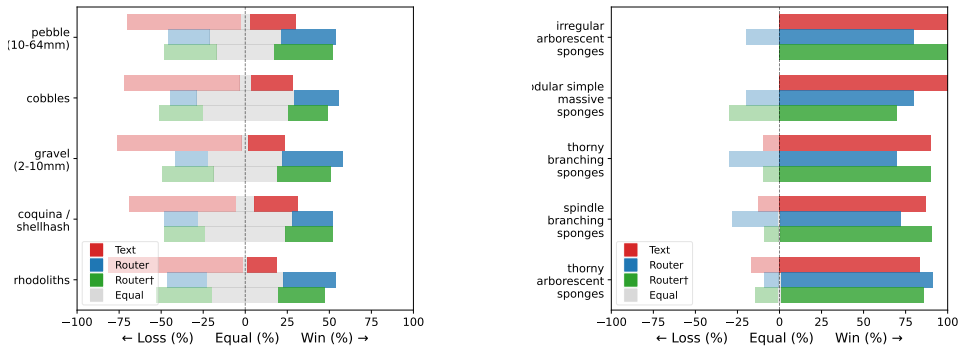
For each of the 293 BN-Bio nodes, the VLM used elsewhere in our pipeline generates six visual descriptors, prompted with the node and its position in the hierarchy and restricted to appearance, giving 1,758 descriptors in total. Each descriptor is embedded once with the CLIP text encoder and scored against every image by cosine similarity, and a node's score is the mean over its six descriptors. Normalisation and per-node thresholds are fitted on the 105,133 training images and applied unchanged to the 37,504 test images, so the baseline receives the same per-node thresholding as every model in Table 1.

The baseline reaches roughly half the macro $F_1$ of the trained image branch and under a tenth of its $d_{\text{inv}}$ (Table 11). Adding the hierarchical constraint leaves macro $F_1$ unchanged and raises macro $d_{\text{inv}}$ from 0.12 to 0.17, while the sample-mean $d_{\text{inv}}$ stays at 0.11, so the gain falls on nodes that carry little weight in the sample mean.

The descriptors account for much of this. Of the 1,758 generated, 22.5% are verbatim duplicates of a descriptor issued for another node: *smooth or slightly rough surface* is assigned to 35 nodes and *pale beige to brownish hue* to 47. Duplication is most common between siblings, where 24% of the 769 sibling pairs share an identical descriptor. For example, *Macroalgae* $>$ *Laminate* $>$ *Brown* and $>$ *Red* receive five identical descriptors out of six, none of which mentions colour, although the prompt supplied both hierarchy paths and asked for distinguishing features. Node scores reflect this overlap: sibling nodes correlate at $r = 0.825$, against $0.469$ for non-siblings, and the correlation among nodes at the same depth rises from $0.36$ at depth one to $0.95$ at depth six. In this domain, appearance-based descriptors separate coarse categories but not the morphological distinctions on which CATAMI subdivides.

Our variables also operate at the clade level; the difference lies in redundancy. A visual descriptor names something a frozen visual encoder already extracts, whereas the properties elicited in section 3 are not recoverable from the image alone. This is consistent with Appendix B, where, at the coarsest level of the hierarchy, generic image descriptions are markedly more redundant with the image embedding than captions built on the elicited variables.

## H BN-Bio (OOD) Case Study




(a) BN-Sub (OOD)

(b) BN-Bio (OOD)

Figure 4: **Modality Information Exclusivity.** Displayed are the top-5 nodes in which text holds the greatest advantages over vision in the BenthicNet datasets in terms of $d_{\text{inv}}$. The bars depict the percentage of samples for each node in which the models perform less than, equal to, or greater than vision.

The correlation analysis between proposed variables and text advantage (Figure 5a) reveals that behavioural and habitat traits (*ambush*, *high mobility*, *benthic habitat*) are significantly positively correlated with text modality performance gains, while morphological plant traits (*light interception*, *leaf morphology*) exhibit significant negative correlations. However, examining text advantage at the level of individual nodes (Figure 5b) reveals that a variable annotated as absent (Property: [0]) for a given sample may still contribute positively to the discriminability of relevant classes. The explicit negation of a property, along with its caption, may serve as part of a combinatorial signature that distinguishes a node from its siblings.

Despite this argument, we emphasise that these correlations are observational and do not establish a causal mechanism by which any individual variable improves classification. While connections can be manually verified in specific cases, we cannot generalise this reasoning to all variable–node pairs. Some variables may contribute indirectly through interactions with other annotations, or may serve primarily as prompt structure that encourages the captioning VLM to attend to relevant image regions. Disentangling these effects would require controlled ablations at the individual variable level, beyond the scope of this work.

### H.1 Node-Wise Analysis

To characterise where the text modality contributes, we compare the text and image branches node by node on every BN-Bio node with at least 20 positive test samples (171 nodes).

On its own, the text branch has lower $F_1$ than the image branch at 157 of the 171 nodes. The gap is recall-driven at 105 of these nodes and precision-driven at 52, so captions more often omit a discriminating property than assert an incorrect one. This matches the clade-level character of the elicited variables: *filter feeding*, for example, separates sponges from other taxa, placing a prediction within the correct subtree without resolving the morphotype within it.

The two metrics diverge in a consistent direction. At 75 of the 171 nodes, the text branch lands closer to the target on $d_{\text{inv}}$ while failing to identify the node on $F_1$; the reverse holds at only 6. Text therefore contributes hierarchical proximity where it does not contribute discrimination.

