HITAG: HIERARCHICAL IMAGE TAGGING WITH HYPERBOLIC VISION-LANGUAGE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image tagging, also known as multi-label image recognition, aims to assign multiple semantic labels to a given image. However, few benchmarks have been tailored for *hierarchical image tagging* to measure the hierarchical classification accuracy, where a concept 'Shiba Inu' should be recognized as both 'dog' and 'animal'. To explicitly capture such hierarchy, we introduce a hierarchical image tagging benchmark, termed HiTag, to evaluate the multi-visual context from a hierarchical perspective. Specifically, we first construct a tree-like hierarchical structure for the tags based on lexical semantic databases, *i.e.,* WordNet and YAGO, including *10 levels* and *3,334 labels*. The hierarchy is consistent with visual perception through optimization by a large model and can be dynamically updated for unexplored tags by locating their positions in WordNet and assessing their validity using a large model. With the designed hierarchical structure, we utilized large language models to annotate 2,872,012 images from CC3M as training data and manually tagged 57,223 images from OpenImage as test data, to advance the exploration of the hierarchical image tagging task. Meanwhile, we develop a pipeline to assess the hierarchical classification capacity of models on multiple levels, including tree edit distance, Jaccard similarity, hierarchical precision, hierarchical recall metrics, *etc.* Furthermore, we embed hierarchical tags, images, and captions into hyperbolic space for modeling, leveraging its inherent suitability for representing tree-structured data. Experimental results on the HiTag confirm that our method not only demonstrates superior performance of zero-shot image tagging, but also achieves state-of-the-art results on hierarchical image tagging modeling. We will release the code and the dataset to support the community.

1 INTRODUCTION

Image recognition, as a fundamental task in computer vision, has made remarkable progress in recent years. This advancement has been driven by the large-scale classification datasets, such as ImageNet Deng et al. (2009), which enable well-trained models to not only recognize the most salient object in an image but also provide robust representations for diverse downstream visual tasks Ren et al. (2016); He et al. (2018a); Long et al. (2015). However, individual images in real world typically contain multiple inherent objects, posing significant challenges to traditional single-label classification methods. To bridge this gap, multi-label image recognition, also known as image tagging, has garnered increasing attention. Concretely, recent efforts such as the RAM series Huang et al. (2023a); Zhang et al. (2024); Huang et al. (2023b) leverage large-scale image-text datasets (*e.g.,* CC3M Sharma et al. (2018), CC12M Changpinyo et al. (2021)) to achieve promising results in predicting multiple fine-grained tags for images. Although the existing image tagging methods can predict most objects in an image, as shown in Figure 1, they ignore the hierarchical relationships that naturally exist among concepts. For example, a 'dog' contained in an image also should be classified as the higher-level concept of 'animal'. In such scenarios, a well-trained and robust classification model should treat categories at different abstraction levels.

In biology, the taxonomy system Ruggiero et al. (2015); Linnaeus (2024) defines and classifies groups of biological organisms based on shared characteristics. Meanwhile, WordNet Fellbaum (1998) from linguistics is a lexical database that links words into semantic relations. The mentioned ImageNet is an image database organized according to WordNet, in which images are depicted by a single label. Inspired by this, we construct a tree-like hierarchical structure according to all WordNet nouns to

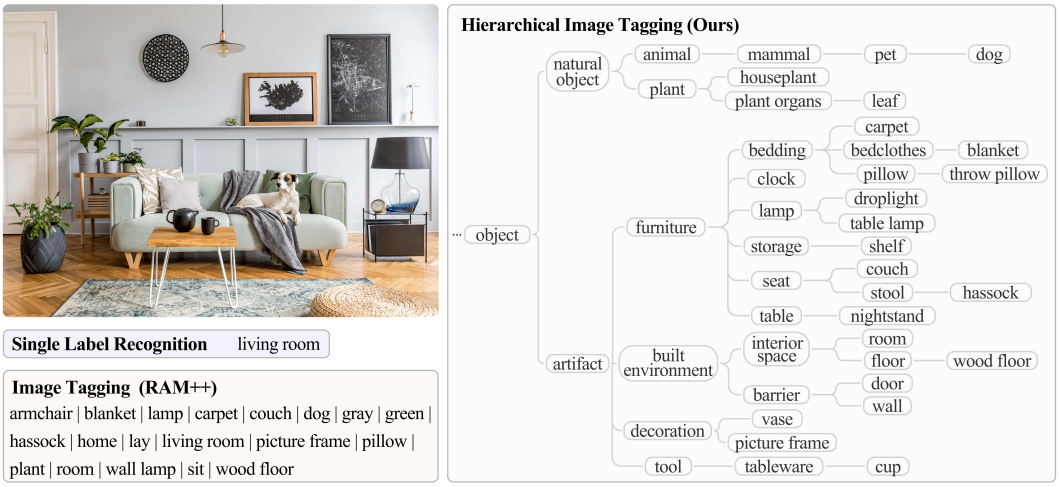

Figure 1: Comparison of Image Recognition tasks. Differing from tasks that only identify tags in an image, the hierarchical image tagging task explicitly models the hierarchical relationships among recognized tags, thereby producing a hierarchical tag tree.

organize the tags in images, providing a reasonable foundation for the hierarchical image tagging task. Although leveraging WordNet's extensive noun coverage has enabled the construction of a relatively complete tag tree, numerous common visual objects, such as '3D glasses' and 'auto showroom', are still absent. To overcome this limitation, we expand the hierarchical tag tree by incorporating new categories from the original tag vocabulary. Moreover, as WordNet is constructed based on psycholinguistic and computational linguistic principles, it contains numerous abstract concepts that are visually unobservable (*e.g.,* 'things' and 'use'). Therefore, we exclude these abstract categories to ensure visual interpretability. Furthermore, to better align the architecture with visual perception and user preferences, we leverage both a large language model (LLM) and YAGO to adjust and refine the structure. The final hierarchical structure of tags contains over 3,334 concepts and 10 levels. Besides, we can expand our hierarchy with unexplored tags by locating their positions in WordNet and using a large model to verify the rationality of the structure and its consistency with visual perception.

To effectively evaluate hierarchical image tagging, we construct a benchmark and develop a pipeline to assess the hierarchical classification performance. Specifically, based on the hierarchical tags structure, we create a benchmark dataset through manually labeling 57,223 images of OpenImage Kuznetsova et al. (2020), where each image is annotated not only with multiple tags but also with the tree-like hierarchical format. During evaluation, we adopt four metrics to evaluate hierarchical tagging performance: Tree Edit Distance, Jaccard Similarity, Hierarchical Precision, and Hierarchical Recall, which jointly measure the completeness and accuracy of predicted hierarchical tags. Moreover, we also use a large language model to label 2,872,012 images from CC3M as training data.

Furthermore, we propose a hierarchical image tagging method that represents, aligns, and models the hierarchical relationships among images, captions, and hierarchical labels in hyperbolic space. Unlike Euclidean embeddings, hyperbolic geometry naturally encodes hierarchical structures due to its exponential volume growth, making it well-suited for representing large-scale hierarchical taxonomies. First, we represent image, tag, and caption features into hyperbolic space and enrich each tag with multi-perspective textual descriptions according to their Lorentz inner product with the images to better align its semantics with visual content. Then, we align tag and image representation using a hyperbolic text-image attention module to extract the visual evidence related to the labels. More importantly, we model the hierarchical structures by leveraging entailment cones in Lorentz space to describe the partial order relationships within the hierarchy. Finally, we achieve state-of-the-art hierarchical tag recognition performance compared with both open-set image tagging methods and hierarchical image classification techniques.

Our main contributions are as follows:

- We construct a comprehensive hierarchical tag structure by integrating WordNet, YAGO, and LLM-based refinement, resulting in a tag tree covering 3,334 tags across 10 levels.
- We develop a hierarchical image tagging benchmark with 3M training and 57,223 test images and an evaluation pipeline to assess the hierarchical classification capacity with four metrics: tree edit distance, Jaccard similarity, hierarchical precision, and hierarchical recall.

• We propose a hierarchical image tagging method in hyperbolic space, which models hierarchical relationships among tags, images, and captions.

2 Related Work

2.1 Image Multi-label Recognition

Image multi-label recognition can be divided into two categories: (i) *top-k prediction*, which returns the k most probable labels; and (ii) *comprehensive tagging*, which aims to recognize *all* labels present in an image. Our work targets the second.

Top-k Prediction. Early works Wei et al. (2016); Wang et al. (2016); He et al. (2018b); Chen et al. (2019); Luo et al. (2019) formulated multi-label recognition task as closed-set recognition over a fixed set of label indices. To move beyond a fixed label set, subsequent methods Ren et al. (2015); Zhang et al. (2016); Lee et al. (2018); Rahman et al. (2020); Huynh & Elhamifar (2020); Narayan et al. (2021); Ben-Cohen et al. (2021) projected images into a word-semantic space using embeddings such as Word2Vec Mikolov et al. (2013) and GloVe Pennington et al. (2014), thereby enabling zero-shot generalization to unseen labels. These approaches, however, are fundamentally constrained by the finite vocabularies and static semantics of the word models. Recent advances in vision–language models (VLMs) like CLIP Radford et al. (2021) and ALIGN Jia et al. (2021) have significantly enhanced open-vocabulary recognition. Subsequent works Sun et al. (2022); Dao et al. (2023); He et al. (2023); Hu et al. (2024); Liu et al. (2024) further extend these capabilities. Nevertheless, any fixed k inevitably omit valid labels or introduce erroneous ones in complex visual scenes.

Comprehensive Tagging. Tag2Text Huang et al. (2023b) achieves strong tagging performance by jointly learning tagging and captioning on large-scale data, but it remains confined to a fixed and predefined label set. Building on this, RAM Zhang et al. (2024) leverages CLIP text encoder to convert tags into semantic label queries, enabling the recognition of unseen categories. RAM++ Huang et al. (2023a) further incorporates large language model (LLM)-generated visual descriptions for each label, boosting open-set generalization. Despite these advances, existing methods remain restricted to flat label sets, failing to capture the inherent hierarchical relationships among labels.

2.2 Hierarchical Representation Learning

Hierarchical representation learning relies on a predefined hierarchy. Current methods and sources for constructing such hierarchies include WordNet Fellbaum (1998) for general classification Bertinetto et al. (2020); Santurkar et al. (2020), the ATUS taxonomy of Labor Statistics (2003-2023) for activity classification Long et al. (2020), biological taxonomy for biological classification Van Horn et al. (2018); Chang et al. (2021); Van Horn et al. (2015), OpenStreetMap for street scene Neuhold et al. (2017), and some hierarchies Maji et al. (2013) that are constructed through manual inspection.

Based on the constructed hierarchy, hierarchical representation learning have three main branches: probability-constraint methods, multi-granular prediction heads, and embedding-based modeling.

Probability-constraint methods Bertinetto et al. (2020); Garg et al. (2022b); Li et al. (2022b) either directly enforce that the predicted probability of a parent category is never lower than that of any of its children Li et al. (2022b), or introduce hierarchical losses such as Hierarchical Cross-Entropy (HXE) Bertinetto et al. (2020) that implicitly preserve the same constraint. Although these techniques guarantee local parent–child consistency, they hard to capture the global hierarchy of the entire label tree. Methods with multi-granular prediction heads Wang et al. (2020); Chang et al. (2021); Garg et al. (2022a); Wang et al. (2023a); Park et al. (2025) construct the output layers of the network according to the branches of the tree hierarchy. Specifically, the output layers of different branches in the methodsChen et al. (2022); Giunchiglia & Lukasiewicz (2020); Zhou et al. (2023) are trained by imposing constraints, such as parent–child probability order, mutual exclusivity at the same level, and parent–child inclusion relationships to learn the probability differences among categories within each branch. Therefore, such methods essentially learn the relationships between index-based probability outputs rather than understanding the semantic relationships between category names. Embedding-based methods Zhang et al. (2022a); Wang et al. (2023b) embed images and hierarchical labels into a shared Euclidean semantic space, comprehensively modeling the tree structure and preserving cross-level relationships (e.g., HiMulConE Zhang et al. (2022a) spreads coarse-grained labels apart while tightly clustering fine-grained labels around their parent).However, embedding

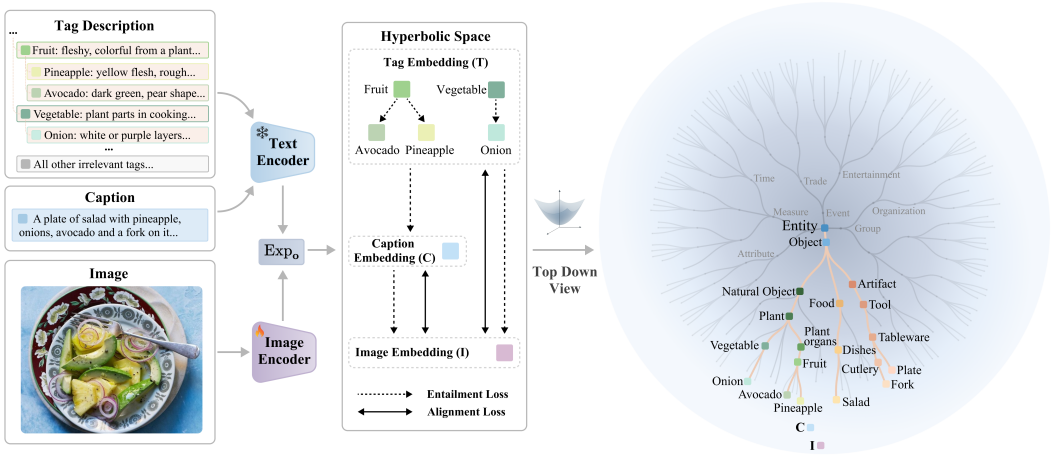

Figure 2: An overview of HiTag. In our model, we use a frozen text encoder and a trainable image encoder to encode text and image separately. We then map these embeddings into hyperbolic space to model hierarchical structures and align visual–textual semantics. As a result, our approach captures the hierarchical relationships among tags, captions, and images, achieving hierarchical image tagging.

exponentially growing tree structures in Euclidean space leads to severe distortion at greater depths, necessitating a more effective hierarchical embedding space.

