ON THE GEOMETRY OF HIERARCHICAL CONCEPT SPACES IN VISION MODELS: TESTING LINEAR AND HYPERBOLIC REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic concepts provide a human-interpretable basis for understanding the representations learned by vision models. Under the Linear Representation Hypothesis (LRH), concepts are modeled as linear directions in a flat embedding space, where they can be composed and translated relative to one another. Meanwhile, prior work has shown that semantic concepts are often organized hierarchically, which raises a fundamental question: is such hierarchical organization geometrically compatible with LRH? We introduce *residual reconstruction*, a geometry-agnostic framework that formalizes geometric compatibility as a quantitative measure, allowing us to assess how well a given geometry captures the structure of hierarchical concepts. We apply it to the flat linear geometry underlying LRH, evaluating the compatibility across four distinct vision models on ImageNet and its associated WordNet hierarchy. Our results show that hierarchical concepts exhibit substantially larger residual reconstruction error, indicating a systematic mismatch between hierarchical organization and the flat additive geometry underlying the LRH formulation. We then investigate whether hyperbolic geometry better captures the structure of hierarchical concepts. Across the same models and hierarchies, hyperbolic representations yield substantially lower residual reconstruction error, providing empirical evidence that the hierarchical concept structures are better captured by hyperbolic geometry than by the corresponding flat geometry. Our work provides a general quantitative framework for studying the geometric compatibility of semantic structures and reveals a systematic tension between hierarchical concepts and linear representations in vision models.

1 INTRODUCTION

The linear representation hypothesis (LRH) (Park et al., 2024) is a central organizing principle in concept-based model interpretability. It posits that human-interpretable semantic concepts can be represented as directions in a model's representation space, with their contributions combining approximately linearly (Papadimitriou et al., 2025). For example, the representation of an image containing a golden retriever may be approximately decomposed as $h(x) \approx b + c_{\text{dog}} + c_{\text{fur}} + c_{\text{gold color}} + \cdots + \epsilon$, where each concept contribution corresponds to a semantic direction. This view underlies practical concept-decomposition approaches based on sparse autoencoders (Huben et al., 2024; Thasarathan et al., 2025; Rajamanoharan et al., 2024), non-negative matrix factorization (Fel et al., 2023a;b), and related methods (Yeh et al., 2020; Stein et al., 2024). Geometrically, the additive formulation of LRH places concepts in a shared flat space, where concept contributions are globally comparable and compose through vector addition.

At the same time, recent work suggests that semantic concepts are not organized independently, but often exhibit hierarchical structure. For example, Bussmann et al. (2025) showed that enlarging a concept dictionary can induce concept splitting and concept absorption, with hierarchical training alleviating these effects. Costa et al. (2026) developed a hierarchical concept extraction procedure based on Matching Pursuit and reported improved functional performance, while Park et al. (2025) formalized geometric relationships among categorical and hierarchical concepts. Appendix A provides a broader discussion of related work on concept-based interpretability, hierarchical concepts, and geometric representations of hierarchy.

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.

A flat additive geometry and a hierarchy impose different structural constraints: in a hierarchy, a finer concept represents a conditional refinement of its parent, so its contribution naturally depends on the ancestor from which the refinement is made. Hyperbolic geometry has long been studied as a natural representation space for hierarchical structures because its geometry can accommodate hierarchical relationships efficiently (Ganea et al., 2018; He et al., 2024; Chen et al., 2021). The key issue, therefore, is not whether hierarchical concepts can be represented, but whether their hierarchical composition is geometrically compatible with the flat additive structure underlying LRH.

This raises the central question: *to what extent does a candidate geometry support the composition of hierarchical semantic concepts?* We formulate this through *residual reconstruction*, a geometry-agnostic framework for quantifying geometric compatibility. We represent each semantic node by a population-level point in a candidate geometric space and each parent–child refinement by a local tangent vector at its parent. A hierarchical path is then composed recursively by transporting each local concept to the current point and applying the exponential map. The resulting endpoint is compared with the observed descendant point, and their normalized geodesic distance defines the residual reconstruction error. The same framework can therefore be evaluated under different candidate geometries without changing the underlying semantic hierarchy.