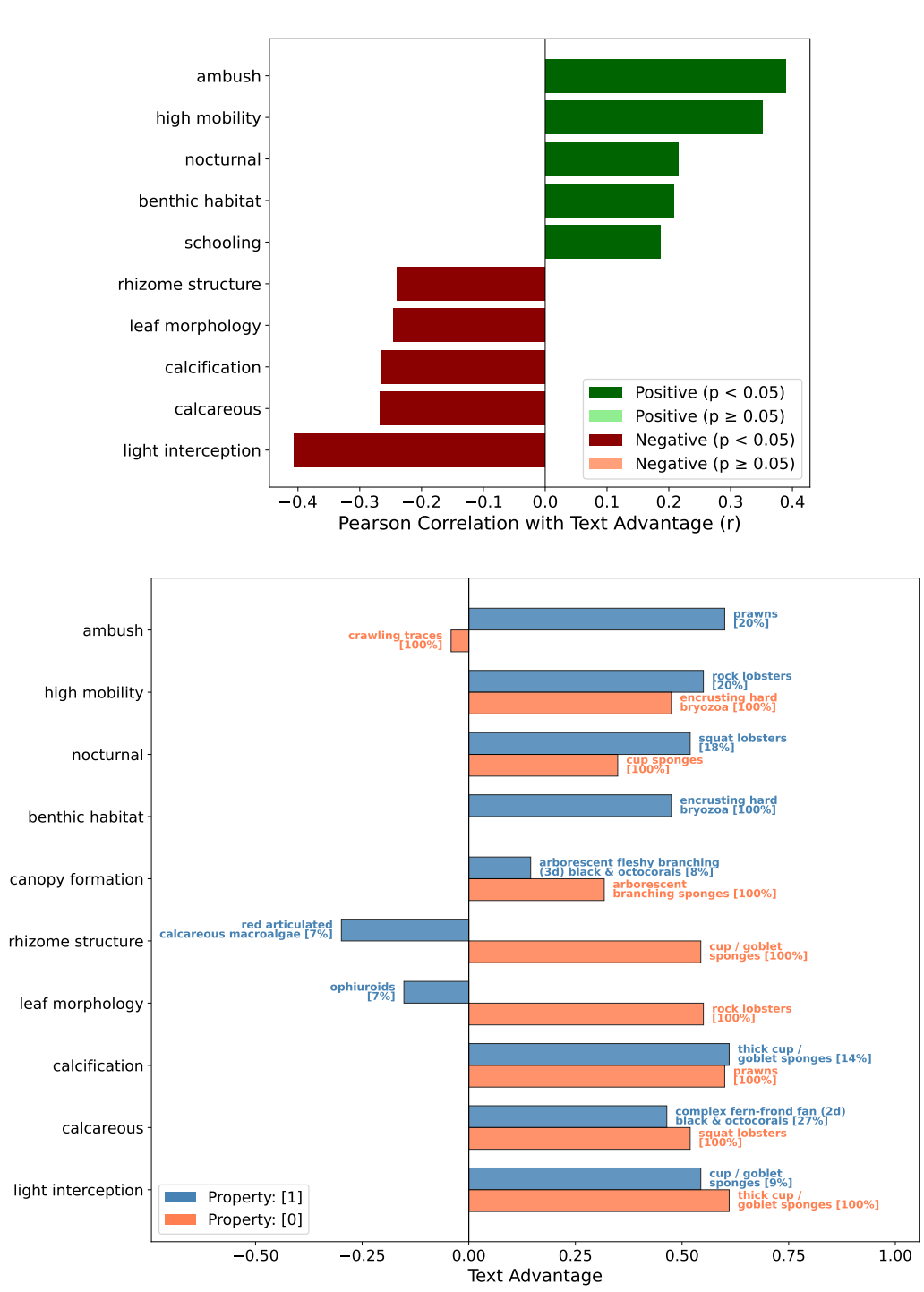



Figure 5: **Variable Correlations with Text Modality Advantage.** (a) Pearson correlation between each variable's positive annotation in captions and text advantage across all nodes. Dark shading indicates statistical significance ($p < 0.05$). (b) For each top-correlated variable (sorted by $r$), we select a representative class where the variable benefits text when it is positive [1] (blue) and one where it is negative [0]. Sample coverage is noted in brackets. A soft diversity penalty discourages repeated selection of the same class across properties. Since the majority of samples have a positive annotation for "benthic habitat", it is the only variable without a negative bar.

Table 12: **Gate Weights under Text Ablation on BN-Bio.** Text gate weight $w_{\mathrm{txt}}$ with clean captions and with every caption replaced by the mean embedding, the relative change, and the number of seeds in which the value decreases.

| | clean | constant text | $\Delta$ | seeds lower |
|---|---|---|---|---|
| mean $w_{\mathrm{txt}}$ | 0.5658 | 0.5455 | $-3.6\%$ | 5/5 |
| per-sample std. of $w_{\mathrm{txt}}$ | 0.1315 | 0.1306 | $-0.68\%$ | — |

In the fused model, routing recovers the text branch's $F_1$ deficit at 130 of the 157 nodes, and at all 21 measurable BN-Sub nodes, with a median gain of 0.018 $F_1$ over the image branch. At the remaining 27 nodes, the median cost is 0.005 $F_1$. These nodes are concentrated in sponges, which make up 24.6% of the nodes examined but 55.6% of those not recovered, a 2.3-fold concentration. CATAMI subdivides sponges by morphology, which our proposal prompt excludes by design (subsection D.1), so these nodes fall where the clade-level character of the variables predicts.

### H.2 Gate Weights

The router assigns per-node weights over the image, text and cross-attention branches. Table 12 reports the mean text weight and its per-sample standard deviation with and without per-sample text information.

The gate does not ignore the text branch. On BN-Bio, the mean text weight is $0.57$. When every caption is replaced by the same payload, the mean text weight falls by $3.6\%$ ($0.566 \to 0.546$), lower at every seed, while its per-sample standard deviation changes by less than $1\%$. The sample-to-sample variation of the gate is therefore driven mainly by the image side of its input rather than by the content of individual captions.

These weights describe how much the fused output draws on each branch, not how the text branch transforms its input. Because the gate mixes branch logits, it does not show how the text embedding is used within the model; that contribution is measured directly by the ablations in subsection 4.3 and subsection G.2.

## I HML Metrics Noise Experiment

**Algorithm 1** Random Hierarchy Generation

**Require:** Number of nodes $|\mathcal{V}|$
**Ensure:** Descendant matrix $\mathbf{D} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$
Select number of root nodes $n_r \in [1, \max(1, \lfloor |\mathcal{V}|/100 \rfloor)]$ uniformly at random
Initialise a forest with root nodes $v = 0, \ldots, n_r - 1$
**for** each remaining node $v = n_r, \ldots, |\mathcal{V}| - 1$ **do**
    Assign $v$ as a child of a node chosen uniformly at random from $\{0, \ldots, v-1\}$
**end for**
$\mathbf{D} \leftarrow \mathbf{0}^{|\mathcal{V}|\times|\mathcal{V}|}$
**for** each node $v \in \mathcal{V}$ **do**
    Set $D_{v,u} \leftarrow 1$ for all descendants $u$ of $v$ (including $v$ itself)
**end for**
**return** $\mathbf{D}$

A natural notion of a "random" model in HML is one that predicts each node independently with uniform probability, analogous to the random baseline in multi-class classification. Such a model produces predictions with expected cardinality near $|\mathcal{V}|/2$, where the number of combinatorial states is greatest. Crucially, the hierarchical constraint forces false positives to propagate upward since activating a node requires activating all of its ancestors. Consequently, higher-level nodes are disproportionately likely to be active. This stacking effect means that even moderate noise inflates cardinality rapidly. Both AP and $F_1$ decay under this noise regime, and do so more steeply as hierarchy size grows, since the propagation affects a larger number of ancestor nodes in absolute

