HIERARCHICAL DICTIONARY LEARNING TREE FOR INTERPRETABLE CONCEPT DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard dictionary learning finds a single set of atoms to sparsely represent all data, conflating coarse and fine structure. We present *Hierarchical Dictionary Learning Tree* (HDLT), a framework that organizes learned sparse dictionaries into a tree whose structure is grounded in Geometric Multi-Resolution Analysis. The key architectural choice, *codes as signals*, passes the sparse code at each node as the input to its children, so each level learns to represent the activation patterns of the level below. Dictionaries are trained end-to-end by backpropagating through the sparse coding optimality conditions, with an optional classification loss that aligns atoms with class boundaries. Our central result is unsupervised concept discovery: trained on CLIP image features with no labels, text, or ontology, HDLT recovers a WordNet-consistent concept hierarchy from a 30-class ImageNet subset (0.98 leaf purity, cophenetic ρ=0.92) so that a single image's route traces its WordNet path (*animal*→*carnivore*→*big cat*→*tiger*). Under an optional task loss, leaf codes approach a raw-CLIP linear-probe baseline on CIFAR-100 and ImageNet while exposing a tree-shaped representation that flat dictionary methods cannot provide.

1 INTRODUCTION

Dictionary learning and interpretability. Sparse dictionary learning seeks $\mathbf{D} \in \mathbb{R}^{D\times K}$ and sparse codes $\mathbf{Z}$ with $\mathbf{X} \approx \mathbf{DZ}$ (Elad, 2010; Aharon et al., 2006); the atoms of $\mathbf{D}$ are interpretable units of representation (e.g. Gabor-like edge detectors for images) and the codes give selective activations for downstream tasks. But a single flat dictionary must represent both global structure and local detail with one unstructured atom set: increasing K scales computation quadratically, sacrifices sparsity, and mixes coarse with fine structure.

Hierarchical structure in data and in learned representations. Natural data often lies near a low-dimensional multi-scale manifold. GMRA (Allard et al., 2012) builds a tree of nested dyadic cells with local SVD bases, giving a data-adaptive compressible basis (error $O(2^{-(1+\alpha)j})$ for $\mathcal{C}^{1+\alpha}$ manifolds). These bases are task-agnostic, orthonormal, and dense — the opposite of the sparse, task-adapted representation we seek. Independently, CLIP embeddings satisfy empirical *hierarchical orthogonality*: child-minus-parent concept differences are nearly orthogonal to the parent (Nguyen et al., 2026), and concepts decompose approximately linearly in CLIP space (Bhalla et al., 2024), suggesting a hierarchical dictionary could *discover* such structure from image features alone.

Task-discriminative learned dictionaries on the GMRA tree. We retain GMRA's tree as a scaffold and replace each SVD basis with a *learned overcomplete sparse dictionary* trained end-to-end via fixed-point differentiation through the sparse-coding step (Sun et al., 2017).[1] Our central design, *codes as signals*, passes the sparse code $\mathbf{z}_i^{(v)}$ as the child's input, encoding *which parent atoms fired* so deeper levels represent increasingly abstract co-activation patterns. An optional classification loss on leaf codes aligns atoms with class boundaries. Routing uses atom-group activation scores with a temperature-annealed softmax (Frosst & Hinton, 2017).

[1] Alternating minimization and LISTA (Gregor & LeCun, 2010) are alternatives; only the fixed-point route carries task-loss gradients *across* levels (Section 3.4).

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

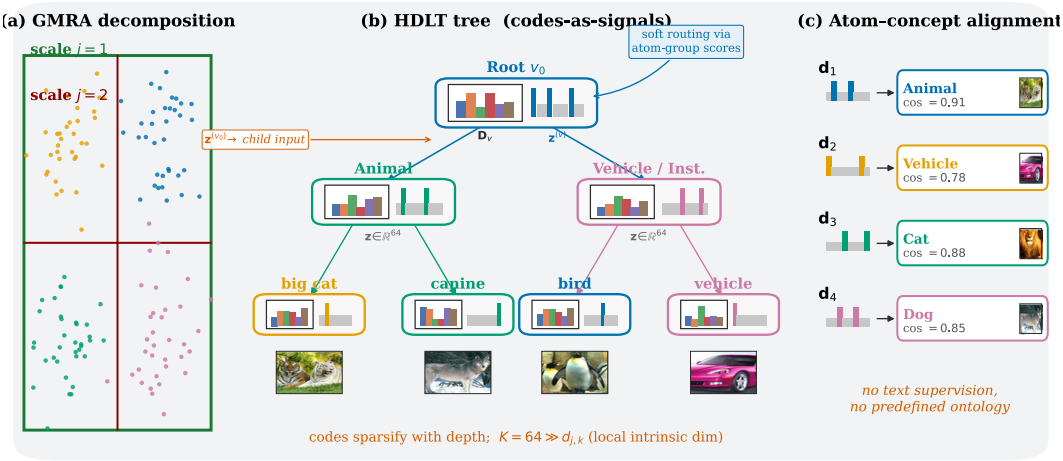

Figure 1: **Overview of HDLT.** (a) GMRA dyadic cell decomposition of data near a low-dimensional manifold: the outer rectangle is the $j{=}1$ cell, which one planar dyadic refinement splits into $2^2{=}4$ sub-cells at $j{=}2$ (solid cross), each receiving one data cluster. (b) HDLT tree: each node v holds a learned overcomplete dictionary $\mathbf{D}_v$ (heatmap glyph) and emits a sparse code $\mathbf{z}^{(v)}$ (bar glyph) that is passed as the child's input signal (*codes as signals*); codes sparsify with depth ($K{=}64 \gg d_{j,k}$); routing uses a temperature-annealed softmax over atom-group activation scores. Representative ImageNet images are shown at the four leaf concepts. (c) Leaf-level atoms align with WordNet concept-difference embeddings in CLIP space (cosine similarities shown), with no text supervision or predefined ontology.