The flat additive formulation of LRH arises as a special case of this framework. In Euclidean space, parallel transport is the identity and the exponential map reduces to vector addition. Consequently, the residual reconstruction error directly measures the extent to which hierarchical concept contributions compose additively. Under ideal flat LRH–hierarchy compatibility, this residual should be close to zero. We contrast this with a hyperbolic instantiation in a Poincaré ball, where the same hierarchical concept structure is composed through geometry-aware geodesic operations. Importantly, the semantic hierarchy and its underlying concept definitions remain fixed across the two settings; only their geometric realization and composition rule change.

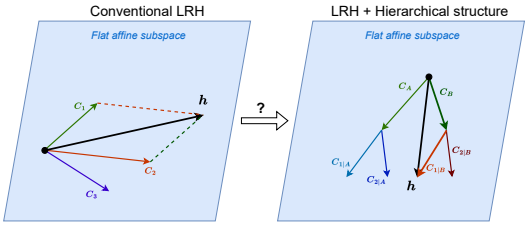

Figure 1: Conceptual illustration of hierarchical concept composition under different geometric structures.

We instantiate the framework using ImageNet-1K, its WordNet hierarchy, and four frozen vision backbones. Across 460 ancestor–descendant pairs, flat additive composition produces residual reconstruction errors that remain substantially above the near-zero ideal predicted by the flat LRH formulation. In contrast, the Poincaré instantiation yields lower residual reconstruction error on every backbone. These results provide empirical evidence that the hierarchical concept structures examined here are not well captured by the tested flat additive geometry and are more naturally represented by the hyperbolic geometry considered in this work.

In summary, our main contributions are as follows:

- *Residual reconstruction*: a geometry-agnostic framework for quantitatively evaluating whether a candidate geometry supports the composition of hierarchical semantic concepts.
- Flat additive LRH is the Euclidean instantiation of this framework, giving a direct, testable criterion for LRH–hierarchy compatibility.
- On ImageNet-1K/WordNet across four frozen vision backbones (460 ancestor–descendant pairs), flat reconstruction error stays well above the near-zero ideal.
- A Poincaré-ball instantiation yields lower residual reconstruction error on all four backbones, indicating hyperbolic geometry better fits this hierarchical structure.

2 Geometric Framework and Hypotheses

2.1 Semantic Concepts and Hierarchies in Vision Models

Concepts. Inspired by the definition of concepts in LLM representations proposed by Park et al. (2025), we refine and generalize this formulation to define concepts in visual models. We define

a concept as a human-recognizable semantic attribute w. At inference time, an input x activates a subset of semantic attributes $\mathcal{W}$ within the model, which causally influence the output y, yielding the conceptual relation $x \xrightarrow{\text{activate}} \mathcal{W} \xrightarrow{\text{lead to}} y$.

We first define a valid semantic attribute w. Semantic attributes may be binary, contrastive, or polytope-based, but all such forms share a categorization-based criterion: a semantic attribute exists only if some subset of samples instantiates it.

A valid semantic attribute can be grounded either in externally specified sample semantics or in the model's internal representation structure; given the diversity of vision backbones studied here, we adopt the sample-based definition below.

Based on this formulation, we define a valid semantic attribute as follows:

Definition 1 (Valid semantic attribute)**.** *Let $\mathcal{X}$ denote the complete input space. For an input $x \in \mathcal{X}$ and a semantic attribute w, we write $x \models w$ if x instantiates the semantics associated with w. The support of w in the input space is then defined as $\mathcal{X}(w) = \{x \in \mathcal{X} : x \models w\}$. If $\mathcal{X}(w) \neq \varnothing$, we say that w is grounded in the input space and constitutes a valid semantic concept.*

Having established that w is a valid semantic attribute grounded in the input space, we next specify when a vector in the model's representation space can be regarded as a representation of w.