**Algorithm 2** Random Hierarchical Data Generation

**Require:** Number of nodes $|\mathcal{V}|$, number of samples $N$, descendant matrix $\mathbf{D}$, sparsity $s$
**Ensure:** Data matrix $\mathbf{Y} \in \{0,1\}^{N \times |\mathcal{V}|}$

Define neighbourhood $\delta = \max(1, \lfloor s/2 \rfloor)$
$\mathbf{Y} \leftarrow \mathbf{0}^{N \times |\mathcal{V}|}$
**for** each sample $i = 1, \ldots, N$ **do**
    Sample $k$ uniformly from $[\max(1, s-\delta),\ \min(|\mathcal{V}|, s+\delta)]$
    Select $k$ nodes uniformly at random from $\mathcal{V}$
    Activate selected nodes: $\mathbf{r} \in \{0,1\}^{|\mathcal{V}|}$
    Propagate through hierarchy: $\mathbf{Y}_i \leftarrow \text{clip}(\mathbf{D} \cdot \mathbf{r},\ 0,\ 1)$ ▷ Enforce ancestor activation
**end for**
**return** $\mathbf{Y}$

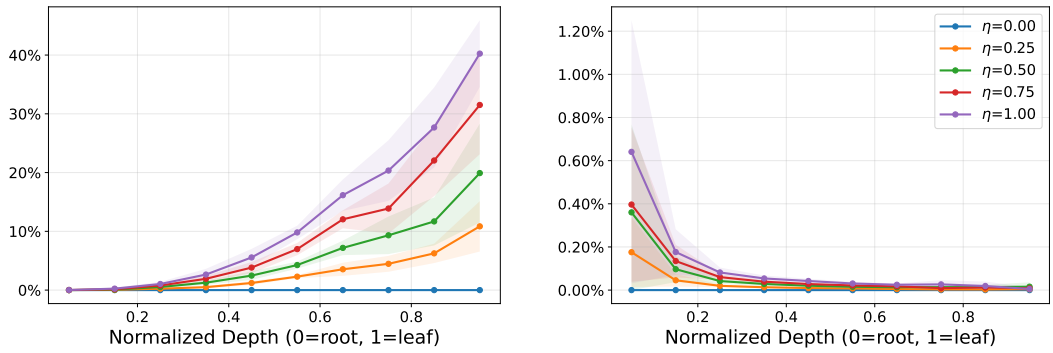
**Algorithm 3** Hierarchical Noise Generation

**Require:** Ground-truth matrix $\mathbf{Y} \in \{0,1\}^{N \times |\mathcal{V}|}$, descendant matrix $\mathbf{D}$, noise level $\eta \in [0,1]$
**Ensure:** Noisy predictions $\hat{\mathbf{Y}}$, prediction scores $\hat{\mathbf{S}}$

Compute depth $\ell_v \leftarrow \sum_u D_{u,v} - 1$ and normalised depth $d_v \leftarrow \ell_v / \max_u \ell_u$ for each node $v \in \mathcal{V}$
Define FN rate per node: $r_v^{\text{FN}} = 0.5\,\eta\, d_v^{1.5}$ ▷ Increases with depth
Define FP rate per node: $r_v^{\text{FP}} = 10^{-4}\,\eta\,\big(1 + 0.5\,(1 - d_v)\big)$ ▷ Higher near root
$\hat{\mathbf{Y}} \leftarrow \mathbf{Y}$ ▷ Initialise predictions from ground truth
**for** each node $v \in \mathcal{V}$ **do**
    **for** each sample $i$ where $Y_{iv} = 1$ **do**
        Flip $\hat{Y}_{iv} \leftarrow 0$ with probability $r_v^{\text{FN}}$ ▷ Inject false negatives
    **end for**
    **for** each sample $i$ where $Y_{iv} = 0$ **do**
        Flip $\hat{Y}_{iv} \leftarrow 1$ with probability $r_v^{\text{FP}}$ ▷ Inject false positives
    **end for**
**end for**
**for** each sample $i = 1, \ldots, N$ **do** ▷ Enforce hierarchical consistency
    $\hat{\mathbf{Y}}_i \leftarrow \text{clip}(\mathbf{D} \cdot \hat{\mathbf{Y}}_i,\ 0,\ 1)$
**end for**
**for** each sample $i$ and node $v \in \mathcal{V}$ **do** ▷ Scores for AP
    **if** $\hat{Y}_{iv} = 1$ **then**
        $\hat{S}_{iv} \sim \mathcal{U}(0.55, 0.95)$
    **else**
        $\hat{S}_{iv} \sim \mathcal{U}(0.05, 0.45)$
    **end if**
**end for**
**return** $\hat{\mathbf{Y}}, \hat{\mathbf{S}}$

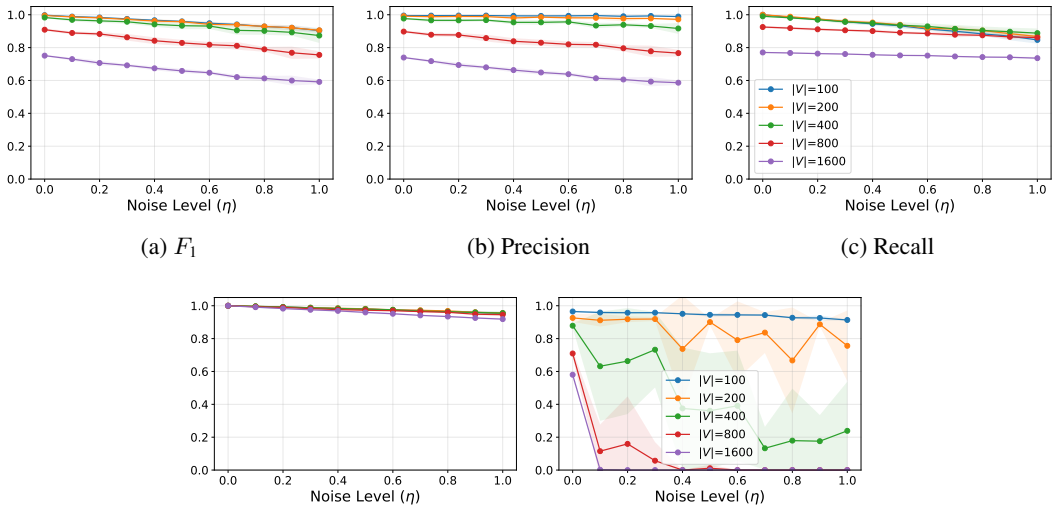
terms. Recall, however, remains high throughout, because the noise mechanism adds false positives without removing true positives. Instead, the decrease in $F_1$ is driven mostly by precision. By contrast, $d_{\text{inv}}$ falls off sharply under this noise, owing to its direct penalisation of excessive prediction cardinality, correctly reflecting that a model whose cardinality has inflated toward $|\mathcal{V}|/2$ retains little predictive value. The key point is not that one metric is right and the other wrong here, but that this regime is the *only* one in which AP behaves acceptably.

