Explaining, Verifying, and Aligning Semantic Hierarchies in Vision-Language Model Embeddings

Gesina Schwalbe[1,2], **Mert Keser**[3,4], **Moritz Bayerkuhnlein**[2], **Edgar Heinert**[5], **Annika Mütze**[5], **Marvin Keller**[2], **Sparsh Tiwari**[2], **Georgii Mikriukov**[6], **Diedrich Wolter**[2], **Jae Hee Lee**[7] and **Matthias Rottmann**[5]
[1]Ulm University, Germany
[2]University of Lübeck, Germany
[3]Technical University of Munich, Germany
[4]AUMOVIO SE, Germany
[5]Osnabrück University, Germany
[6]Leibniz Institute for Agricultural Engineering and Bioeconomy, Germany
[7]University of Hamburg, Germany
gesina.schwalbe@uni-ulm.de, {moritz.bayerkuhnlein, sparsh.tiwari, diedrich.wolter}@uni-luebeck.de, mert.keser@aumovio.com, {edgar.heinert, annika.muetze, matthias.rottmann}@uni-osnabrueck.de, marvin.keller@student.uni-luebeck.de, gmikriukov@atb-potsdam.de, jae.hee.lee@uni-hamburg.de

Abstract

Vision-language model (VLM) encoders such as CLIP enable strong retrieval and zero-shot classification in a shared image–text embedding space, yet the semantic organization of this space is rarely inspected. We present a post-hoc framework to *explain*, *verify*, and *align* the semantic hierarchies induced by a VLM over a given set of child classes. First, we extract a binary hierarchy by agglomerative clustering of class centroids and name internal nodes by dictionary-based matching to a concept bank. Second, we quantify plausibility by comparing the extracted tree against human ontologies using efficient tree- and edge-level consistency measures, and we evaluate utility via explainable hierarchical tree-traversal inference with uncertainty-aware early stopping (UAES). Third, we propose an ontology-guided post-hoc alignment method that learns a lightweight embedding-space transformation, using UMAP to generate target neighborhoods from a desired hierarchy. Across 13 pretrained VLMs and 4 image datasets, our method finds systematic modality differences: image encoders are more discriminative, while text encoders induce hierarchies that better match human taxonomies. Overall, the results reveal a persistent trade-off between zero-shot accuracy and ontological plausibility and suggest practical routes to improve semantic alignment in shared embedding spaces.

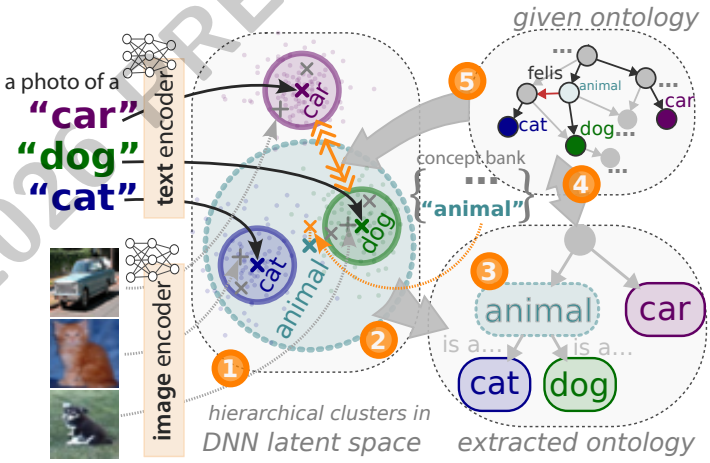


Figure 1: Our post-hoc framework to **explain** (embed 1 given concepts, cluster 2 and name found parents 3), **verify** 4, and **align** 5 the semantic hierarchy induced by a VLM in its embedding space.

1 Introduction

Contrastive vision-language models (VLMs) such as CLIP Radford et al. [2021] demonstrate remarkable capabilities in learning joint representations of images and text [Barraco *et al.*, 2022]. These representations are adopted in various applications, including image retrieval [Baldrati *et al.*, 2022], captioning [Barraco *et al.*, 2022], and zero-shot classification [Radford *et al.*, 2021]. However, their evaluation is still dominated by task-level performance metrics such as classification accuracy that provide little insight into how concepts are internally organized and related. In particular, understanding the semantic

[1]Find supplementary material at: https://arxiv.org/abs/2603.26798
[2]Find code at: https://github.com/gesina/ontological-commitment

hierarchies of these representations has not been thoroughly explored [Lee *et al.*, 2025].

Semantic hierarchies capture how concepts relate via commonalities and differences, and are central to human categorization and symbolic knowledge representation via ontologies [Confalonieri and Guizzardi, 2025; Guarino, 1998]. Extracting the hierarchy induced by a VLM is crucial for explainability [Bhalla *et al.*, 2024; Alper and Averbuch-Elor, 2025; Lee *et al.*, 2025; Wan *et al.*, 2020]. It further enables identifying unintuitive or biased groupings (e.g., grouping `man` and `woman` primarily via `hair` rather than under `person`) and aligning learned similarities with human knowledge [Ribeiro and Leite, 2021; Confalonieri and Guizzardi, 2025]. Lastly, hierarchies provide an inductive bias for more robust and informative classification [Bertinetto *et al.*, 2020; Dhall *et al.*, 2020; Ge *et al.*, 2023], such as explainable hierarchical tree-traversal [Wan *et al.*, 2020].

Meanwhile, key desiderata formulated for VLM embedding spaces are text-to-image alignment, and that classes can be *discriminated* via zero-shot classification. We complement this by two new desiderata: Text and image encoders should also align in their induced hierarchies; and they should be *plausible*, i.e., the hierarchy should match human taxonomies. Note that plausibility and discriminability may diverge, since embedding similarities may accurately discriminate class members, but still violate human expectations.

Into that direction, prior work has explored the interpretability of VLM embeddings [Bhalla *et al.*, 2024] (*Which concepts are encoded?*) and found and reinforced pre-defined semantic hierarchies in them [Alper and Averbuch-Elor, 2025] (*Is a given hierarchy encoded?*). However, there is a lack of systematic methods to answer the following questions for a trained VLM encoder, along which we organize our pipeline (cf. Figure 1):

Q1: Can we **explain** *which semantic hierarchy* is induced by an encoder for a given set of leaf classes? How faithful and reliable are local explanations via explainable hierarchical inference? (Secs. 4.1, 4.2 and 5.2)

Q2: Can we **verify** the plausibility of induced hierarchies relative to reference ontologies? How do they vary across modalities, models, and hierarchy depths? (Secs. 4.3 and 5.3)

Q3: Can we post-hoc **align** an induced hierarchy to a target similarity structure without sacrificing zero-shot accuracy? (Secs. 4.4 and 5.4)

Contributions. We demonstrate the utility of semantic hierarchy extraction:

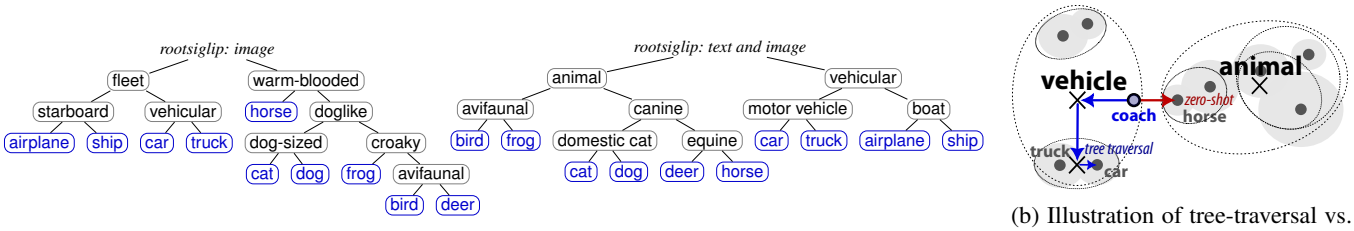
- We propose a post-hoc pipeline to explain, verify, and align semantic hierarchies in VLM embedding spaces. We extract a hierarchy via hierarchical clustering of leaf classes and dictionary-based parent naming, verify it with efficient ontology-based comparison metrics, and align it using UMAP-guided target neighborhoods and a lightweight embedding transformation.
- We empirically study 13 pretrained VLMs across 4 datasets and multiple common-sense ontologies, enabling systematic comparison of encoder semantics.
- We find a consistent modality gap: image encoders yield higher zero-shot accuracy, while text encoders induce more plausible hierarchies; and enriching hierarchical inference with UAES improves semantic correctness, bridging the explainability-faithfulness trade-off.