2.3 Hyperbolic Geometry for Hierarchical Modeling

Compared with Euclidean space, hyperbolic geometry Ratcliffe (2006); Lee (2019); Peng et al. (2022); Cannon et al. (1997) captures exponentially expanding tree-like hierarchies more naturally Krioukov et al. (2010); Sarkar (2011); Nickel & Kiela (2017); Chamberlain et al. (2017). Leveraging this advantage, researchers have begun to redesign core neural-network components in hyperbolic space. The Hyperbolic Neural Networks (HNN) framework Ganea et al. (2018b) first introduced hyperbolic version of fully connected layers, multinomial logistic regression, and recurrent units; subsequent work extended these components to include CNN Shimizu et al. (2021), GNN Liu et al. (2019); Chami et al. (2019); Zhang et al. (2022b), attention Gulcehre et al. (2019); Zhang et al. (2022b), and VAE Mathieu et al. (2019); Skopek et al. (2020). These foundational efforts enabled hyperbolic geometry's adoption in computer vision.

Building on these foundational efforts, hyperbolic embeddings have driven improvements across diverse computer-vision tasks. Across recognition tasks, hyperbolic embeddings better capture label hierarchies, improving performance in single-label classification Khrulkov et al. (2020); Fang et al. (2021); Gao et al. (2021); Xu et al. (2023); Liu et al. (2020); Dengxiong & Kong (2023), image retrieval Khrulkov et al. (2020), person re-identification Khrulkov et al. (2020); Fang et al. (2021), and object detection Kong et al. (2024). For segmentation, capturing hierarchical part–whole relations yields sharper boundaries and more coherent masks (Atigh et al., 2022; Chen et al., 2024; Franco et al., 2024). In vision–language models, MERU Desai et al. (2023) and HyCoCLIP Pal et al. (2024) align visual and textual features in hyperbolic space to represent hierarchical relations across modalities utilizing the entailment loss Ganea et al. (2018a); Le et al. (2019). Despite recent advances, vision–language models that have not been applied to the task of hierarchical multi-tag recognition in hyperbolic space. Therefore, we perform open-set tagging using a hyperbolic vision–language model.

3 Method

We propose a hierarchical image tagging method that represents and models the hierarchical relationships among images, captions, and tags in hyperbolic space. As show in Figure 2, we first encode image feature and text features (tag descriptions and captions) using the image encoder and text encoder, respectively. Then, we project these features into the Lorentz model of hyperbolic space. Finally, we align between texts and images for tag recognition and utilize entailment cones in Lorentz space to model entailment relations in the hierarchy, enabling hierarchical image label recognition.

3.1 Preliminary

Hyperbolic space is a non-Euclidean manifold with constant negative curvature. In contrast to Euclidean space, the volume of a hyperbolic space grows exponentially with respect to its radius, making it well-suited for hierarchical structures. The Lorentz model is particularly favorable for deep

learning in hyperbolic space due to its numerical stability in computing exponential and logarithmic maps. In this work, we adopt the Lorentz model for hyperbolic space representation.

The Lorentz model The Lorentz model L^n is embedded in an $(n+1)$-dimensional Minkowski spacetime $\mathbb{R}^{n+1}$, and can be viewed as the upper sheet of a two-sheeted hyperboloid. Let $\mathbf{x} = (x_0, x_1, \ldots, x_n)$ be a point in L^n, where x_0 is the time component and $x_1, \ldots, x_n$ are spatial components. A Lorentz manifold with constant curvature $-k$ ($k > 0$) is defined as:

$$L^n = \Big\{ \mathbf{x} \in \mathbb{R}^{n+1} \mid \langle \mathbf{x}, \mathbf{x} \rangle_L = -\tfrac{1}{k},\ x_0 > 0 \Big\}, \tag{1}$$

where $\langle \cdot, \cdot \rangle_L$ denotes the Lorentzian inner product. For any $\mathbf{x}, \mathbf{y} \in L^n$,

$$\langle \mathbf{x}, \mathbf{y} \rangle_L = -x_0 y_0 + \langle \mathbf{x}, \mathbf{y} \rangle_E, \tag{2}$$

where $\langle \mathbf{x}, \mathbf{y} \rangle_E$ is the Euclidean inner product. The Lorentz norm is defined by $\|\mathbf{x}\|_L = \sqrt{\big|\langle \mathbf{x}, \mathbf{x} \rangle_L\big|}$.

The exponential and logarithmic maps For any $\mathbf{x} \in L^n$, its tangent space is given by $T_{\mathbf{x}} L^n = \big\{ \mathbf{v} \in \mathbb{R}^{n+1} \mid \langle \mathbf{x}, \mathbf{v} \rangle_L = 0 \big\}$. The exponential map, projecting from the tangent space to the hyperbolic manifold, is given by $\mathrm{Exp}_{\mathbf{x}}(\mathbf{v}) = \cosh \alpha\, \mathbf{x} + \frac{\sinh \alpha}{\alpha} \mathbf{v}$, where $\alpha = \sqrt{k}\, \|\mathbf{v}\|_L$. The Euclidean space can be regarded as the tangent space $T_{\mathbf{o}} L^n = \big\{ (0, \mathbf{u}) \mid \mathbf{u} \in \mathbb{R}^n \big\}$ of the Lorentz manifold at the origin $\mathbf{o} = \big(\frac{1}{\sqrt{k}}, 0, \ldots, 0\big)$. The exponential map $\mathrm{Exp}_{\mathbf{o}}$ from Euclidean to Lorentz manifold denotes:

$$\mathrm{Exp}_{\mathbf{o}}(\mathbf{u}) = \big(\cosh(\sqrt{k}\|\mathbf{u}\|),\ \sinh(\sqrt{k}\|\mathbf{u}\|)\, \tfrac{\mathbf{u}}{\|\mathbf{u}\|} \big). \tag{3}$$

Conversely, the logarithmic map projects points from the Lorentz manifold to the Euclidean space:

$$\mathrm{Log}_{\mathbf{o}}(\mathbf{x}) = \frac{1}{\sqrt{k}}\ \mathrm{arcosh}(k^{-\frac{1}{2}} x_0)\ \frac{(x_1, \ldots, x_n)}{\sqrt{x_0^2 - k^{-1}}} \in \mathbb{R}^n, \tag{4}$$

Note that exponential and logarithmic maps are local inverses, enabling seamless transformations between the Lorentz space and Euclidean space.

3.2 Hyperbolic Feature Representation

As illustrated in Figure 2, we first encode images and texts with an image encoder f_{img} and a text encoder f_{txt} respectively, yielding n-dimensional feature vectors in Euclidean space. Then, we utilize the exponential mapping $\mathrm{Exp}_{\mathbf{o}}$ (Eq equation 3) to project these feature vectors onto the Lorentz hyperbolic surface. Consequently, the hyperbolic representations of the image I, caption C, and tag T are denoted as $\mathbf{x}^{(i)} = \mathrm{Exp}_{\mathbf{o}}\big(f_{\mathrm{img}}(I)\big)$, $\mathbf{x}^{(\mathrm{cap})} = \mathrm{Exp}_{\mathbf{o}}\big(f_{\mathrm{txt}}(C)\big)$, $\mathbf{x}^{(\mathrm{tag})} = \mathrm{Exp}_{\mathbf{o}}\big(f_{\mathrm{txt}}(T)\big)$.

To fully capture the semantics of each tag from different perspectives, we employ a multi-modal large language model (MLLM), Qwen2.5-vl-72B Bai et al. (2025b), to generate h diverse descriptions for each tag. These tag descriptions are likewise mapped into the same Lorentz hyperbolic space as $\mathbf{x}^{(i)}$. To match images precisely with the tags, we measure the similarity between the image and each tag's h descriptions, then perform a weighted aggregation to produce a appropriate tag representation.

Specially, we adopt the Lorentz inner product in Eq equation 2 to measure the similarity between the image representation $\mathbf{x}^{(i)}$ and the h descriptions $\{\mathbf{x}^{(\mathrm{tag}_{j,1})}, \ldots, \mathbf{x}^{(\mathrm{tag}_{j,h})}\}$ of tag j. We then apply a softmax function to obtain the weight for each description:

$$w_{i,j,k} = \frac{\exp\Big(\langle \mathbf{x}^{(i)},\, \mathbf{x}^{(\mathrm{tag}_{j,k})} \rangle_L\Big)}{\sum_{k'=1}^{h} \exp\Big(\langle \mathbf{x}^{(i)},\, \mathbf{x}^{(\mathrm{tag}_{j,k'})} \rangle_L\Big)}, \quad k = 1, \ldots, h. \tag{5}$$

Next, we map the features $\mathbf{x}^{(\mathrm{tag}_{j,k})}$ back to the Euclidean space $\mathbf{x}_E^{(\mathrm{tag}_{j,k})} = Log(\mathbf{x}^{(\mathrm{tag}_{j,k})})$ and generate an aggregated representation of the j-th tag for the i-th image with a weighted combination over h description vectors:

$$\mathbf{x}_E^{(\mathrm{tag}_{i,j})} = \sum_{k=1}^{h} w_{i,j,k}\, \mathbf{x}_E^{(\mathrm{tag}_{j,k})}. \tag{6}$$

The aggregation mechanism emphasizes tag descriptions that are more semantically aligned with the image in hyperbolic space. As a result, the generated tag representations can better capture the hierarchical and contextual semantics present in the visual content.

3.3 HIERARCHICAL RELATIONSHIP MODELING

In hierarchical structures, it is usual to represent partial-order relations, such as "x is a general concept of y." Entailment cones is a common approach for capturing such relations in hyperbolic geometry. Following prior works Ganea et al. (2018a); Desai et al. (2023), we model hierarchical relations by checking if the cone of $\mathbf{x}$ contains $\mathbf{y}$, implying that $\mathbf{x}$ entails $\mathbf{y}$. Notably, the aperture of entailment cone decreases with distance from the origin, naturally encoding increasing specificity. Specifically, given a vector $\mathbf{x}$, we construct an entailment cone in hyperbolic space with $\mathbf{x}$ as its apex, and measure its width by a half aperture. The half aperture $\alpha_{\text{half}}(x)$ is defined by:

$$\alpha_{\text{half}}(\mathbf{x}) \;=\; \sin^{-1}\Big(\frac{2K}{\sqrt{k}\,\|\mathbf{x}\|_E}\Big), \tag{7}$$

where $-k$ is the curvature of the hyperbolic space, $K = 0.1$ is a constant that limits the half-aperture angle near the origin, and $\|\mathbf{x}\|_E$ denotes the Euclidean norm of $\mathbf{x}$. Therefore, as $\mathbf{x}$ moves farther from the origin, $\alpha_{\text{half}}(\mathbf{x})$ becomes smaller.

The hierarchical relations requires a general concept $\mathbf{x}$ to cover its specific concept $\mathbf{y}$. To assess whether $\mathbf{x}$ entails $\mathbf{y}$, we compute the external angle θ_{oxy} as follows:

$$\theta_{\text{oxy}} \;=\; \arccos\left(\frac{y_{\text{time}} \;+\; k\,\langle \mathbf{x}, \mathbf{y}\rangle_L\, x_{\text{time}}}{\|\mathbf{x}\|_E\,\sqrt{k\left(\langle \mathbf{x},\, \mathbf{y}\rangle_L\right)^2 \;-\; 1}}\right), \tag{8}$$

where $x_{\text{time}} = \sqrt{\frac{1}{k} + \|\mathbf{x}\|^2}, \quad y_{\text{time}} = \sqrt{\frac{1}{k} + \|\mathbf{y}\|^2}$ are the time components of $\mathbf{x}$ and $\mathbf{y}$, respectively. If $\theta_{\text{oxy}} < \alpha_{\text{half}}(\mathbf{x})$, then $\mathbf{x}$ entails $\mathbf{y}$. Otherwise, the hierarchy is contradicted. To encourage $\mathbf{y}$ to lie inside the entailment cone of $\mathbf{x}$, we introduce the following entailment loss:

$$\mathcal{L}_{\text{ent}}(\mathbf{x}, \mathbf{y}) = \max\Big\{\,0,\, \theta_{\text{oxy}} - \gamma\,\alpha_{\text{half}}(\mathbf{x})\Big\}, \tag{9}$$

where γ is a threshold controlling the maximum allowable external angle.

Unlike Desai et al. (2023); Pal et al. (2024), which imposes entailment constraints only between an image and its caption or a single tag name, we explicitly employ a hierarchical structure to model hierarchical relations among image I and caption C, image and multiple tags T, caption C and tags T, as well as child tag T_{child} and its parent tag T_{parent}. Therefore, we define the hierarchical loss as:

$$\mathcal{L}_{\text{hier}} = \mathcal{L}_{\text{ent}}(T, I) + \mathcal{L}_{\text{ent}}(C, I) + \mathcal{L}_{\text{ent}}(T, C) + \mathcal{L}_{\text{ent}}(T_{\text{parent}}, T_{\text{child}}). \tag{10}$$

Furthermore, we employ a tag-image attention mechanism and the Asymmetric Loss Ben-Baruch et al. (2020) $\mathcal{L}_{\text{align}}$ to effectively align images with associated tags. Please see Appendix for more details. The total loss of hierarchical image tagging loss is defined by:

$$\mathcal{L}_{\text{HIT}} = \mathcal{L}_{\text{align}} + w\mathcal{L}_{\text{hier}}. \tag{11}$$

where w is the weight for hierarchical loss $\mathcal{L}_{\text{hier}}$.

4 DATASET AND BENCHMARK WITH HIERARCHICAL TAG STRUCTURE

In order to evaluate hierarchical image tagging, we construct a general hierarchical tag tree referencing lexical semantic databases in linguistics, WordNet Fellbaum (1998), and a LLM. The hierarchical tag tree contains 10 levels and 3,334 tags along with their hierarchical relationships.