Indeed, this type of performance degradation is almost never relevant in real-world scenarios. HML classification is a difficult problem because natural imbalances arise at deeper nodes (Xu et al., 2026; Pereira et al., 2021), with detection and recall falling in these specific and rare regions. In contrast, higher-level nodes, being more frequently observed, are rarely confused with one another. Instead, a realistic noise environment sharply increases the rate of false negatives with node depth, and only slightly increases false positives, which decrease with depth. Following these guidelines, we simulate data by generating synthetic hierarchies as in Algorithm 1, with fictitious data generated




(a) False Negative (FN) Rate (b) False Positive (FP) Rate

Figure 6: **Random Noise Profile.** The behaviour of the random noise added is such that it slightly increases the FP rate at higher-level nodes, while substantially increasing the FN rate at deeper nodes, simulating a more realistic type of HML model degradation.




(a) $F_1$ (b) Precision (c) Recall

(d) AP (e) $d_{\text{inv}}$

Figure 7: **HML Metric Behaviours.** The top row displays the behaviour of node-wise $F_1$, and its components in the simulated noisy model experiment across hierarchy sizes $|\mathcal{V}|$ and added $\eta$. The bottom row compares AP to $d_{\text{inv}}$.

using Algorithm 2, and add noise with Algorithm 3. The effects of the added noise are shown in Figure 6. Our synthetic experiment targets an expected cardinality of three (to simulate the low cardinality and sparsity of benthic data). We tested five distinct hierarchies at each of five sizes (25 hierarchies in total), with 500 data points per hierarchy. The noise levels spanned 0.0 to 1.0 in 0.1 increments.

In this more realistic setting (Figure 7), we observe that AP is both insensitive to disruptive noise and misleading as to model quality, leading to an optimistic overestimation of model performance. This behaviour matches results in existing studies (Xu et al., 2024; Giunchiglia & Lukasiewicz, 2020; Pereira et al., 2021; Xu et al., 2026). AP behaves this way because of the sparsity of nodes in large hierarchies. Low ground-truth cardinality means that the vast majority of nodes are inactive for any given sample. Micro-average AP computes precision and recall across all node-sample pairs simultaneously, so by its inability to distinguish the trivial discrimination of inactive nodes from the genuinely difficult and important task of detecting active ones, the metric becomes inflated. On the other hand, $d_{\text{inv}}$ is not susceptible to this inflation because it operates directly on predicted and ground-truth cardinalities. Consequently, $d_{\text{inv}}$ degrades more steeply than both $F_1$ and AP as noise

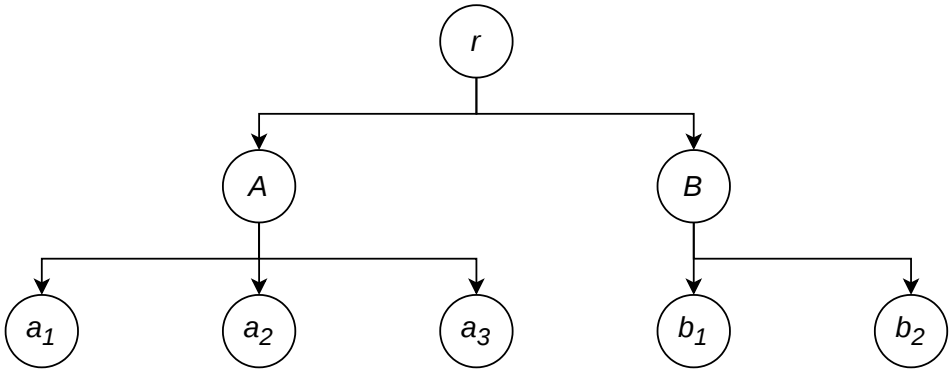



Figure 8: **Sample Hierarchy.** A toy hierarchy of eight nodes, intended as the subject for comparisons between $hF_1$ and $d_{\text{inv}}$.

intensifies, reflecting the practical reality that the model has lost its ability to resolve the fine-grained distinctions important to HML classification.

### I.1 Comparisons to Hierarchical $F_1$

We compare $d_{\text{inv}}$ and $hF_1$ (Lin et al., 2025) on two possible prediction strategies that we dub "sibling hedging" and "depth over-prediction". Throughout this section, we refer to the example hierarchy depicted in Figure 8. Consider a hierarchy with root node $r$, internal nodes $A$ and $B$ as children of $r$, and leaf nodes $a_1, a_2, a_3$ as children of $A$ and $b_1, b_2$ as children of $B$. The hierarchy therefore contains $|\mathcal{V}| = 8$ nodes. Suppose the ground truth annotation for a given sample is $\mathcal{Y} = \{a_1\}$, so $n_{\text{truth}} = 1$. The maximum tree distance from $a_1$ to any other node is $d_{\max} = 4$, attainable by $b_1$ and $b_2$. For reference, the perfect prediction $\hat{\mathcal{Y}} = \{a_1\}$ achieves $hF_1 = 1.0$ and $d_{\text{inv}} = 1.0$.

First, consider sibling hedging, where a risk-averse classifier predicts all siblings of the true label, producing $\hat{\mathcal{Y}} = \{a_1, a_2, a_3\}$ with $n_{\text{pred}} = 3$. To compute $hF_1$, each label is expanded into the set of itself and its ancestors. The expanded ground truth is $\mathcal{Y}^{\uparrow} = \{a_1, A\}$ (ignoring the root) and the expanded prediction is $\hat{\mathcal{Y}}^{\uparrow} = \{a_1, a_2, a_3, A\}$. Precision is $|\mathcal{Y}^{\uparrow} \cap \hat{\mathcal{Y}}^{\uparrow}| \;/\; |\hat{\mathcal{Y}}^{\uparrow}| = |\{a_1, A\}| \;/\; |\{a_1, a_2, a_3, A\}| = 2/4$, and recall is $|\mathcal{Y}^{\uparrow} \cap \hat{\mathcal{Y}}^{\uparrow}| \;/\; |\mathcal{Y}^{\uparrow}| = 2/2 = 1$. The harmonic mean gives $hF_1 = 2 \cdot (1/2) \cdot 1 \;/\; (1/2 + 1) = 1 \;/\; (3/2) = 2/3 = 0.67$. The ancestor overlap substantially attenuates the penalty for the two spurious predictions, because siblings share their full ancestral lineage with the true label. This leniency intensifies as hierarchy depth increases relative to the count of sibling nodes, since the shared ancestral path increasingly dominates the expanded sets.