Definition 2 (Concept vector)**.** *Let w be a valid semantic attribute according to Definition 1, and $v_w \in \mathbb{R}^d$ denote a direction associated with w in representation space; $||v_w||_2 = 1$. For a scalar strength $\alpha \in \mathbb{R}$, we define $\mathbf{c}_w = \alpha v_w$. We call $\mathbf{c}_w$ a concept vector of w if intervening on a representation h by $h \mapsto h + \mathbf{c}_w$ induces a model-level prediction change associated with w.*

When v_w is indeed captured by the backbone, i.e., the intervention based on $\mathbf{c}_w$ is effective, we index its corresponding semantic attribute as w_i and simplify v_{w_i} to v_i. By setting $\alpha = a_i(x)$, where $a_i(x)$ quantifies the activation of semantic attribute w_i for input x, its contribution to the representation of x is given by $\mathbf{c}_i(x) := a_i(x)v_i$.

Hierarchical concepts. We next consider semantic attributes organized hierarchically through inclusion of their input supports. Let $\mathcal{W}$ be a collection of valid semantic attributes. For $t, r \in \mathcal{W}$, we write $t \prec r$ when $\mathcal{X}(t) \subsetneq \mathcal{X}(r)$, so t is a finer attribute nested within r. A semantic hierarchy is a partially ordered collection of concepts whose relations are determined by nested input supports; we denote it by $\mathcal{H} = (\mathcal{W}, \prec)$. To describe the refinement from r to t, we introduce a conditional semantic attribute that distinguishes t within the domain of r.

Definition 3 (Hierarchical semantic attribute and concept vector)**.** *Let t and r be two valid semantic attributes satisfying $t \prec r$. We define a binary variable that distinguishes t from the remainder of the parent population r, as*

$$w_{t|r}(x) = \begin{cases} 1, & x \in \mathcal{X}(t), \\ 0, & x \in \mathcal{X}(r) \setminus \mathcal{X}(t). \end{cases}$$

A vector $\mathbf{c}_{t|r} \in \mathbb{R}^d$ satisfying Definition 2 for this conditional semantic attribute $w_{t|r}$, with interventions evaluated on $\mathcal{X}(r)$, is the hierarchical concept vector of $w_{t|r}$.

For example, if r denotes the semantic attribute of "Dog" and t denotes the semantic attribute of "Golden Retriever", then $w_{t|r}$ could correspond to the conditional attribute of "Golden fur".

If no valid attribute w satisfies $t \prec w \prec r$, then $r \to t$ is a direct edge, and we call t a direct child concept of r, and r the direct parent concept of t. Along a path, we abbreviate its hierarchical concept vector as $\mathbf{c}_{i|i-1}$.

2.2 Compositional-manifold Hypothesis

Motivated by the manifold hypothesis and prior concept-based interpretability work, we posit that backbone representations concentrate near a low-dimensional Riemannian manifold, and that human-interpretable concepts admit geometry-respecting realizations on it whose composition reconstructs the manifold component of a representation. Our goal is thus to find a candidate manifold whose intrinsic composition rule lets identified concepts explain backbone representations.

Formally, let $h : \mathcal{X} \to \mathbb{R}^d$ be the backbone representation, let $(\mathcal{M}, g)$ be a connected Riemannian manifold, and let $\pi_{\mathcal{M}} : \mathbb{R}^d \to \mathcal{M}$ denote the geometry-dependent map that realizes an ambient representation in $\mathcal{M}$. For an input x, write $p_x = \pi_{\mathcal{M}}(h(x))$. Given a common base point $p_0 \in \mathcal{M}$ shared by the representations under consideration, a finite collection of concept realizations $U_x = \{u_i : i \in I_x\}$, and a geometry-dependent composition operator $\mathrm{Comp}_{\mathcal{M}}$, the compositional-manifold hypothesis requires

$$\widehat{p}_x := \mathrm{Comp}_{\mathcal{M}}(p_0; U_x) \approx p_x, \qquad \text{or equivalently} \qquad d_{\mathcal{M}}(\widehat{p}_x, p_x) \approx 0. \tag{1}$$

Here, u_i denotes the geometry-specific realization of the fixed ambient concept vector $\mathbf{c}_i$ under the candidate manifold $(\mathcal{M}, g)$. The realization of a hierarchical concept vector $\mathbf{c}_{t|r}$ is denoted by $u_{t|r} \in T_{p_r}\mathcal{M}$, as defined formally below.