2 Related Work

Explaining Learned Concepts. The field of concept-based XAI (c-XAI) has evolved around the goal to associate vectors in the latent representations with semantic concepts [Lee *et al.*, 2025]. Such associations have found manifold applications in the inspection of a DNN's learned ontology, such as: which concepts from a given ontology are learned? [Ribeiro and Leite, 2021], how similar are pairs of concepts? [Fong and Vedaldi, 2018; Schwalbe, 2021], and how are concepts from different layers or models correlated? [Kim *et al.*, 2018; Mikriukov *et al.*, 2023]. Notably, we here build upon the approach by Yuksekgonul et al. [2022] to name parent embeddings. They utilize the text-to-image alignment in VLMs to match vectors to the closest text embeddings from a dictionary. To go from concepts to hierarchical concept relations, Wan et al. [2020] employed hierarchical clustering on latent representations, but unlike our approach restricted to the last layer of a classifier. And despite rising interest in verification and steering of DNNs against ontologies [Ribeiro and Leite, 2021; Ponomarev and Agafonov, 2022; Badreddine *et al.*, 2022; Alper and Averbuch-Elor, 2025], it remains largely unexplored how to post-hoc explain and control the learned semantic similarity structure within DNN embedding spaces [Confalonieri and Guizzardi, 2025].

Understanding VLM Embedding Spaces. The CLIP-based learning scheme [Radford *et al.*, 2021; Jia *et al.*, 2021] of concurrent VLMs have been shown to emerge deep concept understanding [Bhalla *et al.*, 2024; Zhao *et al.*, 2025], and even class hierarchies [Alper and Averbuch-Elor, 2025; Ge *et al.*, 2023]—though existing hierarchy extraction techniques rely on costly parent supervision and don't align with ontology concept dictionaries. Moreover, previous findings that integrating hierarchy information into the inference process makes classification more explainable [Wan *et al.*, 2020] and reliable [Bertinetto *et al.*, 2020; Dhall *et al.*, 2020], were shown to transfer to VLMs [Novack *et al.*, 2023]. We therefore adapt the hierarchy tree-traversal approach from Wan et al. [2020] for VLM classification, newly adding early stopping. Despite these impressive properties, the modality alignment in CLIP embedding spaces remains suboptimal, mapping text and image to distinct latent regions [Liang *et al.*, 2022; Eslami and de Melo, 2024], and requiring prompt finetuning to boost zero-shot performance [Radford *et al.*, 2021; Ma *et al.*, 2025; Ge *et al.*, 2023]. Other than prior studies, we here for the first time investigate the text-to-image alignment with respect to the primarily learned semantic hierarchies of the encoders.

Ontologies and Ontology Matching. An ontology reflects a perspective on a domain [Guarino, 1998], formalized as concepts, relations, and assumptions thereon [Guarino, 1998]. By now a variety of both specialized and general domain



(a) Comparison of two extracted hierarchies over the CIFAR-10 leaves (shown in blue) from the SigLIP model, using image-only leaf embeddings (left) versus combined text&image leaf embeddings (right).

(b) Illustration of tree-traversal vs. zero-shot inference in embedding space of the text & image tree.

Figure 2: Running toy example on CIFAR-10 leaf classes.

ontologies are available [Confalonieri and Guizzardi, 2025; Mascardi *et al.*, 2007]. The latter include closed-source ones [Lenat, 1989; Forbus and Hinrich, 2017], large-scale but less principled ones like ConceptNet [Speer *et al.*, 2017] or DBPedia [Lehmann *et al.*, 2015]. We here use the well-established, large, general domain ontologies SUMO [Niles and Pease, 2001], OpenCyc [Matuszek *et al.*, 2006], and Yago [Mahdisoltani *et al.*, 2015] as reference for human knowledge on plausible hierarchies. Ontology matching techniques aim to compare or integrate divergent ontologies [Otero-Cerdeira *et al.*, 2015]. A common distance measure for graphs is *edit distance*, which transforms one instance into another via minimal edit operations [Zhang and Shasha, 1989; Akutsu, 2010]. For our special case of semantic hierarchy trees, tree edit distance measures can be applied [Zhang *et al.*, 1992]. Aside from this global comparison, we here also aggregate local edge-wise scores based on matching lowest common ancestors (LCAs) [Bender *et al.*, 2005]. However, existing works have not yet used ontology matching techniques to assess DNN-learned ontologies.

3 Background

The goal of our approach is to discover a learned semantic hierarchy, constituted of latent embeddings of parent concepts. In the following we formalize semantic hierarchies, and recap the classical techniques for concept discovery as well as underlying key assumptions for hierarchy extraction.

3.1 Ontologies

A semantic hierarchy is a slice of a broader ontology.

Definition 1 (Ontology)**.** *An* ontology $\mathcal{O}$ *is a description of a domain by a set of class membership functions (the* concepts*), instance* relations*, and a logic theory of assumptions thereon. [Guarino, 1998]*

Semantic similarity is an example for relations between instances. This paper only considers taxonomic *is-a* assumptions on concepts (hypernymy/subclass relations [Fellbaum, 2010]) for verification and alignment. We denote these as $\texttt{IsParentOf}(\texttt{P},\texttt{C}) := (\forall v\colon \texttt{C}(v) \Rightarrow \texttt{P}(v))$, meaning that instances of C are also instances of P (e.g., $\texttt{cat} \Rightarrow \texttt{animal}$). A semantic hierarchy is a tree formed by such relations; our goal is to extract such a hierarchy from a VLM's embedding space.

3.2 Assumed Properties of VLM Latent Spaces

We rely on a few high-level assumptions that are standard in concept-based interpretability [Lee *et al.*, 2025], where often single vectors are used as prototypes for concepts [Kim *et al.*, 2018; Fong and Vedaldi, 2018] (Ass. 1); and in the typical VLM training and zero-shot inference schemes which employ cosine similarity as a proxy for semantic similarity of latent representations (Ass. 2). These assumptions are formalized in the following.

First, we simplify the representation of a visual concept from its cluster of concept instance embeddings to a point in embedding space.

Assumption 1 (Concept Embeddability)**.** *A human-interpretable visual concept* C *can be represented by a point* $e(\texttt{C})$*, its* concept embedding*, in the VLM embedding space [Kim* et al.*, 2018; Fong and Vedaldi, 2018].*

Notable work introducing and validating this modeling scheme of concepts as single vectors were the concept activation vectors by Kim et al. [2018] and concept embeddings by Fong et al. [2018]. More recently, Mikriukov et al. [2025] confirmed that despite the more distributed nature of concept instance representations, single vectors are usually a good approximation.