To compute $d_{\text{inv}}$ for the same scenario, we first construct the score matrix $\mathbf{D}_{\text{inv}} \in \mathbb{R}^{n_{\text{truth}} \times n_{\text{pred}}} = \mathbb{R}^{1 \times 3}$, where each entry is the normalised similarity $(d_{\max} - d_{\text{pred}})/d_{\max}$. The tree distances from $a_1$ to each prediction are $d_{\mathcal{T}}(a_1, a_1) = 0$, $d_{\mathcal{T}}(a_1, a_2) = 2$, and $d_{\mathcal{T}}(a_1, a_3) = 2$, giving $\mathbf{D}_{\text{inv}} = [(4-0)/4 \quad (4-2)/4 \quad (4-2)/4] = [1.0 \quad 0.5 \quad 0.5]$. Solving the rectangular assignment problem yields $f_{\text{JV}}(\hat{\mathcal{Y}}, \mathcal{Y}; \mathbf{D}_{\text{inv}}) = 1.0$, so $s_{\max} = f_{\text{JV}}/n_{\text{truth}} = 1.0$, since the optimal match assigns predicted $a_1$ to true $a_1$ at a score of 1.0. For the cardinality factor, we have $n_{\text{E}} = n_{\text{pred}} - n_{\text{truth}} = 3 - 1 = 2$ excess predictions and $n_{\text{ME}} = |\mathcal{V}| - n_{\text{truth}} = 8 - 1 = 7$. Evaluating $\gamma(2;\, k_{\text{P}}{=}0.5,\, n_{\text{ME}}{=}7)$ numerically via Equation 4 gives $\gamma \approx 0.248$. The final score is therefore $d_{\text{inv}} = s_{\max} \cdot \gamma = 1.0 \cdot 0.248 \approx 0.25$. While $hF_1$ awards 0.67 to this prediction, $d_{\text{inv}}$ assigns only 0.25, reflecting the intuition that predicting all children when only one is correct constitutes a significant cardinality violation.

Now consider depth over-prediction, where the classifier predicts every leaf: $\hat{\mathcal{Y}} = \{a_1, a_2, a_3, b_1, b_2\}$ with $n_{\text{pred}} = 5$. For $hF_1$, the expanded prediction becomes $\hat{\mathcal{Y}}^{\uparrow} = \{a_1, a_2, a_3, b_1, b_2, A, B\}$, containing seven elements, while the expanded ground truth remains $\mathcal{Y}^{\uparrow} = \{a_1, A\}$. Precision is $|\{a_1, A\}| \;/\; |\{a_1, a_2, a_3, b_1, b_2, A, B\}| = 2/7$, recall is $2/2 = 1$, and $hF_1 = 2{\cdot}(2/7){\cdot}1 \;/\; (2/7{+}1) = (4/7) \;/\; (9/7) = 4/9 \approx 0.44$. Despite predicting every leaf, the metric remains well above zero because several nodes share intermediate ancestors.

Table 13: **Decomposition of the $d_{\text{inv}}$ Change under Mean Replacement.** BN-Bio, five seeds (Table 3), rescored with alternative cardinality factors on the same matched predictions; setting $\gamma \equiv 1$ leaves only the matched proximity $s_{\max}$. Per-node: nodes that degrade out of nodes evaluated, with $p$-value. Without the text signal, mean predicted endpoints rise from $10.16$ to $10.29$ (ground truth: $2.34$).

| score | clean | ablated | $\Delta$ | per-node |
|---|---|---|---|---|
| $d_{\text{inv}}$, $k_P = 0.5$ | 0.0391 | 0.0306 | $-\mathbf{21.7}\%$ | $214/248$, $1.6 \times 10^{-32}$ |
| min/max ratio | 0.2553 | 0.2452 | $-4.0\%$ | $193/248$, $1.1 \times 10^{-21}$ |
| $F_1$-style count penalty | 0.3695 | 0.3594 | $-2.7\%$ | $183/248$, $2.8 \times 10^{-16}$ |
| $hF_1$ | 0.3857 | 0.3811 | $-1.2\%$ | $158/246$, $4.7 \times 10^{-6}$ |
| proximity, $\gamma = 1$ | 0.8371 | 0.8408 | $+0.43\%$ | $60/246$, $p = 1.0$ |

For $d_{\text{inv}}$, the score matrix is $\mathbf{D}_{\text{inv}} \in \mathbb{R}^{1\times 5}$. The tree distances from $a_1$ are $d_{\mathcal{T}}(a_1, a_1) = 0$, $d_{\mathcal{T}}(a_1, a_2) = 2$, $d_{\mathcal{T}}(a_1, a_3) = 2$, $d_{\mathcal{T}}(a_1, b_1) = 4$, and $d_{\mathcal{T}}(a_1, b_2) = 4$, yielding $\mathbf{D}_{\text{inv}} = [1.0 \quad 0.5 \quad 0.5 \quad 0.0 \quad 0.0]$. The optimal assignment again matches predicted $a_1$ to true $a_1$, giving $f_{\text{JV}} = 1.0$ and hence $s_{\max} = f_{\text{JV}}/n_{\text{truth}} = 1.0$. Now $n_{\text{E}} = 5 - 1 = 4$ and $n_{\text{ME}} = 7$, so the cardinality factor is $\gamma(4;\, 0.5,\, 7) \approx 0.057$, and the final score is $d_{\text{inv}} = 1.0 \cdot 0.057 \approx 0.06$. Where $hF_1$ still awards $0.44$ to a degenerate prediction that labels every leaf in the hierarchy, $d_{\text{inv}}$ reduces the score to near zero, faithfully capturing that such a prediction carries almost no discriminative information.

The divergence between the two metrics in these scenarios stems from a fundamental design difference. Because $hF_1$ operates on ancestor-expanded sets without solving an explicit assignment between predicted and true labels, any overlap in ancestral paths contributes positively to both precision and recall, regardless of whether those shared ancestors reflect genuine predictive accuracy or are simply artefacts of the hierarchical structure. In shallow hierarchies with few shared ancestors this effect can be minor, but in deep or broad hierarchies it can substantially dampen the penalty for spurious predictions. By contrast, $d_{\text{inv}}$ isolates the best possible matching through the assignment problem and then applies the cardinality factor $\gamma$, which decays monotonically with the number of excess predictions. This ensures that cardinality violations are penalised in proportion to their severity, independently of where in the hierarchy the spurious predictions happen to fall. We view this property as particularly important for applications where downstream decisions depend on the precise set of predicted labels rather than merely on their neighbourhood in the hierarchy.