Relation to hierarchical concept pursuit. HCEP (Nguyen et al., 2026) constructs a fixed hierarchical dictionary from CLIP embeddings over a *given* WordNet taxonomy and pursues concept paths via beam OMP. HDLT addresses the complementary problem: it learns both the hierarchy and the dictionary from data, recovering structure HCEP takes as given — applicable where no ontology exists. We defer broader positioning to Section 4.

Contributions.

1. **GMRA-grounded learned dictionaries.** We use GMRA's dyadic cell decomposition as a geometrically principled scaffold and replace each SVD basis with a learned overcomplete dictionary, deriving an exact backpropagation rule through the FISTA sparse-coding step via fixed-point differentiation (Sun et al., 2017).
2. **Codes-as-signals hierarchy.** Rather than passing residuals to children, we pass each node's sparse *code*. The code records *which parent atoms fired*; the child learns to represent these co-activation patterns, producing an increasingly abstract summary with depth — validated by a monotone decay of code support.
3. **Task-discriminative training.** An optional classification loss on leaf codes pushes atoms toward class-relevant directions, a regime reconstruction-only methods (SVD, K-SVD) cannot access.
4. **Concept discovery without supervision.** Trained on unlabelled CLIP image features, HDLT induces a tree that reproduces the WordNet branching shape (cophenetic $\rho{=}0.92$, leaf purity 0.98) on a 30-class ImageNet subset with no labels, text, or ontology provided.

2 Method

2.1 Geometric Tree Decomposition

GMRA (Allard et al., 2012) organizes data near a d-dimensional manifold into a rooted tree $\mathcal{T}$ of *dyadic cells* $\{C_{j,k}\}$, built by recursive k-means bisection. Each node (j,k) holds $\mathcal{I}_{j,k} = \{i : \mathbf{x}_i \in$

$C_{j,k}\}$ and fits a rank-$d_{j,k}$ SVD basis $\Phi_{j,k}$ — reconstruction-optimal but task-agnostic and dense. We retain this tree but replace each $\Phi_{j,k}$ with a *learned overcomplete sparse dictionary*.

2.2 Local Sparse Dictionary Learning

At each node $v = (j, k)$ of depth j, let $\mathbf{X}_v \in \mathbb{R}^{m_j \times |\mathcal{I}_v|}$ denote the matrix of input signals for the data points in $\mathcal{I}_v$, where m_j is the signal dimension at depth j. We associate to each node a *learned dictionary* $\mathbf{D}_v \in \mathbb{R}^{m_j \times K}$ whose K columns $\mathbf{d}_{v,1}, \ldots, \mathbf{d}_{v,K}$ are the atoms, and seek sparse codes $\mathbf{Z}_v \in \mathbb{R}^{K \times |\mathcal{I}_v|}$. The dictionary is *overcomplete relative to the local intrinsic dimension*: K exceeds the local intrinsic dimension $d_{j,k}$ of the cell (the number of SVD directions GMRA would use; see below), though not necessarily the input dimension m_j.

$$\min_{\mathbf{D}_v,\, \mathbf{Z}_v} \|\mathbf{X}_v - \mathbf{D}_v \mathbf{Z}_v\|_F^2 + \lambda \|\mathbf{Z}_v\|_1 \quad \text{s.t.} \quad \|\mathbf{d}_{v,k}\|_2 \leq 1 \; \forall k \in [K]. \tag{1}$$

Relationship to GMRA. The GMRA basis solves equation 1 with $K{=}d_{j,k}$, $\lambda{=}0$, and orthonormality — reconstruction-optimal but task-agnostic. Our formulation relaxes all three: $K > d_{j,k}$ (overcomplete), $\lambda > 0$ (sparse codes), column-norm replaces orthonormality.

The sparse coding step of equation 1 (with $\mathbf{D}_v$ fixed) is solved using ISTA (Daubechies et al., 2004) or FISTA (Beck & Teboulle, 2009); the dictionary update (with $\mathbf{Z}_v$ fixed) uses projected gradient descent onto the column-norm constraint.

2.3 Codes as Signals: A Learned Wavelet Hierarchy

In GMRA, after computing the scale-j approximation $\mathbf{x}_j = \mathbb{P}_{j,x}(\mathbf{x})$, the *geometric wavelet* coefficients $q_{j+1,x} = \Psi_{j+1,x}^\top (\mathbf{x}_{j+1} - c_{j+1,x})$ encode the residual between successive scale approximations, where $\Psi_{j+1,x}$ is an orthonormal basis for the wavelet subspace $W_{j+1,x} = (I - P_{j,x})V_{j+1,x}$. These coefficients satisfy $\mathbf{x} \approx \mathbb{P}_{j_0,x}(\mathbf{x}) + \sum_{\ell \geq j_0} Q_{\mathcal{M}_{\ell+1}}(\mathbf{x})$ and decay quadratically with scale (Allard et al., 2012, Thm. 2.3).