Second, cosine similarity between concept embeddings is a reasonable proxy for semantic relatedness, which is our key to find parent embeddings. This also is the same assumption used by zero-shot classification where we assign x to the closest concept embedding under cosine distance $d(u,v) := 1 - \frac{u \cdot v^T}{\|u\| \cdot \|v\|}$ in embedding space [Radford *et al.*, 2021].

Assumption 2 (Embedding Interpolatability)**.** *A concept* C *is more similar to* $\texttt{C}'$ *than to* $\texttt{C}''$ *iff* $d(e(\texttt{C}), e(\texttt{C}')) \leq d(e(\texttt{C}), e(\texttt{C}''))$ *[Fong and Vedaldi, 2018; Ghorbani* et al.*, 2019; Posada-Moreno* et al.*, 2024].*

This interpolatability property was empirically confirmed for concept embeddings both in language [Mikolov *et al.*, 2013] and vision model latent spaces [Fong and Vedaldi, 2018; Ghorbani *et al.*, 2019; Posada-Moreno *et al.*, 2024].

If cosine similarity is used to assign an instance's embedding v to a concept C ($\texttt{C}(v)$ from above), a parent concept for given children concepts is the one most similar to all instances of its children simultaneously. Hence, a direct consequence of Ass. 2 is to assume that DNNs learn hierarchies in terms of similarity structures: Concept vectors of parent

concepts can be approximated by aggregating their children's concept vectors (e.g., by averaging [Mikriukov *et al.*, 2025; Wan *et al.*, 2020]), which motivates our use of agglomerative clustering.

4 Methodology

This section introduces our pipeline to *explain* the semantic hierarchy induced by a VLM as binary tree of concepts (Sec. 4.1) and via enhanced explainable hierarchical classification (Sec. 4.2), which allows to **verify** the VLM plausibility against references (Sec. 4.3), and semantically **align** its embedding space to a target similarity structure (Sec. 4.4).

Example 1. *To make the pipeline concrete, we use the CIFAR-10 leaf classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) as a running example throughout the paper. Figure 2a shows two hierarchies extracted for the same ten leaves from SigLIP, and foreshadows the modality-dependent differences we quantify later (Sec. 5.3).*

Experimental Setup and Notation. We assume a trained VLM encoder F and a dataset of leaf classes (e.g., CIFAR-10 [Alex, 2009]) with train/test splits. Leaf concepts can be represented via images, text descriptions (e.g., LLM-generated prompts [Ma *et al.*, 2025]), or both. We further assume a concept bank $\mathcal{C}$ of candidate parent concepts, also with text descriptions, that map to a given reference ontology $\mathcal{O}$. For $\mathcal{C}$ we here consider the large-scale lexical database WordNet [Fellbaum, 2010], which conveniently maps to many general domain ontologies like SUMO [Niles and Pease, 2001].

4.1 Hierarchy Extraction

We aim to extract a labeled binary tree T that summarizes in a human-inspectable manner how the encoder groups the given leaf classes in embedding space. Leaves correspond to the leaf dataset classes; each internal node represents an extracted parent concept from the concept bank together with an embedding. Edges represent predicted `IsParentOf` relations (cf. Figure 2a).

Example 2 (Parent Relations)**.** *In Figure 2a, image-only leaf embeddings group* `bird` *and* `deer` *under* `avifaunal` *as captured by the extracted binary tree structure.*

Approach. We build T in three steps, as shown in Algorithm 1. First, we compute a centroid embedding $e(\mathtt{C})$ for each leaf class $\mathtt{C}$ from embeddings of train samples for $\mathtt{C}$ (image, text, or both). Second, we run *agglomerative hierarchical clustering* [Ward Jr., 1963] on the leaf centroids using cosine similarity and average linkage; each merge creates a parent node whose embedding is the normalized mean of its children's embeddings. Third, we name internal nodes by matching each parent embedding p to the closest concept-bank embedding $e(\tilde{\mathtt{P}})$, using a one-to-one matching (linear sum assignment [Crouse, 2016], LSA) to optimally avoid duplicate names.

Rationale. Intuitively, each concept $\mathtt{C}$ is defined by the set of samples x to which $\mathtt{C}$ applies. Via F this translates to a *cluster of embeddings* $F(x)$; for leaf concepts we approximate this cluster using the embeddings of their training samples. By definition, a parent concept $\mathtt{P}$ should cover the cluster of each of its children concepts $\mathtt{C}$: If for all x holds $\mathtt{C}(x) \implies \mathtt{P}(x)$, then $\{x \mid \mathtt{C}(x)\} \subset \{x \mid \mathtt{P}(x)\}$. This motivates iteratively merging the most similar clusters, as achieved by the agglomerative hierarchical clustering. Choosing cosine similarity reflects the objective used in CLIP training, differing from Euclidean-linkage choices in prior work on clustering concept embeddings in neural network latent spaces [Wan *et al.*, 2020; Mikriukov *et al.*, 2025]. Intuitively, a good parent embedding for a parent cluster should minimize the distances to all of its children equally; for cosine distance on normalized vectors, the mean of children centroids fulfills this, which we assume to be the children's concept embeddings (Ass. 2):

$$e(\mathtt{P}) \approx \underset{C \in \text{Children}(P)}{\text{mean}} e(\mathtt{C}) \tag{1}$$

Averaging additionally robustifies the assignment of concept representatives and robustness increases with the number of samples in the leaf class datasets. Since clustering is applied to those robust means, it naturally becomes robust against fluctuations in the leaf datasets with more samples. Empirically we at most observed swaps between nearest and second-nearest siblings, as confirmed in qualitative UMAP visualizations (refer to supplementary for details).

Algorithm 1 Hierarchy extraction and parent naming

Require: encoder F, leaf concepts $\{\mathtt{C}_i\}_{i=1}^{k}$ with specification sets $\text{Spec}(\mathtt{C}_i)$ (image, text, or both), concept-bank descriptions Descs, cosine distance d

{*1. Obtain normalized leaf concept centroids*}
for $i = 1, \ldots, k$ **do**
  $\bar{e}_i \leftarrow \text{mean}_{x \in \text{Spec}(\mathtt{C}_i)} F(x)$
  $e(\mathtt{C}_i) \leftarrow \bar{e}_i / \|\bar{e}_i\|$
end for

{*2. Recover hierarchy using cosine average linkage*}
$T \leftarrow \text{AgglomerativeClustering}($
  $\{e(\mathtt{C}_i)\}_{i=1}^{k},\ \text{metric} = d,\ \text{linkage} = \text{average})$
for each internal node $\mathtt{P}$ of T in bottom-up order **do**
  $\bar{e}_{\mathtt{P}} \leftarrow \text{mean}_{\mathtt{C} \in \text{Children}(\mathtt{P})} e(\mathtt{C})$
  $e(\mathtt{P}) \leftarrow \bar{e}_{\mathtt{P}} / \|\bar{e}_{\mathtt{P}}\|$
end for
ParentEmbs $\leftarrow \{e(\mathtt{P}) \mid \mathtt{P} \in \text{InternalNodes}(T)\}$

{*3. Name parent embeddings by one-to-one matching*}
$e(\text{desc}) \leftarrow F(\text{desc}) / \|F(\text{desc})\|$ for each desc $\in$ Descs
assignment cost $L_{\text{assign}} : (p, \text{desc}) \mapsto d(p, e(\text{desc}))$
$A \leftarrow \text{LSA}(\text{ParentEmbs}, \text{Descs}, L_{\text{assign}})$

return tree T with node embeddings $e(\cdot)$ and parent name assignment A

4.2 Explainable and Faithful Tree-Traversal Inference with UAES

Standard zero-shot classification of a sample x directly picks the leaf class with the most similar embedding to $F(x)$. Wan et al. [2020] instead suggest to use a hierarchy tree of embeddings for inference via *tree-traversal*: One hierarchically traverses the tree from root to leaf, at each node picking the child with the closest embedding to $F(x)$. This results in a

more explainable classification, since the full path of specialization is available (instead of just `cat`, "`animal` → `mammal` → `cat`"). Interestingly, predictions don't necessarily coincide with zero-shot classification: if a leaf lies at the boundary of its superclass $\mathtt{P}$, the centroid of a sibling $\mathtt{P}'$ may be closer to $F(x)$, so traversal proceeds into the subtree of $\mathtt{P}'$ (cf. Figure 2b). From a semantic perspective this can be desirable: assuming interpolatability, a parent covers the area between its children, thereby preserving abstraction for underrepresented superclasses.