As shown in Figure 3, to construct the hierarchical structure, we first project all original tags onto a WordNet-based tree. During this process, we remove redundant synonyms (*e.g.,* "autumn" and "fall") and abstract concepts (*e.g.,* "adaptation" or "approach"), while enriching certain branches with familiar objects already present in WordNet, including "black swan" and "jackfruit". Although WordNet covers a broad vocabulary, it still lacks several common tags, such as "3D glasses" and "auto showroom". Therefore, we extend the tree with these tags from original list. Furthermore, since WordNet is grounded in psycholinguistic theories and organizes categories based on semantic relationships, its structure does not always align with visual perception or user expectations.

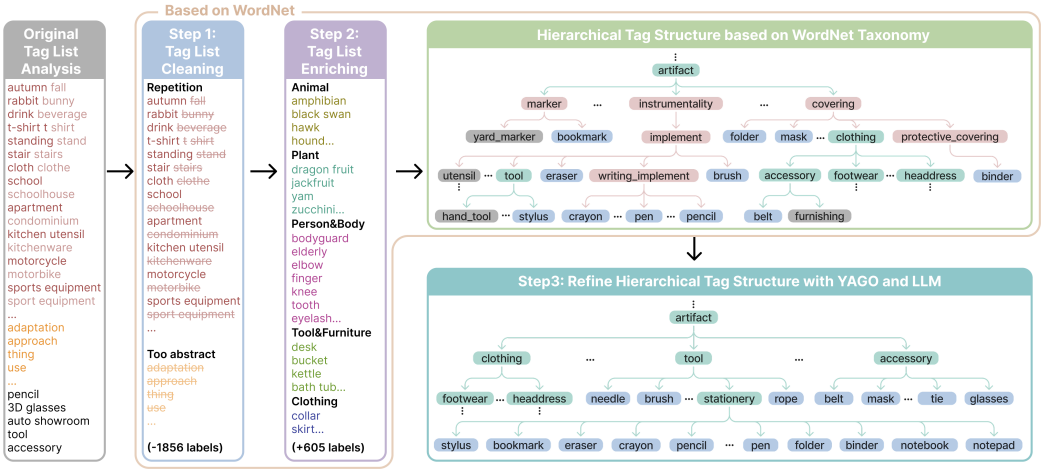

Figure 3: Process of constructing the hierarchical tag structure. Firstly, we refine and enrich the RAM++ tag list to build an hierarchical tag structure using the lexical semantic databases in linguistics, WordNet. Futhermore, the structure is optimized with reference to YAGO and a LLM to better align with visual perception and human cognition.

Thus, we reorganize and refine the structure with guidance from YAGO and a LLM. Figure 4 illustrates the subtree of the hierarchical tag structure, highlighting the tags in the primary branches. Overall, the complete structure contains 3,334 tags distributed across 10 levels, with 1, 7, 63, 230, 203, 655, 1194, 748, 201, and 32 tags from Level 1 to Level 10, respectively. The full tree and tag are provided in the appendix B.1. Our predefined taxonomy is flexible, and the tree-structured label space can be dynamically updated. This flexibility benefits from the well-organized structure of WordNet and the visual perception capabilities of large models in open-world settings. When encountering an unexplored tag in the open world, we first locate its position in WordNet, then assess its validity using a large model, and finally update the structure accordingly. For example, British Shorthair can be identified in WordNet as a subcategory of cat. After verifying the correctness of British Shorthair's hierarchical position with the large model, we update the taxonomy accordingly.

Figure 4: Visualization of a 6-level subtree with 113 tags from the full hierarchical tag structure.

Building on this hierarchical tag structure, we used a multimodal large model Qwen2.5-vl-72B Bai et al. (2025b) to annotate images in the CC3M dataset Sharma et al. (2018), and manually checked and corrected the annotations. However, the annotation from large model is low-quality and noisy due to potential hallucinations and biases from the large model. We spent significant 3 months reviewing and correcting the annotations. Each image's annotation includes its corresponding tags as well as the hierarchical relationships among those tags. In addition, we perform preliminary annotations for Open-Images Common dataset Kuznetsova et al. (2020), followed by manual checks of the annotation results. In the resulting benchmark, each image is labeled with a tag set and a tag tree derived from the hierarchy of these tags. Table 1 provides detailed benchmark statistics and comparisons with common multi-label classification datasets.

We construct an evaluation pipeline employing Tree Edit Distance (TED), Jaccard Similarity (J), Hierarchical Precision (P_H) and Hierarchical Recall (R_H) to evaluate the predicted hierarchical structures. These metrics provide a comprehensive assessment of both the accuracy and completeness of hierarchical structure prediction.

Table 1: **Comparison of multi-label classification datasets in terms of images, labels and levels.**

Dataset	Images	Classes	Tree-like	Levels	Train	Test
Cityscapes Cordts et al. (2016)	5,000	19	✗	1	2,975	1,525
Pascal VOC Everingham et al. (2010)	11,540	20	✗	1	5,717	–
COCO Lin et al. (2014)	123287	80	✗	1	82,783	40,775
NUS-WIDE Chua et al. (2009)	269,648	81	✗	1	161,789	107,859
Objects365 Shao et al. (2019)	738,000	365	✗	1	600,000	100,000
OpenImages Kuznetsova et al. (2020)	9,177,125	600	✗	1	9,011,219	125,436
RAM Zhang et al. (2024)	14,000,000	4,585	✗	1	14,000,000	–
SubPartImageNet Myers-Dean et al. (2024)	10,387	254	✓	3	8,828	1,040
Mapillary Vistas 2.0 Neuhold et al. (2017)	25,000	144	✓	3	18,000	5,000
HiTag (ours)	2,929,235	3,334	✓	10	2,872,012	57,223

Table 2: **Comparison of hierarchical image tagging performance among open-set image tagging models, vision language models and hierarchical image classification models.** 'w/Hier' indicates whether the method performs hierarchical classification. 'TED', 'J', 'P_H', and 'R_H' refer to the tree edit distance, Jaccard similarity, hierarchical precision, and hierarchical recall metrics respectively.

Methods	Training Images	*w/* Hier.	TED↓	J↑	P_H ↑	R_H ↑
MKT He et al. (2023)	162K	✗	47.45	0.24	0.38	0.46
BLIP Li et al. (2022a)	129M	✗	38.93	0.24	0.53	0.37
DiHT Radenovic et al. (2023)*	438M	✗	43.77	0.22	0.53	0.35
CLIP Radford et al. (2021)	400M	✗	50.12	0.04	0.12	0.08
MERU Desai et al. (2023)	12M	✓	41.77	0.21	0.50	0.32
HyCoCLIP Pal et al. (2024)	23.5M	✓	35.13	0.22	0.56	0.30
RAM Zhang et al. (2024)*	14M	✗	37.09	0.33	0.49	0.55
RAM++ Huang et al. (2023a)*	14M	✗	34.88	0.35	0.52	0.57
GPT-4o	-	✗	34.85	0.14	0.61	0.16
Gemini-2.0	-	✗	34.21	0.18	0.64	0.21
Qwen2.5-vl-72B Bai et al. (2025a)	-	✗	38.34	0.44	0.71	0.56
RAM++ Huang et al. (2023a)	3M	✗	48.15	0.28	0.38	0.57
HiTag (ours)	3M	✓	**24.83**	**0.56**	**0.64**	**0.82**

5 EXPERIMENT

Baselines. We train our model on HiTag training dataset, which contains 3M images from CC3M with tags and their hierarchical relations. For fair comparison, we train RAM++ on the CC3M dataset without fine-tuning. Furthermore, we compare our model against many other models He et al. (2023); Li et al. (2022a); Radenovic et al. (2023); Radford et al. (2021); Desai et al. (2023); Pal et al. (2024); Zhang et al. (2024); Huang et al. (2023a) trained on much larger datasets Cordts et al. (2016); Everingham et al. (2010); Lin et al. (2014); Chua et al. (2009); Shao et al. (2019); Kuznetsova et al. (2020); Zhang et al. (2024); Myers-Dean et al. (2024); Neuhold et al. (2017). Both our method and the baseline methods, such as RAM++, generate diverse descriptions for each image using the large model, making the experimental comparisons fair.

Implementation Details. Similar to RAM++ and RAM Huang et al. (2023a); Zhang et al. (2024), we employ $Swin_{Large}$ Liu et al. (2021) as the image encoder for fair comparison. We employ an off-the-shelf text encoder of HyCoCLIP Pal et al. (2024) to extract caption and tag description embeddings. The input size for training images is 224×224, and the image size remains unchanged during evaluation. We use AdamW optimizer and step lr scheduler with a base learning rate of 1e-4 and a weight decay of 0.9. We choose 10 as our hierarchical loss weight. Training our model for 10 epochs on 8 NVIDIA A800 GPUs takes approximately 11 hours.

5.1 HIERARCHICAL IMAGE TAGGING

We evaluate hierarchical tagging performance on the manually annotated images with hierarchical tags from the Open Images Common dataset. We employ four tree-based hierarchical metrics: TED, J, P_H, and R_H.The comparison of HiTag with other open-set image tagging models, vision language models and hierarchical image classification models is shown in Table 2. Across all four metrics, HiTag demonstrates substantial improvements over state-of-the-art image tagging approaches with much larger training datasets. Although LLMs exhibit the ability to construct tree-like structure, they only achieves comparable performance with image tagging approaches without hierarchy in certain settings. Moreover, even when compared to hierarchical classification methods trained on large-scale datasets, HiTag still shows superior hierarchical tagging performance.

Table 3: **Image tagging performance comparison of state-of-the-art open-set and closed-set image tagging models on mAP.** Closed-set models are fully supervised and trained on vertical-domain datasets. "Inference prompt" indicates the text prompt used during inference, such as Hand-Written: "A photo of a dog" or LLM Tag Description: "A dog has a furry body, a tail. . .". Models marked with $*$ have external fine-tuning.

Methods	Training Images	Inference Prompt	OpenImages Common	OpenImages Uncommon	ImageNet-Multi
Closed-Set Models					
ML-Decoder Ridnik et al. (2023)	9M	-	85.8	79.5	-
Tag2Text* Huang et al. (2023b)	4M	-	82.9	-	-
Tag2Text* Huang et al. (2023b)	14M	-	83.4	-	-
Open-Set Models					
MKT He et al. (2023)	162K	Hand-Written	77.8	63.5	49.8
BLIP Li et al. (2022a)	129M	Hand-Written	75.7	61.1	39.0
DiHT* Radenovic et al. (2023)	438M	Hand-Written	71.3	62.4	-
CLIP Radford et al. (2021)	400M	Hand-Written	72.8	65.6	56.0
RAM* Zhang et al. (2024)	14M	LLM Tag Des	82.2	65.9	-
RAM++* Huang et al. (2023a)	14M	LLM Tag Des	86.6	75.4	63.3
RAM++ Huang et al. (2023a)	3M	LLM Tag Des	84.3	68.5	53.4
HiTag (ours)	3M	LLM Tag Des	**85.6**	**68.6**	**57.8**

Table 4: Ablation study on the effect of entailment loss for hierarchical image tagging.

Hyperbolic representation	Entailment loss				Hierarchical image tagging			
	$L_{\text{ent}}(T,I)+L_{\text{ent}}(C,I)$	$L_{\text{ent}}(T,C)$	$L_{\text{ent}}(T_{parent},T_{child})$		TED (↓)	J (↑)	P_H (↑)	R_H (↑)
✓					25.57	0.54	0.63	0.81
✓	✓				25.51	0.54	0.63	0.82
✓	✓	✓			25.53	0.55	0.63	0.82
✓	✓	✓	✓		24.83	0.56	0.64	0.82

5.2 Image Tagging

For image tagging, we compare with state-of-the-art open-set (*e.g.,* RAM++, RAM) and closed-set (*e.g.,* ML-Decoder, Tag2Text) models on three multi-tag recognition tasks: OpenImages Common, OpenImages Uncommon and ImageNet-Multi Yun et al. (2021). To assess the performance of image tagging, we use the commonly used mean Average Precision (mAP) as the evaluation metric. The results in Table 3 show that HiTag outperforms the state-of-the-art open-set image tagging model RAM++ with the same 3M training images as HiTag and significantly surpasses most leading open-set models across all tasks. Moreover, HiTag achieves comparable performance to RAM trained on 14M images and to the closed-set model Tag2Text on the OpenImages Common and Uncommon benchmarks.

5.3 Ablation Study

We conduct a comprehensive ablation study to analyze the contribution of hyperbolic representation and each component in our proposed entailment loss framework, as shown in Table 4. First, we observe that hyperbolic representation provides a respectable performance on hierarchical image tagging (TED = 25.57, J = 0.54). Introducing the entailment loss $L_{\text{ent}}(T,I)+L_{\text{ent}}(C,I)$ and $L_{\text{ent}}(T,C)$ brings marginal gains in hierarchical tagging accuracy (*e.g.,* R_H from 0.81 $\rightarrow$ 0.82). Further incorporating $L_{\text{ent}}(T_{parent},T_{child})$ consistently enhances hierarchical tagging metrics, highlighting the utility of modeling parent–child entailment relationships among tags. Finally, combining the hyperbolic representation with all three entailment components achieves the best overall performance on hierarchical image tagging tasks (TED = 24.83, J = 0.56, P_H = 0.64), verifying the effectiveness of jointly enforcing hyperbolic geometry and multi-level entailment constraints.

6 Conclusion

In this work, we introduce HiTag, the largest and most structurally rich benchmark for hierarchical image tagging, featuring 3M annotated images with 3,334 hierarchically organized tags across 10 levels. Beyond establishing a new benchmark, we demonstrate the effectiveness of hyperbolic space for modeling hierarchical relationships between images, captions, and tags. Our hierarchical image tagging approach utilize entailment cones in Lorentz space to model hierarchical relations and achieves state-of-the-art performance in image tagging tasks.

7 Ethics Statement

This work does not involve human subjects, personally identifiable data, or sensitive information. The datasets used in our experiments (CC3M and OpenImages) are publicly available, and our use strictly follows their licenses. To ensure data quality and reproducibility, our authors manually checked the annotation results and manually annotated the test set. These efforts were strictly limited to verifying and refining image–label associations. Manual inspection and annotation were carried out to reduce bias and ensure that the benchmark reflects semantically consistent and visually coherent categories, supporting fair evaluation of hierarchical image tagging models.