We build a multi-scale hierarchy in *sparse code space*. While GMRA passes a residual to the next scale, we pass the parent's *sparse code*: the input to a child node v' at depth $j+1$ is the sparse code produced at the parent v at depth j:

$$\mathbf{x}_i^{(j+1)} = \mathbf{z}_i^{(v)}, \quad v = \text{parent of } v'. \tag{2}$$

Interpretation. The sparse code $\mathbf{z}_i^{(v)}$ encodes *which atoms* of the parent fired; the child learns to represent these co-activation patterns. Setting $K \ll D$ at the root and decreasing K with depth, each level performs a dimensionality reduction analogous to GMRA's ambient-to-local compression. Monotone decay of code support with depth certifies that each level *abstracts* the one below; a flat profile would signal redundant re-encoding. We use this as a diagnostic in Section 3.3.

Remark 1 (Scope). *HDLT's contribution is empirical: we show a WordNet-consistent hierarchy is* discoverable *(Section 3.2) but do not claim exact-recovery guarantees. The enabling conditions — well-separated concept cells and hierarchical orthogonality of CLIP embeddings (Park et al., 2025; Nguyen et al., 2026) (verified for the discovered tree in Appendix L) — are inherited geometry; turning them into finite-sample identifiability guarantees is left to future work.*

2.4 Routing via Sparse Codes

After solving equation 1 at node v with C children, point i is routed by its sparse code $\mathbf{z}_i^{(v)}$. Given a partition $\{\mathcal{A}_c\}_{c=1}^C$ of the K parent atoms into C groups, the score for child c is the total activation mass on its atoms, $s_c(\mathbf{z}_i^{(v)}) = \sum_{k \in \mathcal{A}_c} |z_{ik}^{(v)}|$, and the soft assignment is the temperature-τ softmax $p(c \mid i) = \exp(s_c/\tau) / \sum_{c'} \exp(s_{c'}/\tau)$, which reduces to hard routing $c^*(i) = \arg\max_c s_c$ as $\tau \to 0$. Annealing τ during training thus interpolates between a fully differentiable surrogate and an interpretable partition; at convergence, $\{\mathcal{A}_c\}$ defines a learned partition of the parent dictionary that aligns with the geometric partition of $C_{j,k}$ into child cells $\{C_{j+1,k'}\}$.

2.5 Atom-Group Update

The partition $\{\mathcal{A}_c\}_{c=1}^{C}$ governs routing and can in principle be updated during training — a greedy k-means-like re-assignment or a differentiable soft variant with a learnable $\mathbf{\Pi} \in \mathbb{R}^{K \times C}$, both derived in Appendix A. All experiments below use a static round-robin partition $\mathcal{A}_c = \{k : k \bmod C = c\}$: the codes-as-signals coupling already drives atoms toward depth-aligned roles under a fixed partition (Section 3.3), and a dynamic update is left to follow-up work.

2.6 End-to-End Objective

Under soft routing, the contribution of point i to node v is the product, over the ancestors u on the root-to-v path $\pi(v)$, of the probability of routing i into $\mathrm{ch}_u(v)$, the unique child of u that also lies on $\pi(v)$:

$$w_i^{(v)} = \prod_{u \in \pi(v)} p(\mathrm{ch}_u(v) \mid i, u)\,. \tag{3}$$

The full tree objective is:

$$\mathcal{L} = \sum_{v \in \mathcal{V}} \sum_{i=1}^{N} w_i^{(v)} \left(\left\| \mathbf{x}_i^{(\ell_v)} - \mathbf{D}_v \mathbf{z}_i^{(v)} \right\|_2^2 + \lambda_v \left\| \mathbf{z}_i^{(v)} \right\|_1 \right), \tag{4}$$

where ℓ_v is the depth of node v and $\lambda_v > 0$ is a per-node sparsity parameter. Because the routing weights $w_i^{(v)}$ depend on sparse codes at ancestor nodes, which in turn depend on those ancestors' dictionaries, gradients flow through the entire tree jointly.

Task loss. When class labels $\{y_i\}$ are available, we augment equation 4 with a classification term on the leaf-level sparse codes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{tree}} + \mu\, \mathcal{L}_{\text{task}} \left(\{\mathbf{z}_i^{(v)}\}_{v \in \mathcal{V}_{\text{leaf}}},\, \{y_i\} \right), \tag{5}$$

where $\mathcal{L}_{\text{task}}$ is cross-entropy applied to a linear classifier over leaf codes and $\mu > 0$ balances reconstruction and discrimination. Setting $\mu = 0$ recovers a purely geometric decomposition analogous to GMRA. Adding $\mathcal{L}_{\text{task}}$ pushes the learned atoms away from principal components toward class-relevant directions, producing codes that are sparse, locally reconstructive, *and* predictive. The SVD basis cannot achieve this trade: it is reconstruction-optimal by construction, with no mechanism to align atoms with class boundaries.