This leniency has been identified independently. Falis et al. (2021) observe the same structural property in set-based hierarchical evaluation, of which $hF_1$ is an instance. Because the expanded sets are unions, several predicted descendants of a common ancestor reduce to a single entry, and the measure does not consider how many predictions reach the ancestor. They address this by propagating descendant counts rather than truth values, and report that the resulting measure returns consistently lower scores than set-based evaluation across three trained models on the MIMIC-III ICD-9 task, attributing the gap to over- and under-prediction within families of codes that set-based evaluation cannot express.

Our own results are in agreement. In Table 13, the degradation observed under text ablation is registered most weakly by $hF_1$ among the cardinality-aware scores, and disappears entirely once $\gamma$ is fixed to one, indicating that the change being measured is one of prediction cardinality. That two measures constructed for different domains exhibit the same insensitivity suggests it follows from ancestor expansion itself rather than from any particular hierarchy or dataset.

### I.2 Decomposing the $d_{\text{INV}}$ Ablation

Removing the per-sample text signal lowers $d_{\text{inv}}$ on BN-Bio by 21.7% (Table 3), but this alone does not show which component of $d_{\text{inv}}$ is responsible. $d_{\text{inv}}$ is the product of a proximity term $s_{\max}$, the assignment-matched hierarchical similarity, and a cardinality factor $\gamma$, which penalises excess predictions. We therefore rescore the BN-Bio mean-replacement runs under alternative cardinality factors, holding the matched assignment fixed so that only the cardinality term changes (Table 13).

The degradation appears under every cardinality-aware score, including three that we did not introduce: a min/max ratio, an $F_1$-style count penalty, and $hF_1$, which registers it most weakly. Setting $\gamma \equiv 1$, which leaves proximity alone, removes the degradation, while the mean number of predicted endpoints rises from 10.16 to 10.29 against a ground-truth mean of 2.34. The change is therefore attributable to excess prediction cardinality rather than to matched proximity. This is the signature expected of a sub-tree prior. Since $s_{\max}$ scores only the best-matched prediction for each ground-truth endpoint, suppressing predictions in incorrect subtrees can only come through in $\gamma$. This concerns the fused model with and without per-sample text, and is distinct from the proximity advantage in subsection 4.2, which compares the text branch with the image branch.

The weak response of $hF_1$ matches subsection I.1. Because $hF_1$ expands each label to itself and its ancestors, spurious predictions that share a lineage with the target are only lightly penalised, so a change expressed through cardinality is precisely the kind to which $hF_1$ is least sensitive.

# J Additional Details for Inverted HML Score

This section contains edge case definitions and proofs for the main claims made in the introduction of the inverted HML score ($d_{\text{inv}}$) in section 3. These claims and proofs serve to demonstrate the mathematical behaviour and robustness of the cardinality component of $d_{\text{inv}}$.

## J.1 Edge Cases

Three edge cases for $d_{\text{inv}}$ are defined as follows. Where no path exists between a ground-truth endpoint and a predicted one, as is the case between disjoint roots, the pairwise similarity is set to zero rather than left undefined. Where the prediction set is empty, $d_{\text{inv}} = 0$. Where the prediction is deficient rather than excessive, $n_{\text{E}} \leq 0$ and $\gamma = 1$, so under-prediction is penalised instead in $s_{\max}$ through unmatched ground-truth components, rather than in the cardinality factor.

**Theorem 1** (Critical Value for the Trivial Root of $g$)**.** *Given $n_{ME} > 1$ and $k_P \in (0, 1)$, then at $k_P = k_P^* := \frac{n_{ME}-1}{n_{ME}}$, the only root of:*

$$g(b) = (k_P - 1)e^{-bn_{ME}} + e^{-b} - k_P, \tag{16}$$

*is the trivial solution $b = 0$.*

*Proof.* We begin by observing that $b = 0$ is always a root of $g$, since $g(0) = (k_{\text{P}} - 1) + 1 - k_{\text{P}} = 0$ for every $k_{\text{P}}$.

We then compute the first derivative of $g(b)$ at $b = 0$:

$$\frac{\mathrm{d}g}{\mathrm{d}b} = -n_{\text{ME}}(k_{\text{P}} - 1)e^{-bn_{\text{ME}}} - e^{-b} \tag{17}$$

$$\left(\frac{\mathrm{d}g}{\mathrm{d}b}\right)_{b=0} = -n_{\text{ME}}(k_{\text{P}} - 1) - 1 \tag{18}$$

Setting $\left(\frac{\mathrm{d}g}{\mathrm{d}b}\right)_{b=0} = 0$, we have:

$$k_{\text{P}}^* = \frac{n_{\text{ME}} - 1}{n_{\text{ME}}}, \tag{19}$$

as the form $k_{\text{P}}^*$ must take for $b = 0$ to be a critical point of $g$. Plugging this back into $\frac{\mathrm{d}g}{\mathrm{d}b}$, we can observe the behaviour of the function to the left and right of $b = 0$:

$$\frac{\mathrm{d}g}{\mathrm{d}b} = n_{\text{ME}}\frac{1}{n_{\text{ME}}}e^{-bn_{\text{ME}}} - e^{-b} \tag{20}$$

$$= e^{-bn_{\text{ME}}} - e^{-b}. \tag{21}$$

Since $n_{\text{ME}} > 1$, for $b > 0$ we have $bn_{\text{ME}} > b$ and hence $e^{-bn_{\text{ME}}} < e^{-b}$, giving $\frac{\mathrm{d}g}{\mathrm{d}b} < 0$. Symmetrically, for $b < 0$ we have $bn_{\text{ME}} < b$ and hence $e^{-bn_{\text{ME}}} > e^{-b}$, giving $\frac{\mathrm{d}g}{\mathrm{d}b} > 0$. This means that for

$k_{\mathrm{P}}^*$, $b = 0$ is a maximum, and the function monotonically increases on its left to meet this value and decreases monotonically on its right.

For completeness, we can examine the bounds of the function. It is observable from the form presented in Equation 16 that:

$$\lim_{b\to\infty} g(b) = 0 + 0 - k_{\mathrm{P}}^* = -k_{\mathrm{P}}^* < 0. \tag{22}$$

For the other direction, we factor $g$ as:

$$g(b) = e^{-b}\big[(k_{\mathrm{P}} - 1)e^{-b(n_{\mathrm{ME}}-1)} + 1\big] - k_{\mathrm{P}}. \tag{23}$$

As $b \to -\infty$, the prefactor $e^{-b} \to +\infty$. Inside the bracket, $e^{-b(n_{\mathrm{ME}}-1)} \to +\infty$ (since $n_{\mathrm{ME}} - 1 > 0$), and its coefficient $(k_{\mathrm{P}} - 1) < 0$ drives the bracket to $-\infty$. The product of $e^{-b} \to +\infty$ with a bracket tending to $-\infty$ gives:

$$\lim_{b\to-\infty} g(b) = -\infty. \tag{24}$$

Therefore, for $k_{\mathrm{P}}^*$, $b = 0$ is the global maximum with $g(0) = 0$, so $g(b) \leq 0$ for all $b \in \mathbb{R}$ with equality only at $b = 0$. Thus, $b = 0$ is the only root of $g(b)$. □

**Theorem 2.** *At $k_P = k_P^*$, the exponential decay function $\gamma(n_E; k_P, n_{ME}) = ae^{-bn_E} + c$ reduces to a linear function of $n_E$.*

*Proof.* From Theorem 1, at $k_{\mathrm{P}} = k_{\mathrm{P}}^*$ the only solution is $b = 0$, so it suffices to evaluate $\lim_{b\to 0} \gamma(n_{\mathrm{E}}; k_{\mathrm{P}}, n_{\mathrm{ME}})$.