Importantly, no human subjects were involved in model training, experimental evaluation, or system assessment, and no sensitive, private, or personally identifiable data were collected or processed. Our research adheres to the principles of responsible stewardship, upholds high standards of scientific excellence, and supports the public good by releasing a high-quality benchmark for the community. No additional ethical concerns are associated with this paper.

8 Reproducibility Statement

To facilitate reproducibility, we provide anonymous repository links containing code [1] and benchmark [2]. The core method and the construction of the hierarchy and benchmark are specified separately in the Method section and the Dataset and Benchmark with Hierarchical Tag Structure section. Implementation details and evaluation protocols containing models, baselines, metrics, and train are in both the Experiment section and the Appendix. The Appendix also includes the construction details of the hierarchy and benchmark, additional method details, further performance comparisons, and more experiments.

References

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4453–4462, June 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025a. URL https://arxiv.org/abs/2502.13923.

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

Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification, 2020.

Avi Ben-Cohen, Nadav Zamir, Emanuel Ben-Baruch, Itamar Friedman, and Lihi Zelnik-Manor. Semantic diversity learning for zero-shot multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 640–650, October 2021.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

James W. Cannon, William J. Floyd, Richard Kenyon, and Walter R. Parry. Hyperbolic geometry. In *Flavors of Geometry*, volume 31 of *MSRI Publications*, pp. 59–115. Mathematical Sciences Research Institute, 1997. URL http://library.slmath.org/books/Book31/files/cannon.pdf.

[1] Our code is available at https://anonymous.4open.science/r/HiTag.

[2] Our benchmark is available at https://huggingface.co/datasets/hitag001/HiTag-Dataset.

Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. Neural embeddings of graphs in hyperbolic space, 2017. URL https://arxiv.org/abs/1705.10359.

Ines Chami, Aditya Wolf, Federico Ruiz, Christopher Ré, and William L. Hamilton. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4868–4879, 2019. URL https://dl.acm.org/doi/10.5555/3454287.3454725.

Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your "flamingo" is my "bird": Fine-grained, or not. In *Computer Vision and Pattern Recognition*, 2021.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3558–3568, June 2021.

Bike Chen, Wei Peng, Xiaofeng Cao, and Juha Röning. Hyperbolic uncertainty aware semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(2):1275–1290, 2024. doi: 10.1109/TITS.2023.3312290.

Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification, 2022. URL https://arxiv.org/abs/2201.03194.

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584805. doi: 10.1145/1646396.1646452. URL https://doi.org/10.1145/1646396.1646452.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Son D. Dao, Dat Huynh, He Zhao, Dinh Phung, and Jianfei Cai. Open-vocabulary multi-label image classification with pretrained vision-language model. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2135–2140, 2023. doi: 10.1109/ICME55011.2023.00365.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Xiwen Dengxiong and Yu Kong. Ancestor search: Generalized open set recognition via hyperbolic side information learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4003–4012, January 2023.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. 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. 7694–7731. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/desai23a.html.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Kernel methods in hyperbolic spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10665–10674, October 2021.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Luca Franco, Paolo Mandica, Konstantinos Kallidromitis, Devin Guillory, Yu-Teng Li, Trevor Darrell, and Fabio Galasso. Hyperbolic active learning for semantic segmentation under domain shift. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=hKdJPMQvew.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. 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. 1646–1655. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/ganea18a.html.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5345–5355, 2018b. URL https://dl.acm.org/doi/10.5555/3327345.3327440.

Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. Curvature generation in curved spaces for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8691–8700, October 2021.

Ashima Garg, Shaurya Bagga, Yashvardhan Singh, and Saket Anand. Hiermatch: Leveraging label hierarchies for improving semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1015–1024, January 2022a.

Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 252–267, Cham, 2022b. Springer Nature Switzerland. ISBN 978-3-031-20053-3.

Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks, 2020. URL https://arxiv.org/abs/2010.10151.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. Hyperbolic attention networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJxHsjRqFQ.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018a. URL https://arxiv.org/abs/1703.06870.

Shiyi He, Chang Xu, Tianyu Guo, Chao Xu, and Dacheng Tao. Reinforced multi-label image classification by exploring curriculum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018b. doi: 10.1609/aaai.v32i1.11770. URL https://ojs.aaai.org/index.php/AAAI/article/view/11770.

Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Xiujun Shu, Bo Ren, and Shu-Tao Xia. Open-vocabulary multi-label classification via multi-modal knowledge transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):808–816, Jun. 2023. doi: 10.1609/aaai.v37i1.25159. URL https://ojs.aaai.org/index.php/AAAI/article/view/25159.

Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3450–3462, 2024. doi: 10.1109/TPAMI.2023.3346405.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pp. arXiv–2310, 2023a.

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023b.

Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text 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. 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Fanjie Kong, Yanbei Chen, Jiarui Cai, and Davide Modolo. Hyperbolic learning with synthetic captions for open-world detection, 2024. URL https://arxiv.org/abs/2404.05016.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, Sep 2010. doi: 10.1103/PhysRevE.82.036106.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL http://dx.doi.org/10.1007/s11263-020-01316-z.

Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings, 2019. URL https://arxiv.org/abs/1902.00913.

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

J.M. Lee. *Introduction to Riemannian Manifolds*. Graduate Texts in Mathematics. Springer International Publishing, 2019. ISBN 9783319917542. URL https://books.google.com.hk/books?id=UIPltQEACAAJ.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022a.

Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1246–1257, June 2022b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Carolus Linnaeus. *Systema naturae, 1735*, volume 8. Brill, 2024.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/103303dd56a731e377d01f6a37badae3-Paper.pdf.

Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=sHswzNWUW2.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015. URL https://arxiv.org/abs/1411.4038.

Teng Long, Pascal Mettes, Heng Tao Shen, and Cees G. M. Snoek. Searching for actions on the hyperbole. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yan Luo, Ming Jiang, and Qi Zhao. Visual attention in multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. URL https://arxiv.org/abs/1306.5151.

Emile Mathieu, Charline Le Lan, Chris J. Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous hierarchical representations with poincaré variational auto-encoders. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Josh Myers-Dean, Jarek Reynolds, Brian Price, Yifei Fan, and Danna Gurari. Spin: Hierarchical segmentation with subpart granularity in natural images. In *Proc. European Conf. Computer Vision (ECCV)*, 2024.

Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8731–8740, October 2021.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf.

Bureau of Labor Statistics. *American Time Use Survey (ATUS): Arts Activities*. Inter-university Consortium for Political and Social Research, 2003-2023. URL https://doi.org/10.3886/ICPSR36268.v8.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. 2024. URL https://arxiv.org/abs/2410.06912.

Seulki Park, Youren Zhang, Stella X. Yu, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification, 2025. URL https://arxiv.org/abs/2406.11608.

Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044, December 2022. doi: 10.1109/TPAMI.2021.3136921. URL https://doi.org/10.1109/TPAMI.2021.3136921.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6967–6977, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Shafin Rahman, Salman Khan, and Nick Barnes. Deep0tag: Deep multiple instance learning for zero-shot image tagging. *IEEE Transactions on Multimedia*, 22(1):242–255, 2020. doi: 10.1109/TMM.2019.2924511.

J. Ratcliffe. *Foundations of Hyperbolic Manifolds*. Graduate Texts in Mathematics. Springer New York, 2006. ISBN 9780387331973. URL https://books.google.com.hk/books?id=JV9m8o-ok6YC.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. URL https://arxiv.org/abs/1506.01497.

Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Multi-instance visual-semantic embedding, 2015. URL https://arxiv.org/abs/1512.06963.

Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 32–41, 2023.

Michael A Ruggiero, Dennis P Gordon, Thomas M Orrell, Nicolas Bailly, Thierry Bourgoin, Richard C Brusca, Thomas Cavalier-Smith, Michael D Guiry, and Paul M Kirk. A higher level classification of all living organisms. *PloS one*, 10(4):e0119248, 2015.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift, 2020. URL https://arxiv.org/abs/2008.04859.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In Marc J. van Kreveld and Bettina Speckmann (eds.), *Graph Drawing – 19th International Symposium, GD 2011, Eindhoven, The Netherlands, September 21–23, 2011, Revised Selected Papers*, volume 7034 of *Lecture Notes in Computer Science*, pp. 355–366, Berlin, Heidelberg, 2011. Springer. ISBN 978-3-642-25878-7. doi: 10.1007/978-3-642-25878-7_34.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, October 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Ryohei Shimizu, YUSUKE Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ec85b0tUwbA.

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g6xeSKDS.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30569–30582, 2022.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 595–604, 2015. doi: 10.1109/CVPR.2015.7298658.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8769–8778, 2018.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Sinan Wang, Xinyang Chen, Yunbo Wang, Mingsheng Long, and Jianmin Wang. Progressive adversarial networks for fine-grained domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Wenhao Wang, Yifan Sun, Wei Li, and Yi Yang. TransHP: Image classification with hierarchical prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=vpQuCsZXz2.

Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1901–1907, 2016. doi: 10.1109/TPAMI.2015.2491929.

Shu-Lin Xu, Yifan Sun, Faen Zhang, Anqi Xu, Xiu-Shen Wei, and Yi Yang. Hyperbolic space with hierarchical margin boosts fine-grained learning from coarse labels. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 71263–71274, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e17e11960843febbc2dd22d3c7d79144-Paper-Conference.pdf.

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2340–2350, 2021.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16660–16669, June 2022a.

Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yiding Zhang, Xiao Wang, Chuan Shi, Xunqiang Jiang, and Yanfang Ye. Hyperbolic graph attention network. *IEEE Transactions on Big Data*, 8(6):1690–1701, 2022b. doi: 10.1109/TBDATA.2021.3081431.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1724–1732, 2024.

Rixin Zhou, Jiafu Wei, Qian Zhang, Ruihua Qi, Xi Yang, and Chuntao Li. Multi-granularity archaeological dating of chinese bronze dings based on a knowledge-guided relation graph, 2023. URL https://arxiv.org/abs/2303.15266.

Appendix

A Usage of Large Language Models

We leverage large language models to refine and adjust the hierarchical tag structure so that it better aligns with visual perception and ensures semantic consistency across different levels of the taxonomy. Based on this optimized hierarchy, we further employ a vision–language large model to automatically annotate tags for the images in the training sets of the benchmark. To guarantee annotation accuracy and reliability, all generated labels were subsequently manually inspected and corrected by the authors, ensuring that the final benchmark data maintains both structural consistency and high-quality annotations.

B Benchmark Details

B.1 Hierarchy Structure

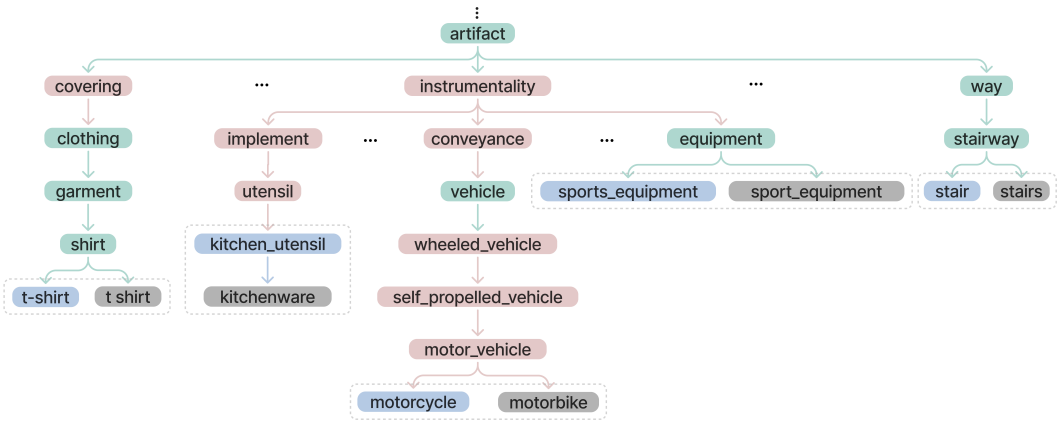

Figure 5: Step 1 of the process for constructing the hierarchical tag structure.

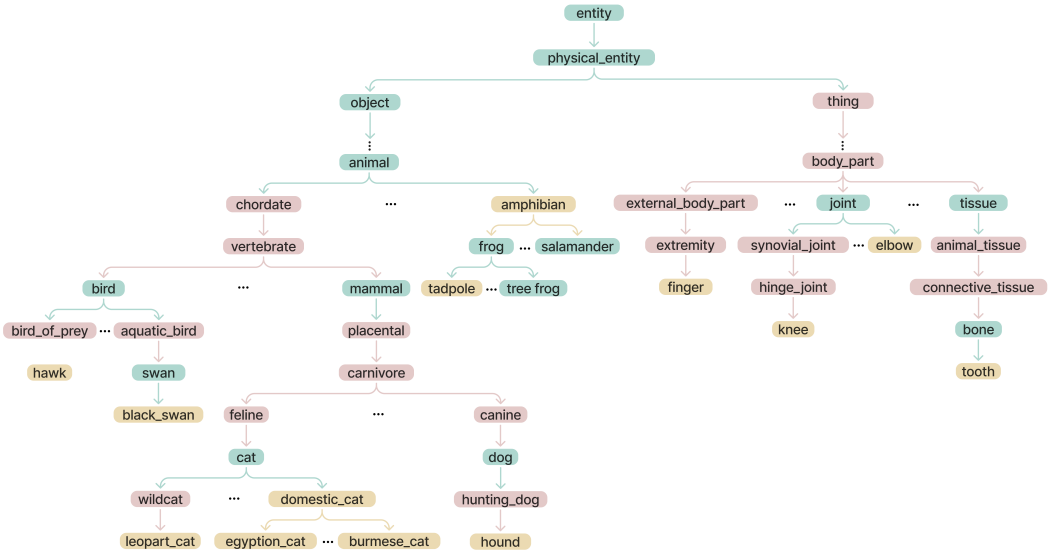

Figure 6: Step 2 of the process for constructing the hierarchical tag structure.