2.7 Backpropagation Through Sparse Coding

To compute $\partial \mathcal{L} / \partial \mathbf{D}_v$ we differentiate through the sparse-coding optimality condition (Sun et al., 2017; Gregor & LeCun, 2010; Amos & Kolter, 2017): with active set Λ, the optimal codes satisfy $\boldsymbol{\alpha}_\Lambda^* = (\mathbf{D}_\Lambda^\top \mathbf{D}_\Lambda + \lambda_2 \mathbf{I})^{-1} (\mathbf{D}_\Lambda^\top \mathbf{x} - \lambda_1 \mathbf{1})$. Implicit differentiation and the chain rule yield the dictionary gradient

$$\frac{\partial \mathcal{L}}{\partial \mathbf{D}_v} = -\mathbf{D}_v \boldsymbol{\gamma}_\Lambda \boldsymbol{\alpha}_\Lambda^\top + (\mathbf{x}_v - \mathbf{D}_v \boldsymbol{\alpha}) \boldsymbol{\gamma}^\top, \quad \boldsymbol{\gamma}_\Lambda = (\mathbf{D}_\Lambda^\top \mathbf{D}_\Lambda + \lambda_2 \mathbf{I})^{-1} (\partial \mathcal{L} / \partial \boldsymbol{\alpha})_\Lambda, \;\; \boldsymbol{\gamma}_{\Lambda^c} = \mathbf{0}. \tag{6}$$

Because only active atoms contribute, the linear solve scales with $|\Lambda|$ and is efficient. The sparsity parameters λ_v are themselves treated as learnable.

2.8 Algorithm and Training Curriculum

Each epoch we (i) compute $\mathbf{Z}_v$ by FISTA (Beck & Teboulle, 2009), (ii) update $\mathbf{D}_v$ via equation 6, and (iii) anneal $\tau \leftarrow \tau / \rho$. Since unreliable early codes corrupt deeper levels, we use a three-phase λ_v curriculum: **Phase 1** (λ_v=0, dense PCA-like warm-up); **Phase 2** (anneal λ_v: $0 \to \lambda_v^*$); **Phase 3** (fix $\lambda_v = \lambda_v^*$, full FISTA). Dictionaries initialise from random Gaussian columns normalised to unit norm.

3 EXPERIMENTS

3.1 SETUP

Data and features. All experiments use CLIP ViT-B/32 image embeddings (Radford et al., 2021) (D=512, L2-normalised). For CIFAR-100 (Krizhevsky, 2009) we use the official 50 000 / 10 000 split (100 classes); for ImageNet-1k (Deng et al., 2009) we sample 100 images per class for a 100 000-feature subset (1000 classes), then take an 80 / 20 deterministic split for evaluation.

Tree and dictionary. The GMRA tree is built by recursive k-means bisection (the implementation of (Allard et al., 2012)). Our default configuration uses binary bisection (b=2, J=4, 8 leaves); a wider b=4 variant (4-way k-means at every node, 64 leaves) and a per-node adaptive-arity variant, closer to the dyadic refinement prescription of GMRA at higher input dimensionality, are reported in Appendix D. Each node carries a dictionary of K=64 atoms with the column-norm constraint $\|\mathbf{d}_{v,k}\|_2 \leq 1$, overcomplete in the sense of Section 2.2 (K=64 $\gg d_{j,k}$, the cell's local intrinsic dimension). Root input dimension is D=512; subsequent levels take their parent's K=64-dimensional sparse code as input (codes-as-signals), so non-root dictionaries are square in the input dimension while remaining overcomplete relative to the intrinsic dimension.

Training curriculum. We follow the three-phase schedule of Section 2.8, where Phase 1 spans the first E_1 epochs and Phase 2 runs from E_1 to E_2 (Phase 3 is the remainder). We set E_1=10% and E_2=50% of the total epoch budget (120 epochs for CIFAR-100, 80 for ImageNet). λ_v anneals linearly from 0 to λ_v^* in Phase 2 and stays at λ_v^* in Phase 3. Routing temperature starts at τ_0=1 and is annealed by ρ=1.05 per epoch. Sparse coding uses FISTA with $T_{\max}$=20 inner iterations. Optimisation is Adam with learning rate 10^{-2}. Configurations are referenced as (λ^*, μ) pairs.

Baselines. **Raw CLIP linear probe**: logistic regression on CLIP features directly — the upper bound any sparse-code method must approach. **K-SVD** (Aharon et al., 2006): a flat overcomplete dictionary (K=512) trained on the same features by alternating FISTA and projected-gradient updates with λ-continuation, then a linear probe on the codes. **Hierarchical NN**: route each test point to a leaf by nearest cluster centre at every depth and predict the leaf's majority training label. **HCEP HB-OMP → probe**: codes from Hierarchical Beam OMP on a fixed concept-difference dictionary of CLIP *text* embeddings of each class's WordNet hypernym chain (Nguyen et al., 2026), fed to logistic regression (support precision/recall in Appendix B). **Greedy HDLT**: our model with the parent code detached before children, blocking cross-level gradients (Section 3.4).