We extend the input-level formulation above to the conditional population $\mathcal{X}(w)$ of any valid semantic attribute w. Its population representation is $H^w := \mathbb{E}[h(X) \mid X \in \mathcal{X}(w)]$, and applying the same geometry-dependent realization map gives $p_w := \pi_{\mathcal{M}}(H^w) \in \mathcal{M}$. Replacing the input-level quantities $(h(x), p_x, U_x)$ with their population-level counterparts (H^w, p_w, U_w) then yields

$$\widehat{p}_w := \mathrm{Comp}_{\mathcal{M}}(p_0; U_w) \approx p_w, \qquad \text{or equivalently} \qquad d_{\mathcal{M}}(\widehat{p}_w, p_w) \approx 0. \tag{2}$$

2.3 Path Composition and Residual Reconstruction

Path composition. We now consider the setting in which concepts are organized hierarchically. Suppose that the concepts are organized by a semantic hierarchy $\mathcal{H} = (\mathcal{W}, \prec)$ satisfying Definition 3. Restricting attention to a single ancestor–descendant path fixes the admissible order of composition. We write Exp and PT for the exponential map, and parallel transport on $(\mathcal{M}, g)$, respectively.

Definition 4 (Geometric realization of a semantic hierarchy)**.** *A realization of $\mathcal{H}$ in $(\mathcal{M}, g)$ associates every semantic node w with its population-level point $p_w = \pi_{\mathcal{M}}(H^w) \in \mathcal{M}$, and realizes every fixed ambient hierarchical concept vector $\mathbf{c}_{t|r}$ as a parent-local vector $u_{t|r} \in T_{p_r}\mathcal{M}$ for each direct edge $r \to t$. Thus, p_w is the geometric realization of a semantic node, while $u_{t|r}$ is the geometric realization of its conditional refinement.*

For any contiguous subpath from w_a to w_b, let $\mathcal{C}^{\mathcal{M}}_{a:b} : \mathcal{M} \to \mathcal{M}$ denote the composition map induced by successively applying the realized edge concepts along that subpath, with $\mathcal{C}^{\mathcal{M}}_{a:a} = \mathrm{Id}$. For an ancestor $r = w_q$ and descendant $t = w_k$, we evaluate only the suffix composition starting from the observed ancestor point p_r. This removes reconstruction error accumulated along the prefix from p_0 to p_r, isolating the geometric consistency of the descendant refinements from r to t. Appendix B derives this conditional rollout from the factorization of the global path.

For a path $P(r, t) : w_q = r \to w_{q+1} \to \cdots \to w_k = t$, we therefore set $\widehat{p}_q = p_r$. At step i, the edge concept is defined at its true parent, whereas the rollout is at $\widehat{p}_{i-1}$. We transport the concept along the selected minimizing geodesic and apply the exponential map:

$$\widetilde{u}_{i|i-1} = \mathrm{PT}_{p_{w_{i-1}} \to \widehat{p}_{i-1}}(u_{i|i-1}), \qquad \widehat{p}_i \;= \mathrm{Exp}_{\widehat{p}_{i-1}}(\widetilde{u}_{i|i-1}), \quad i = q+1, \ldots, k. \tag{3}$$

The endpoint of this recursion is $\widehat{p}_k = \mathcal{C}^{\mathcal{M}}_{q:k}(p_r)$. We consider rollouts for which these operations are well defined and the selected minimizing geodesics are unique. The geometry determines transport, stepping, and distance; the composition principle is shared.