We show the processes of step 1 and step 2 in Figure 5 and Figure 6, respectively. In Step 1, mapping the original tag list to WordNet revealed redundant synonyms, which we removed. We then refined the list by eliminating abstract terms that are visually ambiguous. In Step 2, we identified additional common tags in WordNet, particularly under the animal branch, and integrated them into our tag tree.

We compare the coverage of general categories across hierarchical classification datasets in Table 5. SubPartImageNet contains a small number of animal and transportation categories, together with many of their subcategories. Although Mapillary Vistas 2.0 covers certain category aspects, it offers very few categories within each aspect. For example, its animal aspect includes only the categories bird and ground-animal. Our HiTag dataset encompasses a more wider range of categories.

Table 5: Hierarchical coverage comparison of the hierarchical classification datasets.

(a) Natural object and Natural phenomenon

Dataset	Natural object							Natural phenomenon
	Plant	Animal	Person	Biological parts	Microorganism	Celestial body	Geographical feature	
SubPartImageNet	✗	✓	✗	✓	✗	✗	✗	✗
Mapillary 2.0	✗	✓	✓	✗	✓	✗	✓	✗
HiTag (ours)	✓	✓	✓	✓	✓	✓	✓	✓

(b) Event, Mythological creatures, Group, Attribute and Food

Dataset	Food	Event					Mythological creatures	Group	Attribute	Food
		Action	Competitive event	Celebratory event	Catastrophic event	Life event				
SubPartImageNet	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Mapillary 2.0	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
HiTag (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

(c) Artifact

Dataset	Artifact									
	Built environment	Tool	Equipment	Furniture	Clothing	Transportation	Accessory	Decoration	Weapon	Communication media
SubPartImageNet	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Mapillary 2.0	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
HiTag (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

We present the complete hierarchical tag structures of the SubPartImageNet dataset, the Mapillary Vistas 2.0 dataset and our HiTag dataset in Figure 7, Figure 8 and Figure 9, respectively. Compared to other structures, the SubPartImageNet structure is much simpler. Besides, the Mapillary Vistas 2.0 hierarchy exhibits extensive duplication of category names at levels 3 and 4. Compared with these structures, our architecture covers a richer set of tags and exhibits a more complete overall structure.

Since displaying the full hierarchy would require more than 60 pages, we have placed the complete hierarchical tag structure of the dataset in the anonymous repository [3] of the provided code. It includes the name and definition of each category and is saved in two formats: hitag/-data/hire_tree_file/object_structure.json and hitag/data/hire_tree_file/object_structure.txt. We show all the 3,334 tags of HiTag in 10 levels:

- Level 1 : [entity]
- Level 2 : [physical entity, natural phenomenon, mythological creatures, group, measure, attribute, event]
- Level 3 : [object, weather, snow, rain, wave, sunrise, sunset, eclipse, hot spring, flame, flood, earthquake, aurora, bolt, dew, lightning, flare, moonlight, sunshine, fire, wildfire, icicle,

[3] Our code is available at https://anonymous.4open.science/r/HiTag.

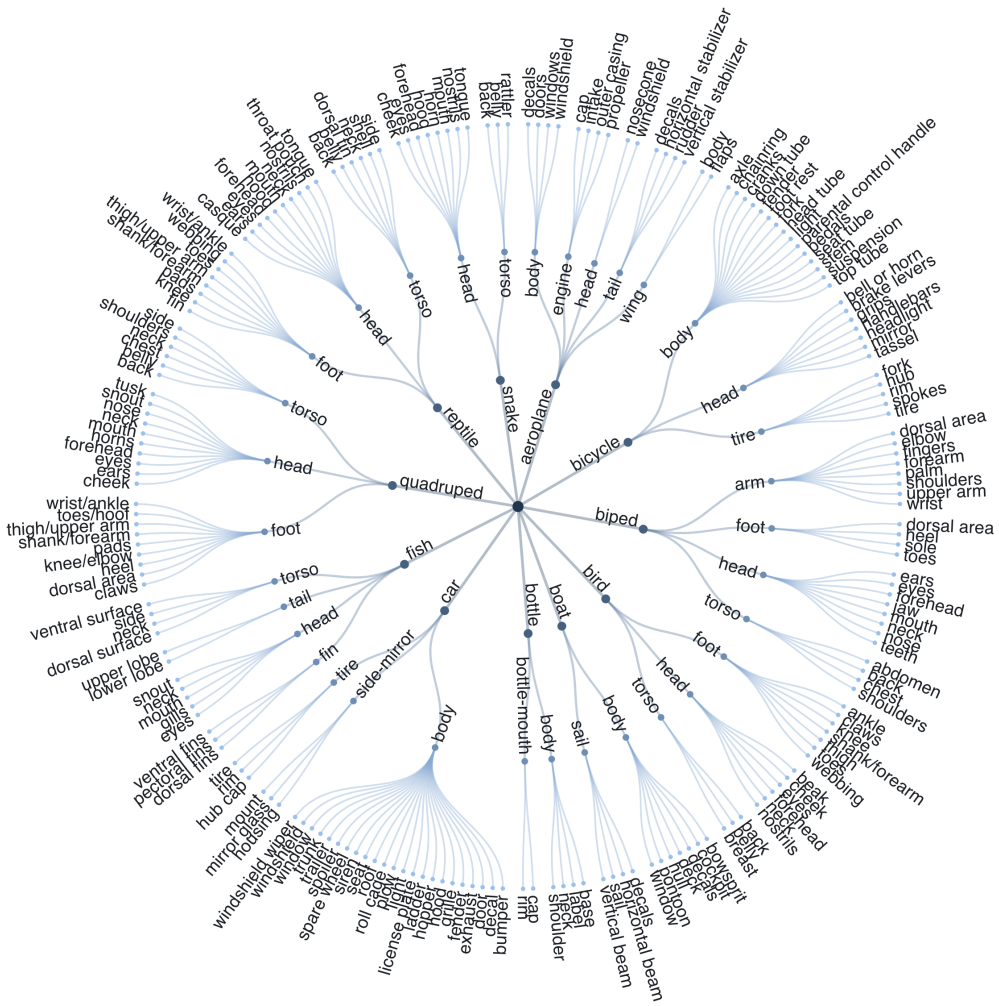

Figure 7: Hierarchical tag structure of SubPartImageNet.

frost, light, shadow, darkness, sparkle, rainbow, reflect, fairy, monster, magic, ghost, queue, social group, organization, collective assembly, time, time period, score, number, numeric quantity, water level, part, scene, appearance, view, state, quality, cognition, recycling, prayer, activity, test, trade, action, competitive event, celebratory event, catastrophic event, life event, entertainment, healthcare activities, communication activities]

- Level 4 : [natural object, artifact, food, matter, hail, foggy, rainy, snowy, stormy, sunny, warm, windy, storm, snowstorm, tornado, cloudy, santa claus, cupid, angel, loki, leprechaun, vampire, witch, mermaid, unicorn, dragon, lineup, friend, gang, family, couple, tribe, demographic group, community, flock, herd, business group, educational group, government group, military group, artistic group, sports team, charity, team, crew, parade, exhibition, banquet, meeting, gathering, crowd, vacation, calendar, season, night, morning, evening, half, quarter, slice, natural scene, scratch, wound, burn, pawprint, shine, footprint, bite, transparency, freckle, stain, profile, beautiful, texture, style, dark, scar, crack, plaque, symmetry, reflection, wrinkle, color, shape, underwater, horizon, city view, mountain view, night view, sea view, outdoor, street view, harbor view, winter view, village view, wet, bright, pose, rural, shirtless, pollution, relax, hot, sleep, drought, mess, flush, tan, injury, fitness, trim, disease, depression, frozen, bloom, hang, damp, calm, luxury, athletic, business, sale, auction, shopping, move, baking, demolition, aid, cheer, archaeological excavation, grazing,

Figure 8: Hierarchical tag structure of Mapillary Vistas 2.0.

talking, arrest, trial, celebrate, touch, rescue, kiss, throw, roll, dribble, sunbathe, capture, injection, handstand, squat, wrestle, looking, stare, wash, hug, feeding, push, pull, smile, salute, applause, handshake, greeting, help, teach, interaction, putt, landing, catch, casting, kneel, cooking, carry, punch, kick, stab, spin, repair, eat, pour, sing, sit, sew, carving, type, bathing, washing, blow, clutch, clean, swinge, haircut, act, squeeze, sport, competition, festival, new year, halloween, christmas, bonfire, wedding reception, fireworks display, party, ceremony, forest fire, conflict, battle, accident, explosion, disaster, emergency, traffic jam, crash, birth, death, wedding day, anniversary, show, game, camping, boat ride, fishing, therapy, exercise, massage, invitation, speech, interview, debate, live, call, protest, dating]

- Level 5 : [plant, animal construct, biological parts, animal, person, microorganism, celestial body, geographical feature, natural materials, atmosphere, built environment, passage, product, tool, equipment, furniture, clothing, transportation, accessories, decoration, weapon, communication media, recreational objects, money, coconut meat, chocolate, baked goods, meat, feed, drink, fast food, meal, dishes, western food, grain, dessert, sweet, condiment, dairy, egg, macaroni, spaghetti, dumpling, noodle, cereal, tofu, dough, batter, stuffing, tea bag, baby food, bento, street food, canned food, muesli, incense, amber, rust, bubble, peel, fuel, states of matter, material, youth, elderly, town, village, downtown, city, countryside, company, bank, market, school, academy, class, government, court, fire

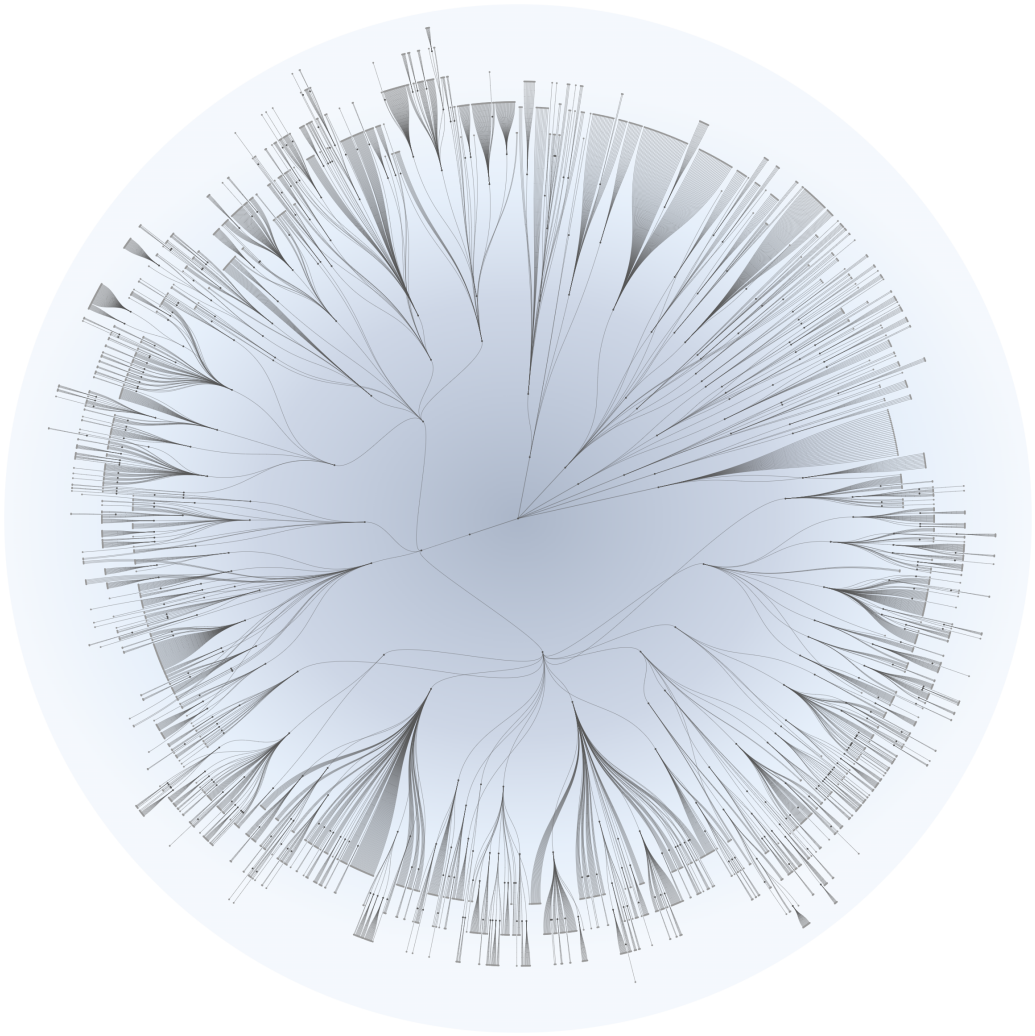

Figure 9: Visualization of the whole hierarchical tag structure which contains 3,334 tags across 10 levels.

department, municipality, police, army, navy, troop, band, choir, ensemble, marching band, baseball team, basketball team, football team, ice hockey team, cricket team, art exhibition, car show, airshow, seminar, winter, summer, spring, autumn, rough, soft, smooth, modern, white, black, blue, yellow, green, pink, red, brown, purple, gray, orange color, point, round shape, figure, cylinder, curve, cube, sphere, standing, upright, industry, agriculture, shipping, jump, climb, ride, sailing, hike, chase, walk, run, flip, takeoff, surfing, taekwondo, walking, tai chi, mountain climbing, high jump, archery, swim, cycling, horse riding, ice skating, yoga, boxing, hockey, football, ping pong, rugby, bullfighting, handball, badminton, polo, karting, motorsport, rock climbing, weightlifting, mountain biking, hockey game, softball, baseball, tennis, volleyball, golf, basketball, bowling, table tennis, karate, football competition, baseball competition, tennis competition, volleyball competition, basketball competition, race, wedding party, birthday party, prom, dinner party, celebration, wedding, funeral, graduation, award ceremony, birthday, performance, juggle, light show, fashion show, talk show, stunt, karaoke, lottery, video game, billiards, card game, playing chess, playing mahjong]