Metrics. *Classification accuracy*: top-1 accuracy of a logistic regression probe trained on the concatenated leaf-level sparse codes (or, for HDLT with $\mu > 0$, the trained classifier head). *Mean code support*: per-tree-depth ℓ_0 count of active atoms in a leaf-level code, averaged over the training data. *Strict monotone* ℓ_0: the central codes-as-signals prediction (Section 2.3); checked across the four tree depths. *Support precision / recall*: for HCEP, fraction of recovered atoms that lie on the planted WordNet root-to-leaf path, and conversely.

3.2 UNSUPERVISED CONCEPT DISCOVERY

This section reports all three criteria (leaf purity, cophenetic ρ, and atom–concept cosine) on a 30-class ImageNet subset, and Fig. 2 exhibits the claim on three images spanning distinct branches of the tree. This is an empirical (discoverability) contribution; we defer formal identifiability to the Limitations (Section 5).

Trained purely reconstructively ($\mu = 0$) on CLIP features of a curated 30-class ImageNet subset (six WordNet supercategories, five classes each), HDLT should (i) segregate supercategories into distinct subtrees, (ii) reproduce the WordNet branching shape, and (iii) learn atoms aligned with WordNet concept-difference directions. The 84-synset hypernym closure is used *only to evaluate* the result; it never enters training, and concept labels are assigned only *after* training.

Reading one image's path. Figure 2 traces a tiger, a king penguin, and a sports car end to end: each image's HDLT route sharpens from root to leaf and matches its WordNet chain rung by rung,

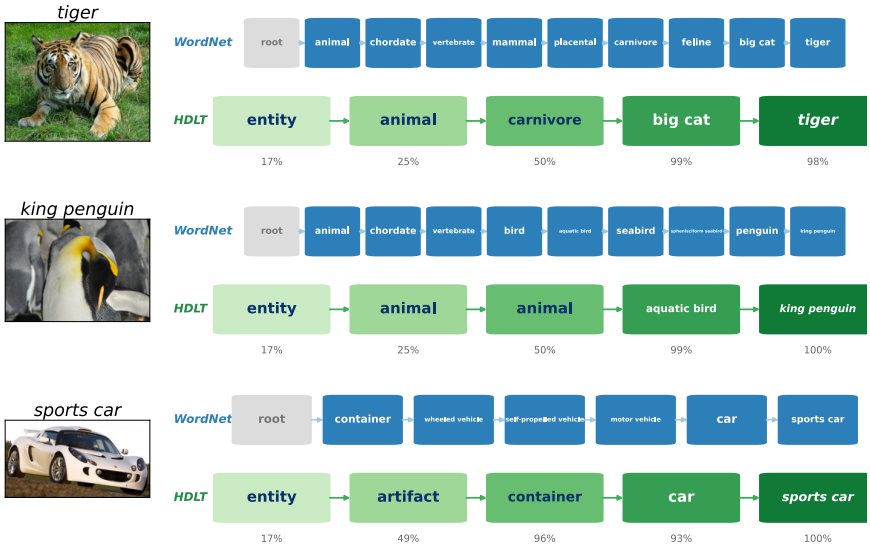

Figure 2: **Three images, end to end.** Each row shows an input ImageNet image (left) and two parallel paths for that *same* image: its ground-truth WordNet hypernym path (top, blue; the generic top-of-hierarchy rungs are collapsed into *root*) and the sequence of cells it routes into within the unsupervised HDLT tree (bottom, green; μ=0, no labels or text), each cell labelled by the WordNet synset nearest its centroid in CLIP space and annotated with its purity. Reading left to right the induced HDLT path recovers the WordNet chain coarse-to-fine for a big cat, a bird, and a wheeled vehicle drawn from different subtrees, though the model was never shown a label or the word "tiger".

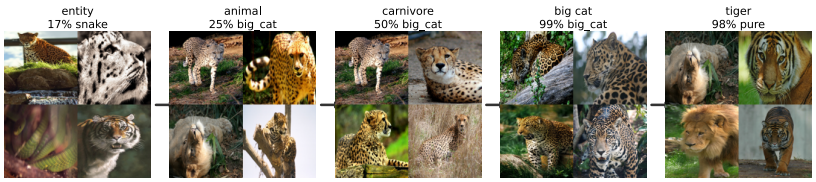

Figure 3: **Coarse-to-fine concept discovery.** Central images (2×2 montage) of each cell along the unsupervised HDLT path to the *tiger* leaf, with the decoded concept (nearest WordNet anchor) and supercategory purity. The cell refines from a maximally mixed root to a 98%-pure tiger leaf — no labels or text were shown to the model.

across the animal/artifact split. Figure 3 shows the actual tiger-path cell images, purity climbing 17%→98%; Appendix Fig. 8 extends this to all six supercategories.

Quantitative recovery. Over the whole 30-class subset (Table 1, Fig. 4) the induced 16-leaf tree attains leaf purity 0.98 and cophenetic $\rho = 0.92$. Composing atoms back to CLIP space and matching to concept-difference embeddings gives mean cosine 0.32 vs. 0.10 random (62% above 0.3); top-k images per leaf form coherent concept groups (Appendix I). Codes-as-signals is what aligns the atoms: residual-passing scores markedly worse (0.22 mean cosine, 14% above 0.3) because the residual sheds concept signal with depth (Appendix B.1).