Example 3. *As illustrated in Figure 2b, in a hierarchy of many* `animal` *and few* `vehicle` *subclasses, tree-traversal can still correctly consider* `coach` *a* `vehicle` *instead of an outlier animal (*`horse`*).*

Nevertheless, a deep traversal into less matching subclasses should be avoided. Conveniently, the calculated similarities at each node may serve as a certainty score that traversal down that node is reasonable. We suggest to *stop early* at a parent node if none of its children would be sufficiently certain, returning the more reliable parent as prediction instead of a wild guess of a leaf class, as specified in Algorithm 2.

Example 4 (UAES)**.** *In Figure 2b,* `vehicle` *should be returned when traversal is uncertain between subclasses* `motor vehicle` *and* `boat-like`*). Early stopping enables more reliable classification both of unknown classes (*`coach`*) as well as known but erroneously classified ones.*

We estimate the cut-off threshold for a node at the pth quantile (here: 0.01) of similarities observed in the training set. The ideal stopping point is the first node where a "wrong turn" would be taken—the lowest common ancestor (LCA) of the ground truth and the leaf node predicted by vanilla tree-traversal. The average walking distance of a prediction to that LCA gives an intuitive measure how much reliability early stopping recovers, here abbreviated as last correct node distance.

Measuring Hierarchy Faithfulness. While vanilla tree-traversal could theoretically surpass zero-shot performance, this is not achieved even on simpler embedding spaces [Wan *et al.*, 2020]. Reasons can be: unfavorable multi-modal, scattered or overlapping concept representations in embedding space violating our assumptions [Mikriukov *et al.*, 2025], or unequal covariances ($\approx$ widths) of sibling class clusters as in unbalanced trees. Hence, we follow the XAI faithfulness notion used by Wan et al. [2020], which answers how well hierarchical-interpretation predictions coincide with vanilla zero-shot ones. They operationalize it by interpreting the ratio $\frac{\text{tree inference accuracy}}{\text{zero-shot accuracy}}$ as a proxy measure how **faithfully** the hierarchy captures the internal similarity structure. To capture how deeply rooted category confusions are, we suggest **soft tree inference accuracy** (the average fraction of the correct root-to-leaf path taken during traversal).

4.3 Verifying Plausibility of Induced Hierarchies

Since internal nodes are named and the structure is a tree, extracted hierarchies support both qualitative inspection and quantitative comparison to a reference hierarchy. We assume a target hierarchy T' represented as a directed acyclic graph, together with a mapping $\mathcal{M}$ from concept-bank entries to nodes in T'. The target T' can be a taxonomic subset of a human-defined ontology $\mathcal{O}$ or the hierarchy of a different encoder. We introduce two ontology-based verification metrics: (i) the distance to the closest valid tree extracted from the ontology (global fit), and (ii) a hierarchical consistency score $\mathcal{S}_{\text{onto}}$ that checks parent-child edges for matching ontology paths (local fit). Let d_T be the integer walking distance between nodes in the underlying undirected graph of T.

Algorithm 2 Tree-traversal with UAES

Require: encoder F, instance x, hierarchy tree $T = (\mathcal{V}, E, \mathtt{P}_{\text{root}})$, embeddings $e(\mathtt{C})$ for all $\mathtt{C} \in \mathcal{V}$, node-wise UAES similarity thresholds $(\tau_{\mathtt{P}})_{\mathtt{P}\in\mathcal{V}}$, cosine distance d
embedding $v \leftarrow F(x)$

{*Traverse while the best child remains sufficiently certain*}
current parent $\mathtt{P} \leftarrow \mathtt{P}_{\text{root}}$
instanceHierarchy $\leftarrow [\mathtt{P}_{\text{root}}]$
while $\mathtt{P}$ is no leaf **do**
  best child $\mathtt{C}^* \leftarrow \arg\min_{\mathtt{C}\in\text{Children}(\mathtt{P})} d(v, e(\mathtt{C}))$
  best child distance $d^*_{\text{child}} \leftarrow d(v, e(\mathtt{C}^*))$

  {*Uncertainty-aware Early Stopping (UAES)*}
  if $d^*_{\text{child}} > \tau_{\mathtt{P}}$ **then**
    break
  end if

  {Traverse down to best child}
  $\mathtt{P} \leftarrow \mathtt{C}^*$
  append $\mathtt{P}$ to instanceHierarchy
end while
return prediction $F_{\text{tree-traversed}}(x) = \mathtt{P}$, instanceHierarchy

Globally Verifying against the Closest Valid Trees. We consider a hierarchy plausible with respect to $\mathcal{O}$, if it is a slice thereof. To quantify how far an extracted hierarchy T is from being plausible, we search for the best matching $T' \subset \mathcal{O}$.

Definition 2. *Let d_{tree} be a tree comparison metric [Akutsu, 2010], like Tree Edit Distance [Zhang and Shasha, 1989] or Tree Alignment [Jiang* et al.*, 1995]. The* closest valid tree *to T (without renaming) with respect to ontology $\mathcal{O}$ and d_{tree} is the subtree*

$$T_{\text{valid}} = \arg\min_{T'\subset\mathcal{O}} d_{\text{tree}}(T, T') \quad \textit{s.t. } T' \textit{ is a tree} \tag{2}$$

T_{valid} can serve as a minimal target to optimize for plausibility, and $d_{\text{tree}}(T, T_{\text{valid}})$ as *global* (yet, costly to obtain) plausibility measure.

Hierarchical Consistency Score. This metric efficiently locally verifies that parent-child edges in the T respect the hypernymy constraints of the reference ontology $\mathcal{O}$. We ground each parent node u to a representative ontology concept

$$\rho(u) = \text{LCA}_{\mathcal{O}}\left(\{\mathcal{M}(l) \mid l \in \text{leaves}(u)\}\right). \tag{3}$$

For each edge $u \to v$, let $\delta(\rho(u), \rho(v))$ be the shortest directed path length in $\mathcal{O}$. We define $S_{\text{edge}}(u \to v) = 0$ if no path exists, else $S_{\text{edge}}(u \to v) = \min(1, (\gamma \cdot \delta(\rho(u), \rho(v)))^{-1})$.