- Level 6 : [vine, houseplant, algae, plant organs, vegetable, spice, tree, moss, lichen, fern, grass, cactus, bush, shrub, wheat, succulent, reed, nest, spider web, cell, animal coverings, body, muscle, wing, tail, skeleton, fin, claw, whisker, antler, penis, paw, limb stump, nipple, tooth, blood, invertebrate, amphibian, fish, reptile, bird, mammal, life stage, family relation, social role, leader, occupation, sportsman, hacker, speaker, camper, tailor, handicapped person, zombie, hiker, mold, fungi, earth, moon, sun, star, galaxy, planet, constellation, nebula, fossil, area, place, territory, surface, shoreline, harbor, rise, corner, hole, angle, habitat, wilderness, body of water, landform, world, ice cube, pebble, jade, coal, petroleum, stone, smoke, mist, fog, sky, cloud, building, infrastructure, facility, grave, estate, construction site, structure, debris, outdoor amenity, barrier, interior spaces, drain, pipe, chimney, corridor, hallway, aisle, elevator shaft, entrance, archway, goods, cargo, medicine, dye, acrylic paint, freshener, detergent, cleaner, cigar, cigarette, spray, clothespin, umbrella, hooks, tableware, container, candle, cassette deck, envelope, mp3 player, software, stamp, ink pad, paper, stationery, toiletry, flashlight, book cover, test tube, stick, personal care, charger, salt shaker, duct tape, hose, rope, tripod, easel, fishing net, crowbar, crutch, cane, ladder, connector, sharpener, brush, oar, paddle, needle, broom, pole, drill, jack, plow, rake, hoe, mower, wrench, screwdriver, saw, shovel, hammer, opener, razor, pocketknife, scissors, axe, 3d glasses, tape measure, plunger, pliers, awl, funnel, tape, ice pack, shaker, basket, bucket, pipe bowl, bin, manger, bundle, tray, ashtray, dustpan, basin, barrel, tub, jar, beaker, bottle cap, fishbowl, yoke, float, palette, ruler, pitchfork, razor blade, horseshoe, floss, rein, leash, screw, lock, clip, roller, rubber band, key, hairdryer, compass, stethoscope, syringe, thermometer, iron, magnifying glass, binoculars, handcuffs, lid, stapler, stirrup, gauge, hourglass, mallet, hanger, magnet, mousetrap, slot, thresher, copperware, match, harvester, pegboard, charcoal burner, dryer, vacuum cleaner, trimmer, concrete mixer, trouser press, appliance, musical instrument, kitchen utensil, screen, projection screen, spirit lamp, anchor, valve, pump, wiper, handle, vane, propeller, bumper, sprinkler, switch, headlight, torch, radar, lightning rod, robot, firefighting equipment, brake, laser, battery, solar panel, sports equipment, camera, kit, heater, pendulum, phone, parachute, poker chip, oil rig, life jacket, lifebuoy, plug, socket, showerhead, scale, juicer, submersible, megaphone, recording device, game controller, keyboard, computer, television, radio, calculator, atm, gear, engine, machine, loom, printer, scanner, typewriter, windmill, grinder, cotton gin, mill, slicer, lathe, snow blower, shredder, fan, water cooler, satellite dish, cd player, amplifier, excavator, generator, record player, monitor, projector, air conditioner, fridge, sundial, microscope, lighting, telescope, remote, medical equipment, dropper, hand dryer, headphones, panel, circuit board, wire, humidifier, jukebox, electric chair, curtain, canopy, window screen, lamp, mirror, clock, watch, storage, seat, table, bedding, washstand, bathtub, toilet, fixture, textile, headdress, garment, footwear, laundry, transportation structure, vehicle, lace, hairnet, belt, broach, mask, earmuff, wristband, hairgrip, hairpin, belt buckle, collar, cigarette case, cuff, dog collar, button, zipper, mousepad, tie, glove, pocket, suspenders, knee pad, elbow pad, glasses, jewelry, mannequin, wallpaper, wind chime, picture frame, totem pole, celebratory items, brass plaque, tinsel, floral arrangement, tapestry, plume, tassel, lanyard, balloon arch, vase, disco ball, antique, ribbon, souvenir, birthday candle, gun, sword, spear,

bomb, sling, dart, bullet, arrow, bow, tomahawk, brass knuckles, pike, flamethrower, cannon, missile, protective gear, net, web page, message, physical media, symbolic representation, visual identity, art, image, time based media, toy, board game, dartboard, frisbee, kite, snowball, seesaw, kaleidoscope, foosball, hula hoop, wand, music box, puzzle, domino, jigsaw puzzle, card, digital gaming, die, digital currency, currency, chocolate bar, white chocolate, cake, bread, zongzi, pie, baozi, eclair, breadstick, gingerbread house, chapati, pastry, shellfish, calamari, octopus tentacles, roe, chicken wing, escargot, sausage, ham, bacon, lamb chop, poultry, beef, pork, hay, pet food, soy milk, coffee, cocoa, tea, alcohol, juice, milk, drinking water, lemonade, milk tea, shake, breakfast, buffet, feast, picnic, caviar, roast squab, mashed potatoes, fries, bibimbap, foie gras, roast chicken, home fries, baked potato, taco, pizza, sashimi, casserole, tempura, paella, burrito, quesadilla, porridge, boiled egg, meatball, sushi, pasta, lasagna, salad, soup, fried egg, curry, fried rice, barbecue, stew, kebab, omelette, french toast, sandwich, potato chip, popcorn, fruit dish, grilled eel, onion ring, scrambled eggs, spring roll, croquette, fried chicken, hotpot, chow mein, fish sticks, meatloaf, flour, cornmeal, rice, paddy, corn cob, oats, wheat grain, corn, rye, egg tart, pudding, tiramisu, chocolate mousse, ice cream, ice cream cone, sugar cube, syrup, honey, candy, jam, jelly, bubble gum, nectar, topping, vanilla, salt, ginger, dip, vinegar, mustard, salsa, sauce, pickle, turmeric, garlic clove, sesame seed, peanut butter, sugar, cooking oil, peppercorn, butter, cream, yogurt, cheese, yolk, egg white, granola, spam, banana peel, orange peel, lemon peel, charcoal, coke, oil, ember, compound, mixture, fluid, solid, glass, plastic, rock, metal, mineral, building material, waste, fiber, dirt, wood, water, wax, insect repellent, dust, resin, latex, sponge, fur, leather, parchment, lava, foam, ash, business team, fish market, flower market, street market, fruit market, night market, vegetable market, art school, campus, university, fleet, triangle, oval, rectangle, cone, dot, rim, hook, arch, loop, dive, bike racing, marathon, slalom, formula 1, concert, circus, poker]

- Level 7 : [seaweed, stem, stump, root, leaf, fruit, flower, tree root, onion, cane sugar, pumpkin, mushroom, zucchini, yam, lettuce, broccoli, cauliflower, brussels sprouts, daikon, turnip, spinach, cucumber, squash, truffle, bean, pea, beet, potato, cabbage, sprout, tomato, eggplant, asparagus, carrot, sweet potato, bell pepper, okra, celery, mint, basil, parsley, thyme, dill, rosemary, clover, geranium, pepper, butterfly pea, lavender, aloe vera, pitcher plant, artichoke, tobacco, garlic, birch, willow, bonsai, beech tree, palm tree, oak, magnolia, mangrove, maple, holly, eucalyptus tree, fig, rowan, cashew, macadamia, hazel shrub, cypress, pine, fir, cedar, ginkgo tree, cypress tree, elm, hawthorn, banana plant, apple tree, cherry tree, albizia, orange tree, olive tree, mulberry, gliricidia, guava tree, lilac, bird nest, hair, shell, feather, skin, ivory, shoulder, belly, chest, joint, head, neck, face, waist, hand, limb, foot, buttock, brain, bone, hip, heart, worm, starfish, sea cucumber, sea urchin, insect, centipede, jellyfish, coral, scorpion, spider, snail, abalone, conch, huitre, clam, tick, isopod, octopus, crab, lobster, squid, termite, sea anemone, millipede, sea spider, horseshoe crab, slug, frog, salamander, snappers, mackerels, porgies, grouper, catfish, tuna, pufferfish, trout, shark, flatfish, swordfish, seahorse, carp, goldfish, clownfish, seabass, snake, turtle, crocodile, dinosaur, lizard, platypus, crow, parrot, swan, seabird, water bird, hummingbird, ostrich, sparrow, hornbill, turkey, chicken, woodpecker, toucan, eagle, owl, vulture, heron, stork, flamingo, swallow, starling, pigeon, osprey, falcon, bluebird, robin, canary, house finch, blackbird, jay, raven, magpie, pheasant, cuckoo, emu, guinea fowl, downy woodpecker, bantam, hawk, crane, spoonbill, roadrunner, wren, mockingbird, titmouse, hoopoe, kookaburra, marine mammal, panda, bear, raccoon, fox, coyote, wolf, lion, serval, margay, lynx, tiger, caracal, cheetah, jaguar, jaguarundi, hyena, leopard, pet, otter, meerkat, rodent, elephant, armadillo, hedgehog, lemur, monkey, ape, pig, warthog, giraffe, deer, alpaca, donkey, sheep, bison, goat, antelope, cattle, water buffalo, yak, sloth, anteater, skunk, mink, mole, bat, kangaroo, rhinoceros, camel, hippo, koala, horse, buffalo, ferret, badger, opossum, cub, baby, child, teenager, adult, boy, girl, sibling, cousin, partner, grandfather, mother, father, daughter, son, customer, patient, client, passenger, bridesmaid, host, hero, flower girl, commuter, pedestrian, tourist, protester, player, king, queen, director, shepherd, princess, prince, president, entertainer, athlete, warrior, matador, boxer, scientist, engineer, writer, farmer, artist, craftsman, driver, astronaut, student, fisherman, airman, sailor, lifeguard, flight attendant, waiter, maid, soldier, commander, lumberjack, miner, cowboy, servant, quarry worker, coachman, policeman, volunteer, bartender, judge, fireman,

guard, detective, bodyguard, businessperson, referee, gardener, street vendor, bullfighter, chef, samurai, educator, nurse, doctor, trainer, architect, butcher, rescuer, pirate, fashion model, lawyer, coach, author, secretary, nanny, worker, cook, hay worker, electrician, reporter, DJ, motorcyclist, rider, snowboarder, surfer, biker, martial artist, archer, wrestler, ground, water surface, dock, leak, sea, lake, waterfall, gulf, bay, stream, river, strait, pool, waterway, pond, fjord, creek, canyon, pasture, hayfield, swamp, hillside, plateau, volcano, snow mountain, ravine, rock arch, ice cave, islet, archipelago, forest floor, continent, oasis, peninsula, woodland, plain, flat, island, hill, mountain, ridge, cliff, forest, beach, cave, shore, wetland, ski slope, headland, valley, iceberg, glacier, ice floe, dune, reef, peak, grassland, desert, quarry, mine, field, orchard, excavation, rainforest, jungle, blue sky, evening sky, night sky, platform, castle, hospital, clinic, pyramid, auto showroom, bell tower, indoor, monument, mausoleum, tower, obelisk, ruins, skyscraper, roof, stair, home, outbuildings, religious buildings, fortification, white house, prison, hydrant, lighthouse, bridge, dam, tunnel, road, port, water tower, landfill, plaza, telegraph pole, foundation, assembly line, heliport, airfield, station, terminal, wind farm, phone box, factory, workshop, mall, supermarket, office, ticket booth, car wash, spa, ktv, bar, restaurant, casino, sports facilities, dormitory, observatory, library, hotel, resort, theater, office building, rink, cinema, kindergarten, post office, zoo, botanical garden, raceway, golf course, swimming pool, gym, football stadium, post, fort, airport, museum, archive, army base, warehouse, payphone, telephone booth, flagpole, playing field, racetrack, street light, train track, well, mailbox, exit, depository, water tank, exhaust hood, lift, faucet, farm, junkyard, pharmacy, stall, store, laboratory, workplace, studio, laundromat, winery, ball pit, sandbox, fountain, playgrounds, campgrounds, maze, park, amusement park, picnic area, garden, campfire, fire pit, backyard, lawn, swing, window, hedge, paling, wall, fence, barricade, gate, door, trench, hurdle, fender, balustrade, graveyard, parking lot, cage, fireplace, aquarium, corral, forge, balcony, ceiling, sink, cockpit, cabin, basement, loft, urinal, floor, room, shower, office cubicle, capsule, soap, shampoo, toothpaste, cutlery, ice cream scoop, plate, bowl, platter, cup, tea set, chopsticks, decanter, gravy boat, napkin, tablecloth, flower pot, cooler, box, bag, bottle, can, wrapping paper, cardboard, newspaper, graph paper, tissue, notepad, blackboard, pen, crayon, pencil, ink, eraser, binder, bookmark, folder, notebook, whiteboard, clipboard, stylus, toilet paper, handkerchief, towel, toothbrush, perfume, makeup, makeup tool, comb, hairspray, shaving cream, nail polish, mascara, eyeliner, lipstick, face powder, eye shadow, mouthwash, condom, earplug, curling iron, ironing board, sanitary pad, chain, hinge, clasp, knot, pin, scrubbing brush, hairbrush, paintbrush, stilt, ski pole, mast, fishing rod, chainsaw, handsaw, bottle opener, can opener, shaver, shopping basket, flower basket, combination lock, latch, padlock, mail slot, stringed instrument, wind instrument, keyboard instrument, drumstick, drum, maraca, plectrum, barrel organ, stove, colander, peeler, masher, rolling pin, grater, whisk, blender, pan, pot, coffeepot, kettle, grill, reamer, spatula, pizza cutter, measuring cup, ladle, pressure cooker, teapot, hotplate, oven, cutting board, rice cooker, toaster, tongs, ice maker, dishwasher, skewer, electric range, waffle iron, microwave, computer monitor, knob, door handle, fire alarm, fire hydrant, fire hose, fire extinguisher, snowboard, surfboard, yoga mat, bridle, horse blanket, saddle, saddlebag, treadmill, ski, hockey stick, basketball hoop, baseball bat, baseball glove, baseball base, racket, diving board, skate, trampoline, high bar, rings, balance beam, wicket, dumbbell, barbell, weight, javelin, mat, golf equipment, tackle, boxing glove, bodyboard, tennis net, roller skates, basketball backboard, punching bag, parasail, finish line, sideline, target, gymnastics apparatus, cricket bat, snowshoe, battle rope, ball, camera lens, shutter, first-aid kit, sewing kit, carpenter's kit, smartphone, corded phone, microphone, desktop computer, laptop, tablet, server, abacus, steam engine, washing machine, slot machine, vending machine, sewing machine, arcade game, coffee maker, fruit machine, milking machine, stone mill, meat grinder, coffee grinder, water mill, flour mill, ceiling fan, floor fan, gramophone, swab, bandage, sunshade, chandelier, lampshade, oil lamp, floor lamp, droplight, wall lamp, table lamp, light bulb, rear-view mirror, hand mirror, alarm clock, wall clock, digital clock, pocket watch, cabinet, dresser, shelf, bookcase, tv cabinet, wine rack, coat rack, drawer, chair, couch, stool, bench, baby seat, stroller, billiard table, kitchen table, counter, desk, workbench, dining table, table tennis table, dressing table, conference table, operating table, glass table, nightstand, altar, bed, bedclothes, mattress, carpet, pillow, cradle, headboard, fabric, cloth, headband, cap, hat, helmet, crown, tiara, wig, sleepwear, sleeve, uniform, leotard, wet suit, swimwear, veil, leggings, vest, suit, protective suit, bib, pants, diaper, robe, shirt, sweater, scarf, cloak, coat, overall, sportswear, dress,