Atom–concept alignment and its attribution. Clustering methods recover a data *partition* but expose no atoms; on the CIFAR-100 20-superclass taxonomy they and HDLT recover superclasses comparably (purity 0.39–0.52; Table 9, Appendix K), so coarse partition recovery does not separate the methods. Learned-dictionary methods (flat SAE, Matryoshka, MP-SAE, cascaded SAE) do expose atoms; we evaluate atom–concept cosine on identical CLIP features (Table 2).

The metric is sensitive to a geometric confound: concept-difference directions are *between-class-mean* directions, so any cluster-preserving representation aligns by construction. The GMRA scaffold

Table 1: Unsupervised concept discovery on a 30-class ImageNet subset (6 WordNet supercategories). All methods use the same CLIP ViT-B/32 features; HDLT is trained with μ=0 (no labels, no WordNet). K-means and Ward are run with 16 clusters on the same features. †Cophenetic ρ: Spearman correlation of pairwise leaf distances vs. WordNet hypernym distances; K-means (Lloyd, 1982) and Ward-16 (Ward Jr, 1963) use a 0/1 same-leaf indicator (no multi-level tree structure). Rows 5–7 are HDLT-only (baselines have no learned atom dictionaries).

Metric	HDLT	K-means (16)	Ward (16)
Leaf purity (6 super.)	0.98	0.99	0.98
ARI / NMI (6 super.)	0.56 / 0.76	0.53 / 0.77	0.59 / 0.76
ARI / NMI (30 classes)	0.45 / 0.74	0.47 / 0.76	0.43 / 0.75
Cophenetic ρ (vs. WordNet)†	**0.92**	0.33	0.34
Atom→concept cosine (rand. 0.10)	0.32	n/a	n/a
Atoms with cos $>$ 0.3	62%	n/a	n/a
Support precision / recall	0.67 / 0.12	n/a	n/a

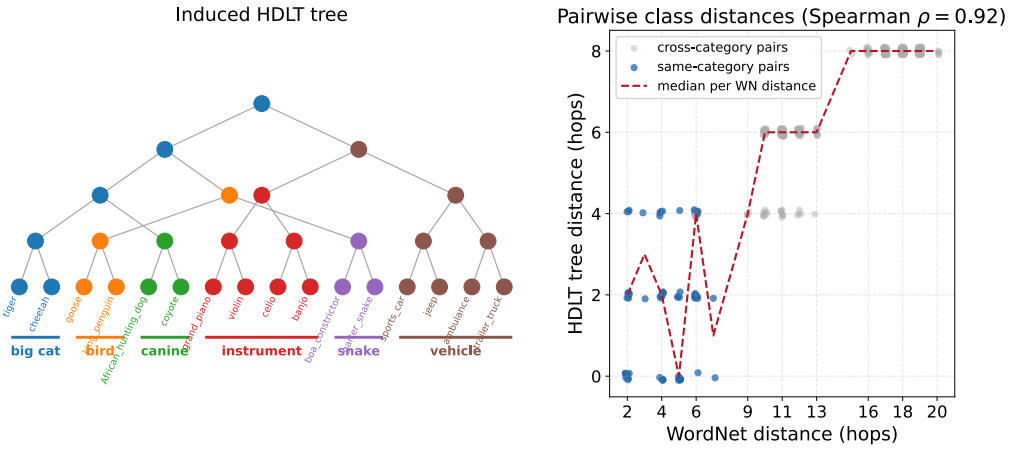

Figure 4: **HDLT recovers the WordNet hierarchy without labels or text.** *Left:* the tree learned from unlabelled CLIP image features. Nodes are coloured by their dominant WordNet supercategory; leaves are labelled by their dominant class and sorted so each supercategory forms a contiguous coloured block. Six distinct coloured subtrees emerge (leaf purity 0.98). *Right:* Spearman correlation between HDLT pairwise tree distances and WordNet pairwise distances over all $\binom{30}{2} = 435$ class pairs ($\rho = 0.92$); same-category pairs (blue) cluster at small distances in both trees.

scores 0.81 with raw centroid diffs (no learned dictionary); class-mean diffs score 0.63 — both above the 0.58 self-similarity ceiling. Reconstruction SAEs remain near the random floor (0.21–0.22). HDLT attains 0.32 by inheriting the scaffold's geometry; the root dictionary alone scores 0.30 $\approx$ full composed 0.32, so end-to-end learned composition adds negligibly to alignment. We therefore frame this as an inductive-bias distinction — cluster-grounded representations retain between-class-mean geometry that reconstruction objectives suppress — not a claim that learned atoms outperform SAEs. The codes-as-signals design yields sparse, hierarchically structured, predictive codes quantified next.

3.3 Produced Codes Are Also Predictive

Concept discovery is our focus, but the leaf codes are also usable for classification — which lets us compare fairly against sparse-coding baselines at matched budgets. Table 3 reports top-1 accuracy and leaf-level mean support $\bar{\ell}_0^{(J)}$ for raw CLIP, K-SVD, and two HDLT settings (identical hyperparameters across datasets bar the epoch budget). For each dataset HCEP HB-OMP is given a fixed WordNet concept-difference dictionary built from CLIP text embeddings of that dataset's class taxonomy (1,812 synsets for ImageNet; Appendix B); designed for concept recovery rather than classification, it trails the learned sparse coders on both datasets.