The factor $\gamma = 0.5$ is adapted from path-length normalization principles [Leacock, 1998] and allows to tolerate common one-level skips. We then aggregate over edges E in T to obtain **Hierarchical Consistency**

$$\mathcal{S}_{\text{onto}}(T) = \frac{1}{|\mathcal{E}|} \sum_{(u \to v) \in \mathcal{E}} S_{\text{edge}}(u \to v). \tag{4}$$

Example 5 (Local Hierarchical Consistency)**.** *In Figure 2a, the edge* `animal` → `canine` *only skips* `mammal` *and receives high* $S_{edge} = 1$*, while* S_{edge}(`canine` → `equine`) $= 0$*.*

4.4 Aligning Induced to Target Hierarchies

The last part of our pipeline aims to post-hoc align the observed similarity structure T of the VLM encoder to a target tree T', like the closest valid tree or a different encoder's semantics. We here suggest an efficient post-hoc, model-agnostic approach that only learns a transformation for the embedding space. Thereby the loss function accounts for preserving the original similarity structure such that the transformation minimally affects zero-shot performance on the given leaf classes Ingredients are the tree structure and leaf labels of T', and a training and test split of the leaf concept dataset. Then, a transformation is obtained in two steps:

(1) Determine suitable **target latent points** to which the instances should be mapped, in order to fulfill the target similarity structure.
(2) Train a **transformation** of the embedding space, here a 2-layer DNN, on the pairs of original and targeted instance embeddings.

This modularization allows to use efficient manifold learning techniques like UMAP [McInnes *et al.*, 2020] in the first step. Via an optimization procedure UMAP finds embeddings of a set of original points in a target space, such that the embeddings match targeted pairwise point distances. This is typically used to map points into a *lower*-dimensional space, whilst targeting the pairwise distances from the original space to *preserve* neighborhoods. We instead map points into the *same* space, and want to *change* neighborhoods to match those of T'. Hence, we initialize the UMAP mapping with the identity, and sum pairwise distances between points x, x' of concepts `C`, `C'` that (α_{orig}) anchor to the original cosine geometry, (β_{onto}) drive pairwise embedding distances to align with the class distances in T' , and (γ_{midp}) determine how much the original class distances should be reduced. Combining the original cosine geometry and the class distances prevents collapsing class representations, i.e.,

$$\begin{aligned} d_{\text{emb}}(x, x') = & \overbrace{\alpha_{\text{orig}} \cdot d(F(x), F(x'))}^{\text{maintain original instances' dist.}} \\ & + \underbrace{\tfrac{1}{Z}\beta_{\text{onto}} \cdot d_{T'}(\texttt{C}, \texttt{C}')}_{\text{match ontology class dist.}} - \underbrace{\gamma_{\text{midp}} \cdot d(e(\texttt{C}), e(\texttt{C}'))}_{\text{discard orig. class dist.}} \end{aligned} \tag{5}$$

for cosine distance d, the undirected tree walking distance $d_{T'}$ in T', the normalization constant $Z := \max d_{T'} / \max_{C_1, C_2} d(e(C_1), e(C_2))$, and scalar parameters α_{orig}, β_{onto} and γ_{midp}. It is a combination of maintaining as much as possible of the *original distances* between instances within a class representation, whilst driving the embedding distances to match the length of the *shortest path between the classes* in T'. During training, the second term is normalized by the maximum path distance and maximum class embedding distance in the dataset. Given these target distances $d_{\text{emb}}(x, x')$, UMAP creates a dataset of pairs each mapping an original embedding of a probing sample to a transformed counterpart embedding vector. The pairwise distances between the transformed embeddings are encouraged to match the target distances. Since the mapping is only valid on the probing samples, the second step then trains a continuous function to extend this mapping to the complete embedding space. This can be applied before zero-shot classification on the latent space to align representations closer with a target class hierarchy.

5 Experiments

We conduct an empirical study to investigate faithfulness and utility of the extracted hierarchies for tree-traversal (Sec. 5.2); compare plausibility of hierarchies across modalities and models (Sec. 5.3), and to measure how easily embedding spaces can be post-hoc aligned to target hierarchies (Sec. 5.4).

The supplementary provides further details on the setup, ablation study, tables with numerical results of all plots, hypothesis tests (p-values are $< 10^{-5}$ *if not reported).*

5.1 Experimental Settings

A hyperparameter study selected WordNet-based over Llama3-based or combined descriptions for generating concept bank embeddings as optimal (see supplementary). WordNet is used for the general, WordNet+Llama3 for the specialized CUB dataset. The following settings apply to the total of 156 main experiments.

Models. We consider in total 13 pre-trained foundational VLMs from 4 families [Miller *et al.*, 2024]: **CLIP** (C) with 9 different backbones [Radford *et al.*, 2021], **ALIGN** [Jia *et al.*, 2021], **FLAVA** [Singh *et al.*, 2022], and **SigLIP** (v2) with 224px and 384px backbones [Zhai *et al.*, 2023].

Leaf Class Datasets. We use CIFAR-10/100 [Alex, 2009] and ImageNet [Deng *et al.*, 2009] (10/100/1000 leaves) as standard datasets of increasing hierarchy complexity, and CUB [Wah *et al.*, 2011] as a specialized 200-class bird dataset. The train split is used for hierarchy extraction and alignment training; the test split is used to assess faithfulness and plausibility.

Ontologies and Concept Bank. We use the WordNet lexical database [Fellbaum, 2010] as a concept bank and SUMO [Niles and Pease, 2001], OpenCyc [Matuszek *et al.*, 2006], and Yago [Mahdisoltani *et al.*, 2015] as reference general domain ontologies, linked via existing mappings. We include nouns and adjectives, and treat `Subclass` and `InstanceOf` as parent relations.

Metrics. We measure **discriminability** as zero-shot performance (ZS, ↑), **faithfulness** via our faithfulness ratio, (soft) tree-traversal inference (tree / tree soft, ↑) and distance to the last correct node (↓), **plausibility** using our hierarchical consistency score (↑), and normalized constrained Unordered Tree Edit Distance (nUTED, ↓) [Zhang, 1996] to obtain the

	tree (↑)	tree soft (↑)	faithfulness (↑)	ZS (↑)
CIFAR10	82 ± 11	89.0 ± 8.3	94.2 ± 6.8	86.6 ± 8.6
CIFAR100	35 ± 14	59 ± 12	55 ± 15	63 ± 13
CUB	22 ± 18	36 ± 26	37 ± 16	49 ± 33
ImageNet	18 ± 11	39 ± 18	26 ± 15	67.5 ± 8.7

Table 1: Faithfulness, tree-traversal accuracy (tree), soft tree-traversal accuracy (tree soft), and zero-shot performance (ZS) by dataset, averaged over models and encoders.

	early stop (ours)	vanilla
CIFAR10	0.35 ± 0.17	**0.33 ± 0.20**
CIFAR100	**2.12 ± 0.56**	2.43 ± 0.59
CUB	**1.11 ± 0.77**	1.66 ± 0.60
ImageNet	**2.47 ± 0.98**	4.3 ± 1.0

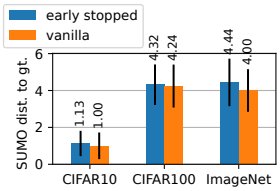


Table 2: Tree-traversal early stopping returns internal nodes instead of incorrect leaves. Left: raw last correct node distance (↓), averaged over encoders and models. Right: average semantic distance (↓) from the predicted node to the ground-truth concept in the SUMO ontology (walking distance). Best per dataset in bold.

closest valid tree. The latter is the minimal number of deletion/insertion/leaf rename operations to turn one tree into the other, normalized to $[0, 1]$ by the operation number of the trivial delete-all-insert-all solution [Rico-Juan and Micó, 2003]. Two randomly constructed binary trees for 1000 leaf nodes have a nUTED of approximately 0.54 (see supplementary).

5.2 Explaining Hierarchies: Faithfulness

A qualitative global explanation of VLM hierarchies is demonstrated in Figure 2a. To quantitatively assess explanation quality in terms of faithfulness, we use the (relative) performance of local explanations via tree-traversal inference as a proxy. We conducted an ablation study across different leaf set complexities, VLM model architectures, and the chosen encoder (text, image, mean of both).