underpants, cheongsam, hanfu, baby clothes, crop top, jumpsuit, underclothes, hoodie, wetsuit, blouse, waistband, hood, costume, coverall, apron, stocking, tights, socks, shoe, bus station, subway station, traffic cone, bus stop, railroad, spacecraft, vehicle structure, bus, train, stretcher, dolly, rocket, escalator, sleigh, construction vehicle, vessel, tank, unicycle, skateboard, bicycle, scooter, baby carriage, tricycle, railcar, cart, wagon, car, ambulance, snowplow, golf cart, motorcycle, truck, go-kart, tractor, forklift, ski lift, chairlift, plane, drone, helicopter, glider, balloon, airship, snowmobile, atv, self-balancing scooter, pickup truck, skibob, horse-drawn vehicle, trailer, cable car, elevator, subway, segway, eye mask, bolo tie, goggles, sunglasses, pendant, necklace, cufflinks, bracelet, anklet, pearl, jewel, earrings, ring, bead, brooch, fireworks, christmas decoration, jack-o'-lantern, snowman, chinese knot, gift, lantern, confetti, wreath, bouquet, rifle, handgun, machine gun, revolver, paintball gun, scimitar, broadsword, rapier, dagger, machete, crossbow, halberd, mortar, gas mask, extinguisher, armor, shield, blueprint, receipt, bulletin board, billboard, menu, gift card, recipe, placard, scoreboard, map, journal, sticker, poster, list, ticket, document, sheet music, handwriting, postcard, bible, fiction, poetry, letter, checkbook, passbook, greeting card, manuscript, publication, magazine, newspapers, book, record, album, cassette, dvd, videotape, fax, reel, sign, signal, icon, avatar, mark, award, design, identity card, credit card, keycard, business card, qr code, barcode, clip art, tattoo, ceramic, fictional character, photography, sculpture, dance, painting, street art, drawing, mosaic, portrait, medical image, photo, illustration, graph, picture, diagram, advertisement, news, press conference, notice, movie, auditory expression, drama, puppet, playhouse, lego, doll, teddy, toy car, toy plane, toy gun, piggy bank, mahjong, chess, game board, playing card, currency unit, currency form, chocolate cake, cupcake, pancake, tortilla, birthday cake, waffle, wedding cake, doughnut, brownie, fruitcake, mooncake, crouton, toast, bagel, croissant, pretzel, muffin, biscuit, baguette, steamed bread, bun, cinnamon roll, profiterole, eel, mussel, scallop, clam meat, lobster meat, oyster, shrimp, crabmeat, whelk, steak, haystack, liquor, wine, beer, cocktail, whisky, orange juice, chocolate milk, ice water, bottled water, broth, tomato soup, hamburger, hot dog, submarine sandwich, easter egg, candy bar, fudge, candy cane, lollipop, cotton candy, caramel, jelly bean, mint candy, whipped cream, frosting, tomato sauce, hot sauce, gravy, gel, gas, powder, crystal, foil, copper, silver, gold, steel, bronze, brass, turquoise, quartz, ruby, sapphire, emerald, gravel, granite, marble, plaster, beam, plank, tile, brick, concrete, cement, tarmac, block, garbage, yarn, wool, cotton, straw, flax, coir, raffia, silk, string, twine, sand, clay, mud, lumber, bamboo, circle, square]

- Level 8 : [trunk, branch, sugarcane, greenery, banana leaf, maple leaf, tobacco leaves, pod, acorn, cherry, peach, olive, pear, apple, seed, raisin, grape, lychee, passion fruit, apricot, citrus fruit, mango, melon, berry, avocado, pomegranate, banana, kiwi, hami melon, dragon fruit, starfruit, nectarine, papaya, prickly pear, mangosteen, plum, durian, pineapple, jackfruit, jujube, calabash, water lily, lotus, iris, daffodil, tulip, lily, marigold, daisy, poppy, anemone, dahlia, petunia, orchid, sunflower, chrysanthemum, peony, carnation, cherry blossom, chinese rose, clivia, hibiscus flower, blossom, rose, rhododendron, fuchsia, poinsettia, oleander, heather, hydrangea, hibiscus, blueweed, violet, thistle, pansy, morning glory, livingstone daisy, dandelion, jasmine, squill, coconut palm, hairstyle, mane, beard, lash, seashell, ankle, elbow, wrist, knee, eye, nose, ear, mouth, finger, arm, leg, lap, toe, barefoot, planarian, leech, pest, larva, beetle, flea, bee, wasp, ant, cricket, grasshopper, stick insect, praying mantis, dragonfly, moth, butterfly, earwig, vespa, hermit crab, king crab, tadpole, tree frog, sardine, whale shark, stingray, koi, python, iguana, gecko, macaw, cockatoo, African grey parrot, budgie, cob, black swan, pelican, cormorant, penguin, gull, auk, mollymawk, goose, duck, hen, bald eagle, barn owl, egret, quail, partridge, peacock, ruffed grouse, seal, walrus, whale, manatee, bear cub, polar bear, brown bear, cougar, mountain lion, snow leopard, cat, dog, rat, squirrel, hamster, mouse, beaver, guinea pig, rabbit, mammoth, baby elephant, squirrel monkey, macaque, baboon, chimpanzee, orangutan, gorilla, piglet, fawn, roe deer, reindeer, moose, lamb, wildebeest, calf, bull, cow, zebra, pony, mule, man, woman, elder, twin, newlywed, wife, husband, magician, musician, joker, dancer, actor, juggler, snake charmer, strongman, climber, hockey player, football player, skater, skateboarder, runner, swimmer, skier, basketball player, baseball player, gymnast, tennis player, rock climber, goalkeeper, motorcycle racer, rugby player, soccer goalkeeper, soccer player, weightlifter, poet, journalist, beekeeper, photographer, sculptor,

painter, conductor, makeup artist, welder, construction worker, plumber, barber, carpenter, blacksmith, mechanic, builder, stonecutter, truck driver, bus driver, taxi driver, racing driver, pilot, police officer, teacher, professor, dentist, surgeon, vet, snowfield, lakeshore, coast, river bank, wheat field, corn field, flower field, rice field, sunflower field, farmland, podium, catwalk, scaffold, deck, stage, boxing ring, gravestone, chinese tower, eiffel tower, dome, tile roof, manor house, log cabin, yurt, house, parking space, beehive, tree house, flower bed, shelter, hangar, courtyard, terrace, pavilion, stable, chicken coop, barn, garage, outhouse, conservatory, shed, dog house, shrine, monastery, church tower, chapel, mosque, temple, church, pulpit, suspension bridge, tower bridge, asphalt road, promenade, sidewalk, curb, street, highway, driveway, railway line, crosswalk, intersection, canal, trail, alley, bike lane, path, crossroad, police station, power station, fire station, gas station, broadcasting station, bowling alley, stadium, bullring, sports field, airport runway, fire escape, silo, treasury, storehouse, repair shop, salon, newsstand, pet store, flower shop, bakery, chemistry lab, music studio, recording studio, art studio, slide, amusement ride, water park, display window, skylight, floor window, city wall, glass wall, climbing wall, seawall, doorplate, screen door, car door, elevator door, glass door, glass floor, tile flooring, wood floor, suite, hotel room, storage room, bedroom, dining room, classroom, hospital room, jail cell, hall, recreation room, gallery, bathroom, engine room, dressing room, kitchen, restroom, sauna, locker, computer room, meeting room, dance studio, hotel lobby, laundry room, toilet bowl, fork, spoon, knife, paper plate, glass plate, glass bowl, drinking glass, teacup, chalice, paper cup, mug, shipping container, carton, pencil case, crate, safe, toolbox, gift box, computer case, briefcase, luggage, shopping bag, backpack, tote bag, sleeping bag, sack, duffel, handbag, shoulder bag, clutch bag, gift bag, paper bag, jug, flask, thermos, gourd, glass bottle, baby bottle, watering can, spray can, fountain pen, sharpie, board eraser, makeup brush, makeup mirror, hair roller, thumbtack, zither, harp, guitar, violin, cello, erhu, recorder, sax, panpipes, flute, bugle, trombone, tuba, whistle, kazoo, trumpet, mouth organ, musical keyboard, piano, harpsichord, accordion, cymbal, snare drum, tambourine, gas stove, wood-burning stove, frying pan, wok, roaster, saucepot, dutch oven, saucepan, water ski, badminton racket, tennis racket, ping pong paddle, roller skate, golf ball, golf club, rugby ball, cricket ball, bowling ball, basketball ball, soccer ball, baseball ball, tennis ball, volleyball ball, billiard ball, adhesive bandage, bookshelf, supermarket shelf, folding chair, rocking chair, swivel chair, feeding chair, wheelchair, throne, barber chair, office chair, sofa, daybed, footrest, folding stool, bar stool, hassock, pew, window seat, park bench, training bench, bunk bed, berth, hammock, hospital bed, dog bed, double bed, single bed, waterbed, canopy bed, crib, quilt, blanket, bedcover, sheet, throw pillow, patch, shower cap, beret, skullcap, swim cap, visor, straw hat, dress hat, witch hat, sun hat, cowboy hat, fedora, baseball cap, party hat, top hat, christmas hat, football helmet, bicycle helmet, safety helmet, racing helmet, military uniform, school uniform, bikini, safety vest, jeans, shorts, sweatpants, kimono, bathrobe, dress shirt, polo shirt, t-shirt, sweatshirt, jersey, cardigan, turtleneck, headscarf, poncho, shawl, jacket, raincoat, overcoat, lab coat, fur coat, sport coat, trench coat, skirt, sari, gown, kilt, lace dress, evening dress, maxi dress, bandeau, corset, cosplay, halloween costume, boot, sandal, running shoe, high heel, skiing shoes, leather shoe, slipper, flip-flop, satellite, space station, space shuttle, windshield, taillight, wheel, tire, steering wheel, seat belt, dashboard, car seat, car trunk, school bus, double-decker bus, train car, locomotive, steam train, raft, yacht, boat, sailboat, passenger ship, warship, pirate ship, cargo ship, mountain bike, exercise bike, caboose, freight car, passenger car, shopping cart, wheelbarrow, hand truck, sports car, taxi, racing car, limo, convertible, jeep, sedan, police car, lorry, tow truck, fire truck, garbage truck, food truck, transporter, police van, bulldozer, biplane, jet, airliner, seaplane, cargo aircraft, bomber, hot air balloon, elevator car, diamond, glass bead, sparkler, firecracker, christmas light, christmas tree, christmas ball, paper lantern, halloween pumpkin, chinese lantern, bulletproof vest, chainmail, checklist, movie ticket, certificate, passport, license, contract, bill, work card, comic book, paperback book, hardback book, atlas, textbook, dictionary, file, register, note, cd, symbol, alarm, armband, air sock, brake light, traffic light, buoy, milestone, label, earmark, watermark, medal, trophy, architecture, pattern, emblem, flag, cross, logo, license plate, letter logo, badge, pottery, porcelain, cartoon character, filming, face close-up, figurine, bronze sculpture, ice sculpture, sand sculpture, statue, ballet, folk dance, mural, oil painting, watercolor painting, portrait painting, monochrome, doodle, sketch, manga, id photo, wedding photo, family photo, selfie, cartoon illustration, fashion illustration, cartoon, documentary, horror film, science

fiction film, animated film, music, audio, laugh, conversation, argument, opera, comedy, play, musical, barbie, water gun, chessboard, dollar, pound sterling, rmb, coin, paper money, check, champagne, sake, steam, ice, snowflake, amethyst, citrine]

- Level 9 : [nut, coffee bean, pumpkin seed, sunflower seed, orange, lime, grapefruit, lemon, watermelon, winter melon, muskmelon, cantaloupe, persimmon, strawberry, blueberry, blackberry, bayberry, raspberry, cranberry, bud, bald, hair color, blonde hair, brown hair, gray hair, red hair, white hair, black hair, curly hair, purple hair, blue hair, green hair, orange hair, pink hair, yellow hair, straight hair, braided hair, ponytail, long hair, short hair, braid, eyebrow, eyelash, tear, beak, fly, mosquito, roach, locust, caterpillar, maggot, mealworm, nymph, ladybird, duckling, wood duck, sea lion, dolphin, killer whale, humpback whale, egyptian cat, kitten, american shorthair, persian cat, british shorthair, burmese cat, ocelot, domestic cat, siamese, abyssinian, leopard cat, puppy, german shepherd, labrador, poodle, chihuahua, dachshund, rottweiler, doberman pinscher, husky, beagle, boxer dog, bulldog, german shorthair pointer, great dane, afghan hound, schnauzer, rhodesian ridgeback, sheepdog, gordon setter, setter, shih-tzu, weimaraner, tibetan mastiff, maltese, basset, malamute, dalmatian, samoyed, ibizan hound, shiba inu, golden retriever, chow chow, bichon frise, corgi, foxhound, hound, saint bernard, bernese mountain dog, pit bull, bride, bassist, guitarist, rocker, violinist, pianist, saxophonist, violist, singer, drummer, ballet dancer, child actor, ice hockey player, diver, catcher, backdrop, apartment, cottage, farmhouse, villa, palace, tent, hut, igloo, zebra crossing, television studio, football field, basketball court, volleyball court, tennis court, granary, ammunition storage, carousel, ferris wheel, pantry, vault, wine cellar, closet, operating room, concert hall, banquet hall, conference hall, auditorium, art gallery, soup spoon, champagne flute, guzheng, denim jacket, leather jacket, miniskirt, ballet skirt, pencil skirt, wedding dress, fishing boat, tugboat, motorboat, ferry, canoe, rowboat, lifeboat, inflatable boat, swan boat, sail, cruise ship, aircraft carrier, cruiser, battleship, submarine, container ship, fighter jet, diploma, traffic sign, stop sign, speed limit sign, parking sign, road sign, warning sign, welcome sign, written symbol, bronze medal, gold medal, silver medal, stripe, banner, pirate flag, national flag, white flag, red flag, pop, jazz, disco]

- Level 10 : [pine cone, peanut, macadamia nut, hazelnut, coconut, chestnut, walnut, rapper, lead singer, ballerina, font, alphabet, check mark, dollar sign, character, percent sign, tricolor, german flag, south korean flag, italian flag, japanese flag, canada flag, brazilian flag, chinese flag, indian flag, mexican flag, australian flag, russian flag, south african flag, argentine flag, union jack, american flag]

B.2 METRICS

We employed four metrics to evaluate the predicted hierarchical structures.