Returning to the exponential decay function and plugging in the expressions for $a$ and $c$ derived earlier, namely $a = (1 - e^{-bn_{\mathrm{ME}}})^{-1}$ and $c = 1 - a$:

$$\gamma(n_{\mathrm{E}}; k_{\mathrm{P}}, n_{\mathrm{ME}}) = \frac{e^{-bn_{\mathrm{E}}}}{1 - e^{-bn_{\mathrm{ME}}}} + 1 - \frac{1}{1 - e^{-bn_{\mathrm{ME}}}} \tag{25}$$

$$= 1 - \frac{1 - e^{-bn_{\mathrm{E}}}}{1 - e^{-bn_{\mathrm{ME}}}}. \tag{26}$$

Now, we can perform a Taylor expansion of the exponential terms around $b = 0$:

$$\gamma(n_{\mathrm{E}}; k_{\mathrm{P}}, n_{\mathrm{ME}}) = 1 - \frac{1 - \big(1 - bn_{\mathrm{E}} + \mathcal{O}(b^2)\big)}{1 - \big(1 - bn_{\mathrm{ME}} + \mathcal{O}(b^2)\big)} \tag{27}$$

$$= 1 - \frac{bn_{\mathrm{E}} + \mathcal{O}(b^2)}{bn_{\mathrm{ME}} + \mathcal{O}(b^2)} \tag{28}$$

$$= 1 - \frac{n_{\mathrm{E}} + \mathcal{O}(b)}{n_{\mathrm{ME}} + \mathcal{O}(b)}. \tag{29}$$

Therefore, we have:

$$\lim_{b\to 0} \gamma(n_{\mathrm{E}}; k_{\mathrm{P}}, n_{\mathrm{ME}}) = 1 - \frac{n_{\mathrm{E}}}{n_{\mathrm{ME}}}, \tag{30}$$

which is linear in $n_{\mathrm{E}}$. As a final note, plugging in $n_{\mathrm{E}} = 1$ into this linear form gives $\frac{n_{\mathrm{ME}}-1}{n_{\mathrm{ME}}}$, which is the form for $k_{\mathrm{P}}^*$ derived in Equation 19, providing a consistency check. □

**Theorem 3** (Non-Trivial Existence of $b$)**.** *For a given $n_{ME} > 1$ and $k_P \in (0, 1) \setminus \{k_P^*\}$, there exists for $g(b)$ a unique non-trivial root. Moreover, this root is positive when $k_P < k_P^*$ and negative when $k_P > k_P^*$.*

*Proof.* We return to our derivative for $g(b)$ from Equation 17, which we restate and set equal to zero:

$$-n_{\mathrm{ME}}(k_{\mathrm{P}} - 1)e^{-bn_{\mathrm{ME}}} - e^{-b} = 0. \tag{31}$$

In this form, we are now able to isolate for $b$, attaining:

$$-n_{\mathrm{ME}}(k_{\mathrm{P}} - 1)e^{-bn_{\mathrm{ME}}} = e^{-b} \tag{32}$$

$$\implies n_{\mathrm{ME}}(1 - k_{\mathrm{P}}) = e^{b(n_{\mathrm{ME}}-1)} \tag{33}$$

$$\therefore b_{\mathrm{crit}} = \frac{\ln\big(n_{\mathrm{ME}}(1 - k_{\mathrm{P}})\big)}{n_{\mathrm{ME}} - 1}. \tag{34}$$

We note that the exponential term in the intermediate Equation 33 is strictly monotonically increasing in the domain $(-\infty, \infty)$; therefore, $b_{\text{crit}}$ must be the unique intersection with the constant on the left-hand side, and hence the unique critical point of $g$. As a quick sanity check, we can observe that $b_{\text{crit}} \to 0$ as $k_{\text{P}} \to k_{\text{P}}^*$.

Now, plugging in $b_{\text{crit}}$ in for the second derivative of $g(b)$, we get:

$$\left(\frac{\mathrm{d}^2 g}{\mathrm{d}b^2}\right)_{b=b_{\text{crit}}} = n_{\text{ME}}^2(k_{\text{P}}-1)e^{-b_{\text{crit}}n_{\text{ME}}} + e^{-b_{\text{crit}}} \tag{35}$$

$$= n_{\text{ME}}^2(k_{\text{P}}-1)\big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}} + \big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{1}{n_{\text{ME}}-1}} \tag{36}$$

$$= \big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}\big(n_{\text{ME}}(1-k_{\text{P}}) + n_{\text{ME}}^2(k_{\text{P}}-1)\big) \tag{37}$$

$$= \big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}\big(n_{\text{ME}}(1-k_{\text{P}}) - n_{\text{ME}}^2(1-k_{\text{P}})\big) \tag{38}$$

$$= \big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}} \cdot (1-k_{\text{P}})\, n_{\text{ME}}(1-n_{\text{ME}}). \tag{39}$$

We see that all the terms above are positive except for $1 - n_{\text{ME}}$, which confirms:

$$\left(\frac{\mathrm{d}^2 g}{\mathrm{d}b^2}\right)_{b=b_{\text{crit}}} < 0, \tag{40}$$

and $b_{\text{crit}}$ as a maximum. Since $b_{\text{crit}}$ is also the only critical point and $g(b)$ is continuous, it is therefore the global maximum.

We can now plug $b_{\text{crit}}$ back into the original $g(b)$:

$$g(b_{\text{crit}}) = (k_{\text{P}}-1)e^{-b_{\text{crit}}n_{\text{ME}}} + e^{-b_{\text{crit}}} - k_{\text{P}} \tag{41}$$

$$= (k_{\text{P}}-1)\big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}} + \big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{1}{n_{\text{ME}}-1}} - k_{\text{P}} \tag{42}$$

$$= \big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}\big(n_{\text{ME}}(1-k_{\text{P}}) - (1-k_{\text{P}})\big) - k_{\text{P}} \tag{43}$$

$$= \big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}(1-k_{\text{P}})(n_{\text{ME}}-1) - k_{\text{P}}. \tag{44}$$

When we insert $k_{\text{P}}^*$ for $k_{\text{P}}$, keeping in mind that $1 - k_{\text{P}}^* = \frac{1}{n_{\text{ME}}}$, it becomes clear that the expression becomes zero.