Table 2: **Atom–concept alignment** on the 30-class ImageNet subset: mean cosine of each atom to its nearest WordNet concept-difference direction (higher = more concept-aligned) and the fraction above 0.3. The metric favours *between-class-mean* geometry: cluster-centroid references (top block) score high without any learned dictionary, reconstruction SAEs (bottom block) remain near the random floor, and HDLT inherits the scaffold's geometry. Its root dictionary alone (0.30) matches the full composed atoms (0.32), localising the alignment to the scaffold rather than the learned composition.

Method	Mean atom–concept cosine	Atoms > 0.3
Cluster-centroid geometry (no learned dictionary)		
GMRA k-means scaffold (centroid diffs)	0.81	100%
Class-mean differences	0.63	100%
Concept-diff self-similarity (ceiling)	0.58	58%
Learned tree of dictionaries (ours)		
HDLT composed atoms	**0.32**	**62%**
HDLT root dictionary only	0.30	50%
Reconstruction SAEs		
Flat SAE (Cunningham et al., 2024)	0.22	21%
Cascaded SAE (Zhao et al., 2026)	0.22	20%
MP-SAE (Costa et al., 2025)	0.22	13%
Matryoshka SAE (Bussmann et al., 2025)	0.21	19%
Random directions	0.10	0%

Table 3: **Main results.** Top-1 accuracy (%), leaf-level mean support $\bar{\ell}_0^{(J)}$, and learned dictionary parameters on CIFAR-100 and ImageNet (100 images/class). HDLT (μ=0.5): mean±std over 5 seeds; others: single run. †Fixed concept-difference directions from WordNet+CLIP; not learned.

		CIFAR-100		ImageNet (100/class)	
Method	Dict. params	Acc. (%)	$\bar{\ell}_0^{(J)}$	Acc. (%)	$\bar{\ell}_0^{(J)}$
Raw CLIP (no sparse coding)	n/a	74.27	n/a	64.90	n/a
Hierarchical NN (tree majority vote)	n/a	6.36	n/a	0.56	n/a
HCEP HB-OMP (Nguyen et al., 2026)	0†	18.20	∼7.3	2.35	∼7.5
K-SVD (Aharon et al., 2006) (K=512)	262K	48.62	108.5	45.51	166.1
HDLT ($\mu = 0$, no task loss)	90K	49.72	31.17	0.29	19.26
HDLT ($\mu = 0.5$, ours)	90K	$\mathbf{73.90 \pm 1.20}$	50.55	$\mathbf{56.55 \pm 2.09}$	50.83

HDLT (μ=0.5) reaches 73.90±1.20% on CIFAR-100 (0.37 pp below raw-CLIP) and 56.55±2.09% on ImageNet, beating flat K-SVD by ∼25/∼11 pp with 3× fewer parameters (90K vs. 262K; Appendix C). Swept against the full HCEP baseline set, HDLT is the most accurate at every sparsity budget without the ontology HB-OMP and CBM consume (Appendix J). Codes-as-signals is the only inter-level mode combining monotone support decay with competitive accuracy; raw-data-passing re-encodes the same input and residual-passing sheds class signal (Appendix F).

3.4 Ablation: End-to-End vs Greedy, and the (λ^*, μ) Trade-off

A (λ^*, μ) sweep (Appendix G) shows larger λ^* buys sparsity at the cost of accuracy, μ the reverse; $(0.025, 0.5)$ gives the best trade-off. A Phase-1 warm-up adds $+4.0$ pp (Appendix G).

End-to-end vs greedy training. To isolate the cross-level fixed-point gradient (Eq. 6) we compare against the greedy variant of Section 3.1 (parent code detached; full table in Appendix E). *Under the task loss* ($\mu > 0$) end-to-end wins decisively ($+17.4$ pp probe on CIFAR-100, $+22.3$ pp on ImageNet): the head signal must flow back across levels to reshape parent atoms, which greedy freezes. *Without it* ($\mu = 0$) greedy wins and stays monotone, as the purely reconstructive objective over-sparsifies the deepest ImageNet leaves until they lose class signal.

Does the hierarchy matter? Collapsing HDLT to a flat $K{=}512$ dictionary (matching the $8{\times}64$ leaf-code dimension) under identical differentiable-FISTA machinery gives 41.9% probe accuracy (reconstruction-only) and 34.9% (with task loss) on a matched CIFAR-100 split, vs. 67.5% for the tree — isolating the *hierarchy*, not sparse coding or dictionary size, as the source of HDLT's accuracy ($\sim$25–33 pp gap).

4 Related Work

Geometric multi-resolution and hierarchical dictionaries. GMRA (Allard et al., 2012) builds a multiscale tree of local linear bases with provable approximation rates; its dictionary-learning extensions add non-asymptotic guarantees (Maggioni et al., 2016) and quadtree-based multiscale dictionaries (Mairal et al., 2008), and hierarchical/tree-structured sparse coding imposes tree penalties on a *single* dictionary (Jenatton et al., 2011). All are reconstruction-driven and task-agnostic; none couples levels by passing codes as signals nor trains end-to-end through the sparse-coding step with a task loss.