- **Tree Edit Distance** TED: Measures the minimum number of edit operations (insertion, deletion, etc.) required to transform the predicted tag tree into the ground-truth tag tree.
- **Jaccard Similarity** J: Treats the set of nodes in the predicted and ground-truth trees as two separate sets, and computes the ratio of their intersection to their union:

$$\mathrm{J} = \frac{|\text{Predicted Nodes} \cap \text{Ground-truth Nodes}|}{|\text{Predicted Nodes} \cup \text{Ground-truth Nodes}|}. \tag{12}$$

- **Hierarchical Precision** P_H: Represents the proportion of correctly predicted nodes among all predicted nodes:

$$\mathrm{P_H} = \frac{|\text{Predicted Nodes} \cap \text{Ground-truth Nodes}|}{|\text{Predicted Nodes}|}. \tag{13}$$

- **Hierarchical Recall** R_H: Represents the proportion of correctly predicted nodes among all ground-truth nodes:

$$\mathrm{R_H} = \frac{|\text{Predicted Nodes} \cap \text{Ground-truth Nodes}|}{|\text{Ground-truth Nodes}|}. \tag{14}$$

B.3 Description Generation

To enrich each tag's semantics, we provide **6** concise visual descriptions:

- **1 CLIP-style description**: the fixed template `a photo of a {tag}`.
- **5 LLM descriptions**: We prepend the tag's curated definition so the model understands its intended meaning. The prompt details are as follows.

```
    llm_prompts = [
        f"Describe concisely what a(n) {tag} looks like:",
        f"How can you identify a(n) {tag} concisely?",
        f"What does a(n) {tag} look like concisely?",
        f"What are the identifying characteristics of a(n) {tag}:",
        f"Please provide a concise description of the visual
                characteristics of {tag}:"
    ]
    ...
    user_prompt = (
        f"Here is the intended meaning of '{tag}':
    {definition}\n\n"
        f"{llm_prompt}"
    )
```

B.4 Data Annotation

We use the following prompts to annotate image with corresponding tags:

```
system_prompt :
    "You are an image classifier."# Output only a valid Python
    list of recognized category names, and nothing else."
user prompt:
    "You need to find which categories exist in this image and
    only output a plain list of category names for the current
    image based on following rules:"
        "1. You need to find which categories exist in this image
    based on the categories and their corresponding definitions in
    the dict {cate_dict}."
        "2. Categories should strictly reflect visible content
    without interpretation or assumption. Choose the most
    appropriate categories for current image from the list and no
    more than 30 categories."
        "3. Only output a list of category names for the image
    without any extra text (e.g. explanations, reasoning). For
    example, the final output should look like:[class1, class2, ...]
```

C Experiment

C.1 Quality comparison of different image tagging methods

In Figure 10, we present a comparison between our method HiTag and other image tagging approaches. The results show that HiTag can recognize a richer set of hierarchical tags.

C.2 Ablation study on entailment weight

We compared the impact of different entailment weights on performance. Based on the results of Figure 11, we conclude that 10 is an appropriate value for entailment weight.

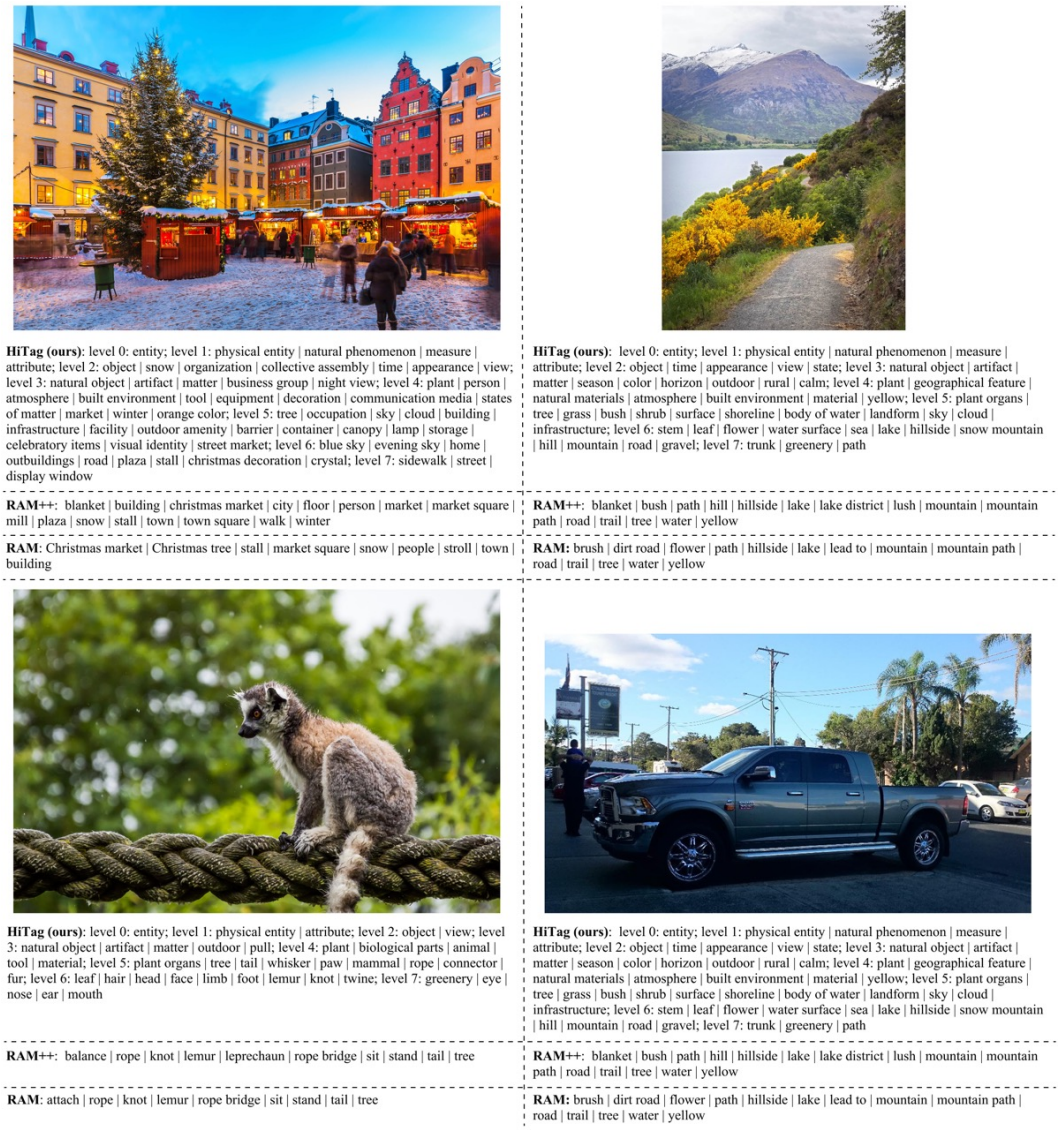

Figure 10: Comparison of Tag Recognition Performance with Image Tagging Methods

C.3 Model Performance With Domain Transfer

In Table6, we present a comparison between our model and the baseline model RAM++ after domain transfer on the COCO dataset.

Table 6: **Comparison of Model Performance With Domain Transfer.** The following datasets are all evaluated under the open-vocabulary setting.

Methods	Training Datasets	Fine-tuning Datasets	COCO
RAM++	CC3M	COCO	75.5
HiTag (ours)	CC3M	COCO	**79.5**

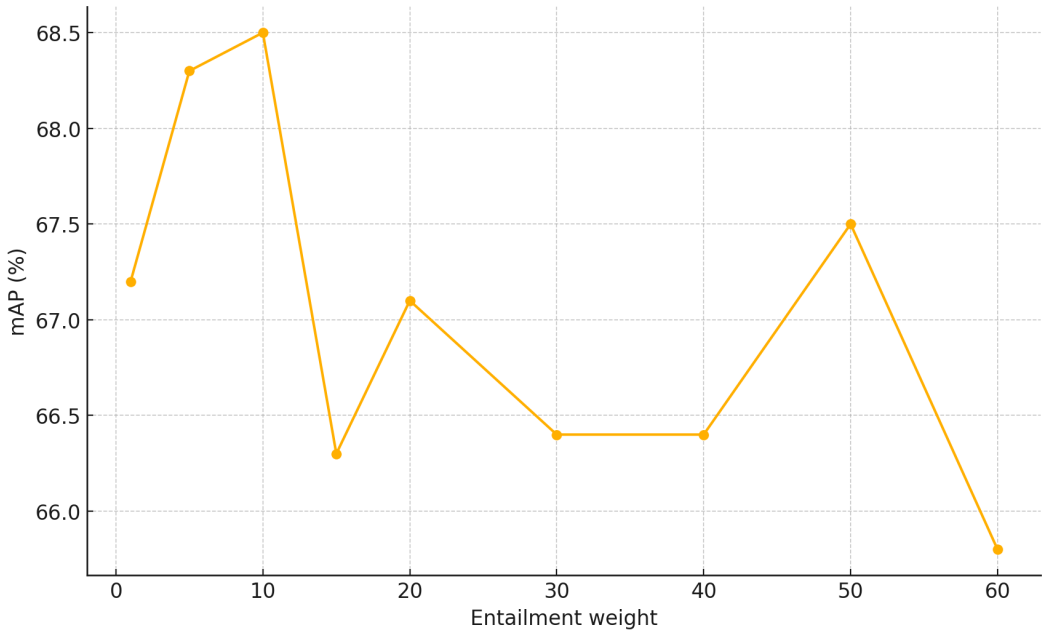

Figure 11: Performance on OpenImages-Uncommon at Different Entailment Weights

D METHOD DETAILS

D.1 HYPERBOLIC SPACE

Hyperbolic space's geodesics exhibit tree-like branching, making it well-suited for hierarchical structures that expand exponentially at each level. Moreover, even low-dimensional hyperbolic spaces can simultaneously capture hierarchical relationships and node similarities, providing a natural and efficient geometric paradigm for representing and learning from hierarchical data. In practice, the Poincaré ball model and the Lorentz model are two commonly used isometric models for hyperbolic spaces. Because the Poincaré ball is bounded by the unit sphere, gradients tend to vanish near the boundary (i.e., as $r \to 1$), which affects training stability. In contrast, the Lorentz model is unbounded, and its exponential/logarithmic maps are more numerically stable, making it more suitable for deep learning.

In the Lorentz model, geodesics are the shortest paths between points. For any $\mathbf{x}, \mathbf{y} \in L^n$, the two-dimensional plane containing $\{\mathbf{x}, \mathbf{y}\}$ and the origin intersects L^n along a unique geodesic. The Lorentz distance is the arc length of this geodesic:

$$d_L(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{k}} \operatorname{arcosh}\big(-k \langle \mathbf{x}, \mathbf{y} \rangle_L\big). \tag{15}$$

D.2 IMAGE-TEXT ALIGNMENT

We define a tag-image attention to model the interaction between tags and images. Specifically, let Q denote the tag features $\mathbf{x}^{(\text{tag})}$, and let K and V represent the image features. The tag-image attention aggregates relevant evidence vectors from the image for each category, which can be formulated as:

$$F_{i \to tag} = \operatorname{softmax}\Big(\frac{\operatorname{Log}(Q)\operatorname{Log}(K)^T}{\sqrt{d}}\Big)\operatorname{Log}(V), \tag{16}$$

where $F_{i \to tag}$ is the tag–image similarity feature encoding the interaction between the tag and the image, and d is a normalization constant. For computational simplicity, we map the features into Euclidean space via the $\operatorname{Log}_o(\cdot)$ function according to Eq equation 4. Furthermore, $F_{i \to tag}$ is passed through a linear layer to obtain the score of each tag, followed by a sigmoid function to obtain the

final prediction probability p_{tag}:

$$p_{tag} = \sigma\Big(\mathbf{W}\, F_{i \to tag} + b\Big), \tag{17}$$

where $\mathbf{W}$ and b are the weight matrix and bias term of the linear layer, respectively, and $\sigma(\cdot)$ denotes the sigmoid function.

Following Ben-Baruch et al. (2020), we employ Asymmetric Loss to compute the tag-image alignment loss $\mathcal{L}_{tag_{align}}$. Likewise, the caption-image alignment loss $\mathcal{L}_{cap_{align}}$ is computed in the same manner. Therefore, the overall image-text alignment loss $\mathcal{L}_{\text{align}}$ is defined as:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{tag}_{\text{align}}} + \mathcal{L}_{\text{cap}_{\text{align}}} \tag{18}$$