We therefore proceed to look at the expression and $b_{\text{crit}}$ itself in the cases where $k_{\text{P}} < k_{\text{P}}^*$ and $k_{\text{P}} > k_{\text{P}}^*$. We rewrite the expression in Equation 44 as:

$$g(b_{\text{crit}}) = n_{\text{ME}}^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}(1-k_{\text{P}})^{1-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}(n_{\text{ME}}-1) - k_{\text{P}} \tag{45}$$

$$= n_{\text{ME}}^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}(1-k_{\text{P}})^{-\frac{1}{n_{\text{ME}}-1}}(n_{\text{ME}}-1) - k_{\text{P}}. \tag{46}$$

Now, to determine how $g(b_{\text{crit}})$ behaves as $k_{\text{P}}$ changes, we take the derivative of $g(b_{\text{crit}})$ with respect to $k_{\text{P}}$[4]:

$$\frac{\mathrm{d}g(b_{\text{crit}})}{\mathrm{d}k_{\text{P}}} = -\frac{1}{n_{\text{ME}}-1}(n_{\text{ME}}-1)n_{\text{ME}}^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}(1-k_{\text{P}})^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}(-1) - 1 \tag{47}$$

$$= n_{\text{ME}}^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}}(1-k_{\text{P}})^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}} - 1 \tag{48}$$

$$= \big(n_{\text{ME}}(1-k_{\text{P}})\big)^{-\frac{n_{\text{ME}}}{n_{\text{ME}}-1}} - 1. \tag{49}$$

We can now attain the following statements:

$$\frac{\mathrm{d}g(b_{\text{crit}})}{\mathrm{d}k_{\text{P}}} = 0 \quad \Longleftrightarrow \quad n_{\text{ME}}(1-k_{\text{P}}) = 1 \quad \Longleftrightarrow \quad k_{\text{P}} = k_{\text{P}}^* \tag{50}$$

$$\frac{\mathrm{d}g(b_{\text{crit}})}{\mathrm{d}k_{\text{P}}} < 0 \quad \Longleftrightarrow \quad n_{\text{ME}}(1-k_{\text{P}}) > 1 \quad \Longleftrightarrow \quad k_{\text{P}} < k_{\text{P}}^* \tag{51}$$

$$\frac{\mathrm{d}g(b_{\text{crit}})}{\mathrm{d}k_{\text{P}}} > 0 \quad \Longleftrightarrow \quad n_{\text{ME}}(1-k_{\text{P}}) < 1 \quad \Longleftrightarrow \quad k_{\text{P}} > k_{\text{P}}^*. \tag{52}$$

[4]Since $b_{\text{crit}}$ has been expressed in terms of $k_{\text{P}}$ alone, this is a total derivative.

It follows that $g(b_{\text{crit}})$ attains a global minimum at $k_{\text{P}}^*$, where it equals zero: to the left of $k_{\text{P}}^*$, the function is descending towards the minimum, while to the right, it is ascending away from it. Therefore, $g(b_{\text{crit}}) > 0$ for all $k_{\text{P}} \in (0,1) \setminus \{k_{\text{P}}^*\}$.

Next, returning to Equation 34, we see that:

$$k_{\text{P}} < k_{\text{P}}^* \implies 1 - k_{\text{P}} > 1/n_{\text{ME}} \tag{53}$$

$$\implies n_{\text{ME}}(1 - k_{\text{P}}) > 1 \tag{54}$$

$$\implies \ln\left(n_{\text{ME}}(1 - k_{\text{P}})\right) > 0 \tag{55}$$

$$\therefore b_{\text{crit}} > 0. \tag{56}$$

If we apply the same logic to the case where $k_{\text{P}} > k_{\text{P}}^*$, we get the opposite result, where $b_{\text{crit}} < 0$.

To summarise, we now have the following relevant information:

1. $b_{\text{crit}}$ is the unique critical point of $g(b)$ and a global maximum, since $g(b)$ is continuous with no other critical points;
2. $b_{\text{crit}}$ is positive for $k_{\text{P}} < k_{\text{P}}^*$, negative for $k_{\text{P}} > k_{\text{P}}^*$, and zero for $k_{\text{P}} = k_{\text{P}}^*$;
3. $g(b_{\text{crit}}) > 0$ for all $k_{\text{P}} \in (0,1) \setminus \{k_{\text{P}}^*\}$, and zero at $k_{\text{P}} = k_{\text{P}}^*$;
4. $g(0) = 0$ for all $k_{\text{P}}$; and
5. $\lim_{b \to -\infty} g(b) = -\infty$ and $\lim_{b \to \infty} g(b) = -k_{\text{P}} < 0$, by the same argument as in Theorem 1 (which holds for all $k_{\text{P}} \in (0,1)$).

Since $b_{\text{crit}}$ is the only critical point, $g$ is strictly increasing on $(-\infty, b_{\text{crit}})$ and strictly decreasing on $(b_{\text{crit}}, \infty)$. In particular, $g$ can cross zero at most once on each of these two intervals.

In the case where $k_{\text{P}} < k_{\text{P}}^*$, we have $b_{\text{crit}} > 0$, so the trivial root $b = 0$ lies in the increasing interval. Since $g$ is strictly increasing on $(-\infty, b_{\text{crit}})$ and $g(0) = 0$, it follows that $g(b) < 0$ for all $b < 0$. On the decreasing interval $(b_{\text{crit}}, \infty)$, $g$ falls from $g(b_{\text{crit}}) > 0$ to $-k_{\text{P}} < 0$, so by the intermediate value theorem and strict monotonicity, there is exactly one root, which is positive.

In the case where $k_{\text{P}} > k_{\text{P}}^*$, we have $b_{\text{crit}} < 0$, so the trivial root $b = 0$ lies in the decreasing interval. Since $g$ is strictly decreasing on $(b_{\text{crit}}, \infty)$ and $g(0) = 0$, it follows that $g(b) < 0$ for all $b > 0$. On the increasing interval $(-\infty, b_{\text{crit}})$, $g$ rises from $-\infty$ to $g(b_{\text{crit}}) > 0$, so by the intermediate value theorem and strict monotonicity, there is exactly one root, which is negative.

Lastly, in the case $k_{\text{P}} = k_{\text{P}}^*$, $b_{\text{crit}} = 0$ and $g(b_{\text{crit}}) = 0$, so only the trivial root exists, recovering the result of Theorem 1. □