More Leaves Diminish Faithfulness. The dominating effect is that a higher number of leaf nodes considerably decreases faithfulness as shown in Table 1: average faithfulness drops from 94.2 ± 6.8 on CIFAR-10 (10 classes) to 26 ± 15 on ImageNet (1000 classes). Similar decrease in soft tree-traversal performance (89 to 38) indicates that tree-traversal inference takes a wrong turn more easily in deeper trees. While the effect is smaller on image-embedded leaves, it is still significant. As a first pathway, this motivates use of our suggested early stopping.

Early Stopping Improves Hierarchical Inference. We evaluate whether uncertainty-aware early stopping (UAES) can correctly capture wrong turns in hierarchical inference. Table 2 reports the undirected tree distance between the predicted output (leaf or early-stopped internal node) and the last correct node, defined as the LCA of the ground-truth leaf and the leaf predicted by vanilla traversal. Early stopping reduces this distance on the larger label sets (CIFAR100/CUB/ImageNet), which means that it often returns semantically appropriate super-categories instead of committing to an incorrect leaf; on CIFAR10 the effect is negligible. Notably, measuring the distance to the ground truth node not in the tree but in the reference ontology SUMO has no significant effect. This validates that early stopping does not change the quality of the semantic alignment against a target ontology, which we inspect in detail next.

5.3 Agreement with Human Ontologies

We quantify plausibility by comparing extracted hierarchies against human reference ontologies (OpenCyc, SUMO, and Yago), using the hierarchical consistency metric introduced in Sec. 4.3. Our evaluation is driven along our desiderata: Can

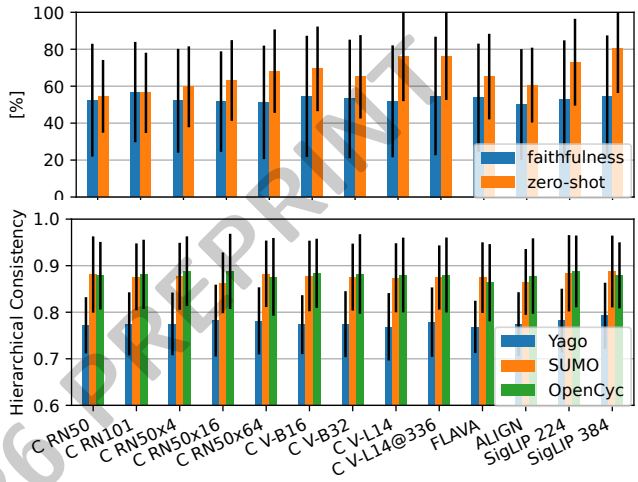
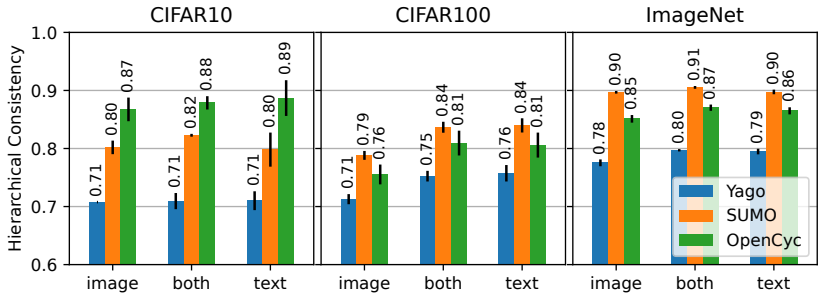


Figure 3: Faithfulness (top) versus plausibility (bottom) results for different models and backbones, averaged over 3 datasets and 3 leaf encoding types (image, text, both).

discriminability and plausibility be achieved simultaneously? And does the text-to-image alignment in CLIP training also induce semantically aligned hierarchies?

Plausibility–Discriminability Trade-off. When considering per-model metrics, zero-shot performance has little predictive value about the plausibility, as visualized in Figure 3. More interestingly zero-shot performance and faithfulness are both significantly negatively correlated with hierarchical consistency across all ontologies (for zero-shot and SUMO/OpenCyc/Yago: p-values < 0.006 and Pearson r $< -0.45/-0.21/-0.29$; for faithfulness: Pearson r$< -0.62/-0.18/-0.61$ and p-value < 0.023 for OpenCyc, else 0). Hence, **there seems to be a consistent trade-off between discriminability of leaf classes, and plausibility of the induced semantic structure.**

Modality Gap: Text versus Image Hierarchies. As shown in Figure 4, image leaf embeddings achieve significantly higher zero-shot and faithfulness scores compared to text leaf embeddings ($+22 \pm 13$ percent points), which get instead consistently higher plausibility values ($+0.015 \pm 0.032$, p-value< 0.002), see Figure 4. This is confirmed by the significantly different trees, measured via average nUTED between trees when switching from image to text leaf embeddings (see Figure 5). Combining text and image information when embed-



(a) Hierarchical consistency (↑) against OpenCyc, SUMO, and Yago across CIFAR10/100 and ImageNet.

	tree	faithful	zero-shot
image	**49 ± 24**	**63 ± 22**	**74 ± 12**
both	44 ± 27	57 ± 28	74 ± 12
text	25 ± 30	39 ± 32	52 ± 31

(b) Faithfulness metrics averaged over model, dataset, and encoder.

Figure 4: Effect of leaf embedding modality: Results of verifying and comparing encoders against a target ontology (Figures 4a and 4b). For results on comparing encoders against each other see Figure 5.

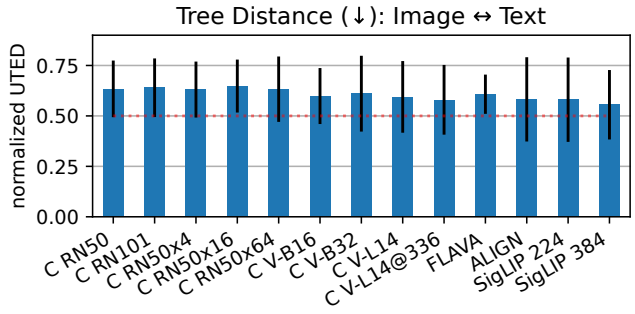


Figure 5: nUTED (↓) between trees induced by **image** and **text** encoders of each the same VLM; averaged over standard datasets. nUTED> 0.5 marks worse than random.

ding leaves offers a useful trade-off between discriminability and plausibility (cf. Figure 2a and Figure 4).

We then ask whether these trends hold consistently across model families and backbones.

Across-Model Comparison. Across VLM families, we observe substantial variation in both faithfulness and ontological plausibility. Figure 3 summarizes these trends and highlights that higher zero-shot accuracy does not necessarily imply a hierarchy that is more consistent with human ontologies.

In the next section we evaluate in how far our alignment procedure also confirms this trade-off between faithfulness and plausibility.

5.4 Post-hoc Alignment to Target Hierarchies

We here investigate whether learned ontologies can successfully be post-hoc corrected, without destroying zero-shot accuracy. For this we consider the following tasks:

- Validation: Swap two leaves in the semantic tree.
- Modality alignment: Turn a tree for image leaf embeddings into the corresponding one for text embeddings.
- Ontological commitment: Steer to the closest valid tree.