Sparse autoencoders and mechanistic interpretability. SAEs recover interpretable features from network activations (Cunningham et al., 2024; Bricken et al., 2023); recent hierarchical variants include nested dictionaries (Bussmann et al., 2025), sequential matching pursuit (Costa et al., 2025), cascaded SAEs (Zhao et al., 2026), and explicit feature trees (Cao et al., 2026). These learn a feature set in a single activation space; HDLT constructively couples a *tree of dictionaries* via codes-as-signals and scores the recovered hierarchy against a gold ontology. Its cluster-grounded scaffold retains the between-class-mean geometry that reconstruction SAEs discard (Section 3.2).

Concept geometry and interpretable-by-design models. The geometry of categorical and hierarchical concepts in LLM representations has been formalized as simplices/polytopes with hierarchical orthogonality (Park et al., 2025), a property CLIP embeddings empirically satisfy (Nguyen et al., 2026). Interpretable-by-design vision models instead attach human concepts through supervision (concept bottlenecks (Koh et al., 2020)) or learned prototypes (ProtoPNet (Chen et al., 2019)). HDLT is complementary: it *discovers* the concept hierarchy from features alone, without a predefined concept set, ontology, or per-concept labels.

5 Conclusion

We presented HDLT, a tree-structured dictionary learning framework grounded in GMRA geometry and coupled by a codes-as-signals design trained end-to-end through the sparse-coding step. Without supervision it recovers a WordNet-consistent concept hierarchy from CLIP image features (leaf purity 0.98, cophenetic $\rho{=}0.92$), and under an optional task loss reaches within 0.4 pp of a raw-CLIP probe on CIFAR-100 while exposing per-image tree-structured codes. Ablations isolate codes-as-signals as the only inter-level mode uniting monotone sparsity decay with competitive accuracy. We hope our work will contribute to future research on unsupervised structure discovery in feature space, and on interpretable-by-design models that are both accurate and easy to interpret.

When does HDLT recover the hierarchy? Our contribution is empirical: we show a WordNet-consistent hierarchy is *discoverable*, without claiming exact-recovery guarantees. Two geometric conditions suffice: (i) the concept cells are well separated, so the GMRA k-means partition aligns with the concept partition; and (ii) concept embeddings satisfy *hierarchical orthogonality* — child-minus-parent differences orthogonal to the parent, which CLIP embeddings empirically satisfy (Park et al., 2025; Nguyen et al., 2026) and we verify for the *discovered* tree (Appendix L). Under both, the composed leaf atoms point along concept-difference directions and the induced tree matches the concept hierarchy.

Limitations. (i) *Robustness*: under feature-space Gaussian noise ($\sigma{=}0.075$) HDLT collapses to chance while a raw-CLIP probe retains 43%, as soft-thresholding reshuffles active atoms across depths (Appendix H). (ii) The 8-pp ImageNet gap persists with a wider $b{=}4$ tree (Appendix D), so the bottleneck is sparse-coding compression, not width. (iii) SVD warm-start, finite-sample identifiability, and adaptive tree growing remain future work.

AI DISCLOSURE

In this work, we used generative AI tools for designing and providing feedback on our experimental methodology, implementing methods, generating synthetic datasets (subset of real dataset) and interpreting results. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for or writing proofs, or proposing or refining hypotheses, and translation and support for qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for drafting and editing parts of the paper to improve readability, creating and editing software code, creating and modifying figures, summarising and searching related literature, brainstorming, and formatting references. We have reviewed all AI-assisted work. All AI-assisted code was verified and tested for correctness by the authors, every reported number was produced by scripts the authors ran and verified, and all AI-assisted text, claims, and figures were checked and edited by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. volume 54, pp. 4311 – 4322, 12 2006. doi: 10.1109/TSP.2006.881199.

William K. Allard, Guangliang Chen, and Mauro Maggioni. Multiscale geometric methods for data sets II: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32 (3):435–462, 2012.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. pp. 693–696, 2009. doi: 10.1109/ICASSP.2009.4959678.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 6077–6101, 2025.

Tue M. Cao, Hoang X. Nhat, Raed Alharbi, Phi Le Nguyen, and My T. Thai. Tree SAE: Learning hierarchical feature structures in sparse autoencoders. *arXiv preprint arXiv:2605.07922*, 2026.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

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, 2009. doi: 10.1109/CVPR.2009.5206848.

Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 399–406, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. 2019.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982.

Mauro Maggioni, Stanislav Minsker, and Nate Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *Journal of Machine Learning Research*, 17(2):1–51, 2016.

Julien Mairal, Guillermo Sapiro, and Michael Elad. Learning multiscale sparse representations for image and video restoration. *Multiscale Modeling & Simulation*, 7(1):214–241, 2008. doi: 10.1137/070697653. URL https://doi.org/10.1137/070697653.

Nghia Nguyen, Tianjiao Ding, and René Vidal. Hierarchical concept embedding & pursuit for interpretable image classification. *arXiv preprint arXiv:2602.11448*, 2026.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pp. 8748–8763, 2021.

Xiaoxia Sun, Nasser M. Nasrabadi, and Trac D. Tran. Supervised deep sparse coding networks. *arXiv preprint arXiv:1701.08349*, 2017.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

Yusong Zhao, Hengyi Wang, Tanuja Ganu, Akshay Nambi, and Hao Wang. Cascaded sparse autoencoders learn multi-level visual concepts in multimodal LLMs. *arXiv preprint arXiv:2606.16193*, 2026.