We measure proximity of a current to a target ontology by turning them into their distance matrices and taking the Frobenius norm of their difference. $\Delta\mathcal{O}$ is defined as the relative change in distance to the target from after manipulation to before. $\mathrm{ZS}^{\mathrm{Orig}}_{\mathrm{text}}$, $\mathrm{ZS}^{\mathrm{Orig}}_{\mathrm{midp}}$, $\mathrm{ZS}^{\mathrm{UMAP}}_{\mathrm{text}}$ and $\mathrm{ZS}^{\mathrm{UMAP}}_{\mathrm{midp}}$ denote zero-shot accuracy before (Orig) and after manipulation (UMAP) and base on (transformed) text embeddings and class midpoints. For this study, we investigate four models, ALIGN, CLIP with ViT-L14@336px backbone, SigLIP-224 and FLAVA (as specified before), on the validation task and run experiments on $2,500$ random training images and evaluate on $10,000$ test images of the CIFAR10 dataset. Averaged over three seeds, 10 random swaps and all four models, according to our hyperparameter grid-search the parameter setting $\alpha_{\mathrm{orig}} = 2.0$, $\beta_{\mathrm{onto}} = 1.0$, $\gamma_{\mathrm{midp}} = 2.0$, $\mathrm{N}_{\mathrm{UMAP}} = 250$, minimizes a combined evaluation score $\Delta\mathcal{O} - \mathrm{ZS}^{\mathrm{UMAP}}_{\mathrm{midp}}$, which balances ontology alignment and zero-shot performance.

Results for the best parameter setting are presented in Table 3. The general finding is that our approach can cheaply initialize transformations from one latent ontology to another, being capable to fully align the original clip embedding with an embedding where automotive and frog have been swapped. There is, however, a trade-off between zero-shot accuracy and successful ontology matching, which reveals how much alignment and class discriminability compete. A comparison against a linear transformation baseline shows how far the true transformation is from being (simply) linear (see supplementary). Results for steering of class hierarchies towards their text embeddings and closest valid trees confirm our findings and can be found in the supplementary as well as the details on the hyperparameter study.

6 Discussion and Outlook

Our experiments revealed a remaining **faithfulness gap** of our explainable surrogates to the true model behavior. Future work should close this gap further, e.g., by shifting to *distribution*-based [Mikriukov *et al.*, 2025; Lee *et al.*, 2025] and *location*-aware [Fong and Vedaldi, 2018] concept representations, and removing any *background bias* in the probing images [Schwalbe *et al.*, 2026]. Nevertheless, it should be noted that even imperfect XAI surrogates remain useful and are widely used for knowledge and test-case discovery [Ribeiro *et al.*, 2016].

Model	$\Delta\mathcal{O}$ ↓	nUTED ↓	ZS^{UMAP}_{text} ↑	ZS^{UMAP}_{midp} ↑	ZS^{Orig}_{text} ↑	ZS^{Orig}_{midp} ↑
$\alpha_{orig} = 2.0$, $\beta_{onto} = 1.0$, $\gamma_{midp} = 2.0$, $N_{UMAP} = 250$						
ALIGN	$0.589 \pm 2.9e\text{-}01$	$0.080 \pm 3.7e\text{-}03$	$0.579 \pm 1.3e\text{-}03$	0.701	0.753	0.804
C V-L14@336	$0.850 \pm 2.8e\text{-}01$	$0.098 \pm 7.1e\text{-}04$	$0.732 \pm 4.5e\text{-}03$	0.858	0.946	0.958
FLAVA	$0.731 \pm 3.9e\text{-}01$	$0.072 \pm 2.2e\text{-}03$	$0.688 \pm 1.1e\text{-}03$	0.786	0.890	0.923
SIGLIP 224	$0.720 \pm 8.6e\text{-}02$	$0.102 \pm 3.7e\text{-}04$	$0.723 \pm 2.4e\text{-}03$	0.837	0.914	0.931

Table 3: Results of the ontology manipulation for two swapped leaves; averaged over 10 random swaps and 3 seeds. We report variances exceeding $1.0e-03$ (see supplementary for all variances).

Interesting extensions of our technique would be *concept combinations* like `black cat`, e.g., using basis decomposition [Zhang *et al.*, 2021]; towards arbitrary vision *models*, e.g., using locked-image text approaches [Zhai *et al.*, 2022]; comparing against specialized ontologies; more types of *assumptions* [Speer *et al.*, 2017] and arbitrary *graph* structures; and *ante-hoc* ontology insertion during training.

Beyond applications to bias removal and model selection, we expect that our technique can *guide specialization of VLMs* using domain specific ontologies, including search for appropriate context prompts for a specific domain. Furthermore, relaxing the strong explainability assumptions on concept distributions in latent space could not only *inform concept entanglement analysis*, but also improve UAES and thus semantic uncertainty-informed VLM zero-shot inference. Further applications could invert the here proposed analysis and use a VLM to validate human ontologies: E.g., *comparing human-defined ontologies* with respect to their coverage of VLM knowledge, and *expanding or correcting human-defined ontologies* from DNN-learned knowledge.

7 Conclusion

This paper presented a principled approach to extract the learned class hierarchy induced by a given set of leaf classes from a VLM. Furthermore, we introduced techniques to measure and improve faithfulness and plausibility, where the latter makes direct use of existing human-defined knowledge bases. We generally could uncover a mismatch between class hierarchies learned by the text and by the image encoders. Our suggested method to post-hoc train ontology manipulating transformations was able to improve the matching of a learned ontology towards a desired one, with only moderate loss on zero-shot accuracy. This suggests that human-interpretable knowledge can directly be used to align VLM representations with human expectations, and potentially help in closing the gap in text-to-image alignment.

Ethical Statement

There are no ethical issues.

Acknowledgments

G.S. and S.T. acknowledge support through the project “chAI” funded by the German Federal Ministry of Research, Technology and Space (BMFTR), grant no. 16IS24058. A.M. and M.R. gratefully acknowledge financial support from the BMFTR within the project “REFRAME” (grant no. 16IS24073C). E.H., A.M. and M.R. also gratefully acknowledge support through the BMFTR junior research group “UnrEAL” (grant no. 16IS22069). J.H.L. was supported by the German Research Foundation (DFG), project no. 551629603. M.K. acknowledges support through the project ”NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung” funded by the German Federal Ministry for Economic Affairs and Energy. The authors are solely responsible for the content of this publication.

Contribution Statement

Authors Jae Hee Lee and Matthias Rottman contributed equally in advising.

References

[Akutsu, 2010] Tatsuya Akutsu. Tree edit distance problems: Algorithms and applications to bioinformatics. *IEICE Trans. Inf. Syst.*, 93-D(2):208–218, 2010.

[Alex, 2009] Krizhevsky Alex. Learning Multiple Layers of Features from Tiny Images. Master’s thesis, University of Toronto, Canada, April 2009.

[Alper and Averbuch-Elor, 2025] Morris Alper and Hadar Averbuch-Elor. Emergent Visual-Semantic Hierarchies in Image-Text Representations. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 220–238, Cham, 2025. Springer Nature Switzerland.

[Badreddine *et al.*, 2022] Samy Badreddine, Artur d’Avila Garcez, Luciano Serafini, and Michael Spranger. Logic Tensor Networks. *Artificial Intelligence*, 303:103649, February 2022.

[Baldrati *et al.*, 2022] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, 2022.

[Barraco *et al.*, 2022] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *CVPR*, 2022.

[Bender *et al.*, 2005] Michael A Bender, Martin Farach-Colton, Giridhar Pemmasani, Steven Skiena, and Pavel Sumazin. Lowest common ancestors in trees and directed acyclic graphs. *Journal of Algorithms*, 57(2):75–94, 2005.

[Bertinetto *et al.*, 2020] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks. In *CVPR*, 2020.

[Bhalla *et al.*, 2024] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE). In *NeurIPS*, 2024.

[Confalonieri and Guizzardi, 2025] Roberto Confalonieri and Giancarlo Guizzardi. On the multiple roles of ontologies in explanations for neuro-symbolic AI. *Neurosymbolic Artificial Intelligence*, 1:NAI–240754, 2025.

[Crouse, 2016] David F. Crouse. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, August 2016.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Dhall *et al.*, 2020] Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical Image Classification Using Entailment Cone Embeddings. In *CVPR Workshops*, 2020.

[Eslami and de Melo, 2024] Sedigheh Eslami and Gerard de Melo. Mitigate the Gap: Improving Cross-Modal Alignment in CLIP. In *ICLR*, 2024.

[Fellbaum, 2010] Christiane Fellbaum. WordNet. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer Netherlands, Dordrecht, 2010.

[Fong and Vedaldi, 2018] Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *CVPR*, 2018.

[Forbus and Hinrich, 2017] Kenneth D Forbus and Thomas Hinrich. Analogy and relational representations in the companion cognitive architecture. *AI Magazine*, 38(4):34–42, 2017.

[Ge *et al.*, 2023] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving Zero-Shot Generalization and Robustness of Multi-Modal Models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11093–11101, 2023.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *NeurIPS*, 2019.

[Guarino, 1998] Nicola Guarino. Formal Ontologies and Information Systems. In *FOIS*, June 1998.

[Jia *et al.*, 2021] 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 *ICML*, 2021.

[Jiang *et al.*, 1995] Tao Jiang, Lusheng Wang, and Kaizhong Zhang. Alignment of trees — an alternative to tree edit. *Theoretical Computer Science*, 143(1):137–148, July 1995.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

[Leacock, 1998] Claudia Leacock. Combining local context and wordnet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*. A Bradford Book, 1998.

[Lee *et al.*, 2025] Jae Hee Lee, Georgii Mikriukov, Gesina Schwalbe, Stefan Wermter, and Diedrich Wolter. Concept-Based Explanations in Computer Vision: Where Are We and Where Could We Go? In *ECCV Workshops*, 2025.

[Lehmann *et al.*, 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, March 2015.

[Lenat, 1989] Douglas B Lenat. *Building Large Knowledge-Based Systems*. Addison-Wesley Pub. Co., 1989.

[Liang *et al.*, 2022] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. In *NeurIPS*, 2022.

[Ma *et al.*, 2025] Pingchuan Ma, Lennart Rietdorf, Dmytro Kotovenko, Vincent Tao Hu, and Björn Ommer. Does VLM Classification Benefit from LLM Description Semantics? In *AAAI*, April 2025.

[Mahdisoltani *et al.*, 2015] Farzaneh Mahdisoltani, Joanna Asia Biega, and Fabian M. Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *CIDR*, 2015.

[Mascardi *et al.*, 2007] Viviana Mascardi, Valentina Cordì, Paolo Rosso, et al. A comparison of upper ontologies. In *Woa*, 2007.

[Matuszek *et al.*, 2006] Cynthia Matuszek, John Cabral, M. Witbrock, and John DeOliveira. An introduction to the syntax and content of cyc. In *AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, 2006.

[McInnes *et al.*, 2020] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020. arXiv:1802.03426.

[Mikolov *et al.*, 2013] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *NAACL-HLT*, 2013.

[Mikriukov *et al.*, 2023] Georgii Mikriukov, Gesina Schwalbe, Christian Hellert, and Korinna Bade. Revealing similar semantics inside cnns: An interpretable

concept-based comparison of feature spaces. In *ECML-PKDD*, 2023.

[Mikriukov *et al.*, 2025] Georgii Mikriukov, Gesina Schwalbe, and Korinna Bade. Local Concept Embeddings for Analysis of Concept Distributions in Vision DNN Feature Spaces. *International Journal of Computer Vision*, May 2025.

[Miller *et al.*, 2024] Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models. In *ECCV*, 2024.

[Niles and Pease, 2001] Ian Niles and Adam Pease. Towards a standard upper ontology. In *FOIS*, 2001.

[Novack *et al.*, 2023] Zachary Novack, Julian Mcauley, Zachary Chase Lipton, and Saurabh Garg. CHiLS: Zero-Shot Image Classification with Hierarchical Label Sets. In *ICML*, 2023.

[Otero-Cerdeira *et al.*, 2015] Lorena Otero-Cerdeira, Francisco J. Rodríguez-Martínez, and Alma Gómez-Rodríguez. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949–971, 2015.

[Ponomarev and Agafonov, 2022] Andrew Ponomarev and Anton Agafonov. Ontology Concept Extraction Algorithm for Deep Neural Networks. In *FRUCT*, 2022.

[Posada-Moreno *et al.*, 2024] Andrés Felipe Posada-Moreno, Nikita Surya, and Sebastian Trimpe. ECLAD: Extracting Concepts with Local Aggregated Descriptors. *Pattern Recognition*, 147:110146, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

[Ribeiro and Leite, 2021] Manuel de Sousa Ribeiro and João Leite. Aligning Artificial Neural Networks and Ontologies towards Explainable AI. In *AAAI*, May 2021.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*, 2016.

[Rico-Juan and Micó, 2003] Juan Ramón Rico-Juan and Luisa Micó. Comparison of aesa and laesa search algorithms using string and tree-edit-distances. *Pattern Recognition Letters*, 24(9):1417–1426, 2003.

[Schwalbe *et al.*, 2026] Gesina Schwalbe, Georgii Mikriukov, Edgar Heinert, Stavros Gerolymatos, Mert Keser, Alois Knoll, Matthias Rottmann, and Annika Mütze. On background bias of post-hoc concept embeddings in computer vision dnns. In Riccardo Guidotti, Ute Schmid, and Luca Longo, editors, *Explainable Artificial Intelligence*, pages 45–71, Cham, 2026. Springer Nature Switzerland.

[Schwalbe, 2021] Gesina Schwalbe. Verification of size invariance in DNN activations using concept embeddings. In *AIAI*, 2021.

[Singh *et al.*, 2022] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A Foundational Language and Vision Alignment Model. In *CVPR*, 2022.

[Speer *et al.*, 2017] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*, 2017.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[Wan *et al.*, 2020] Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Suzanne Petryk, Sarah Adel Bargal, and Joseph E. Gonzalez. NBDT: Neural-backed decision tree. In *ICLR*, 2020.

[Ward Jr., 1963] Joe H. Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[Yuksekgonul *et al.*, 2022] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc Concept Bottleneck Models. In *ICLR Workshop on PAIR2Struct*, 2022.

[Zhai *et al.*, 2022] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-Shot Transfer With Locked-Image Text Tuning. In *CVPR*, 2022.

[Zhai *et al.*, 2023] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[Zhang and Shasha, 1989] Kaizhong Zhang and Dennis Shasha. Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM Journal on Computing*, 18(6):1245–1262, December 1989.

[Zhang *et al.*, 1992] Kaizhong Zhang, Rick Statman, and Dennis Shasha. On the editing distance between unordered labeled trees. *Information processing letters*, 42(3):133–139, 1992.

[Zhang *et al.*, 2021] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for CNN models with non-negative concept activation vectors. In *AAAI*, 2021.

[Zhang, 1996] Kaizhong Zhang. A constrained edit distance between unordered labeled trees. *Algorithmica*, 15(3):205–222, 1996.

[Zhao *et al.*, 2025] Jitian Zhao, Chenghui Li, Frederic Sala, and Karl Rohe. Quantifying Structure in CLIP Embeddings: A Statistical Framework for Concept Interpretation, 2025. arXiv:2506.13831.