Definition 5 (Residual reconstruction error)**.** *For $p_r \neq p_t$, define*

$$\mathcal{E}_{\mathcal{M}}(r, t) := \frac{d_{\mathcal{M}}\big(\mathcal{C}^{\mathcal{M}}_{q:k}(p_r), p_t\big)}{d_{\mathcal{M}}(p_r, p_t)} = \frac{d_{\mathcal{M}}(\widehat{p}_k, p_t)}{d_{\mathcal{M}}(p_r, p_t)}. \tag{4}$$

The realized geometry is ideally compatible with this path when $\mathcal{E}_{\mathcal{M}}(r, t) = 0$.

Proposition 1 (Isometry invariance)**.** *If $\psi : (\mathcal{M}, g) \to (\mathcal{N}, h)$ is a Riemannian isometry and $p'_w = \psi(p_w)$, $u'_{t|r} = d\psi_{p_r}(u_{t|r})$, then $\mathcal{E}_{\mathcal{N}}(r, t) = \mathcal{E}_{\mathcal{M}}(r, t)$.*

An isometry preserves distance, exponential maps, and parallel transport, so residual reconstruction measures the realized structure rather than a chosen coordinate system.

We now specialize this geometry-agnostic framework to the flat Euclidean geometry underlying the additive formulation of the LRH. This specialization yields a directly testable reconstruction criterion for hierarchical concept composition.

2.4 EUCLIDEAN INSTANTIATION UNDER LRH

The LRH admits several formulations of varying strength; we adopt the version underlying most practical interpretability methods, in which concept contributions to an internal representation combine approximately linearly. Let $v_j \in \mathbb{R}^d$ be the direction of the j-th concept and let $a_j(x)$ quantify its strength for input x. Writing the unexplained component explicitly, the practical LRH decomposition is

$$h(x) = b + \sum_{j=1}^{J} a_j(x) v_j + \epsilon(x) = b + Va(x) + \epsilon(x), \tag{5}$$

where $V = [v_1, \ldots, v_J]$, $a(x) = [a_1(x), \ldots, a_J(x)]^\top$, b is a baseline representation, and $\epsilon(x)$ collects the component not explained by the linear concept decomposition. Thus, scalar multiplication represents the strength of an individual concept, while vector addition represents the simultaneous contribution of multiple concepts. The concept-explained component consequently lies in a common flat affine space.

If the LRH and the hierarchical concepts are compatible, then every concept direction in Equation (5) that participates in a hierarchical reconstruction must realize a conditional semantic attribute $w_{t|r}$ from Definition 3. Restricting the LRH dictionary to the concepts used in the hierarchy, compatibility therefore requires $\{v_j\}_{j=1}^{J} = \{v_{t|r} : r \to t\}$. We may therefore reindex these directions as $v_{t|r}$, with $\mathbf{c}_{t|r} = \alpha_{t|r} v_{t|r}$ by Definition 2. The resulting flat realization is

$$\begin{aligned} \mathcal{S} &:= \operatorname{span}\{v_{t|r} : r \to t\} \subseteq \mathbb{R}^d, \qquad \Pi_{\mathcal{S}} : \mathbb{R}^d \to \mathcal{S}, \\ (\mathcal{M}, g) &= \big(\mathcal{S}, \langle \cdot, \cdot \rangle\big), \quad \pi_{\mathcal{M}} = \Pi_{\mathcal{S}}, \quad u_{t|r} = \mathbf{c}_{t|r}. \end{aligned} \tag{6}$$

Here $\Pi_{\mathcal{S}}$ is the orthogonal projector. Accordingly, the population-level point $p_w = \pi_{\mathcal{M}}(H^w)$ is already the component retained by the flat concept manifold; no ambient-space component outside $\mathcal{S}$ enters the reconstruction error.

Every tangent space of $\mathcal{S}$ is canonically identified with $\mathcal{S}$, parallel transport is the identity, and $\operatorname{Exp}_p(u) = p + u$. Thus, the general composition reduces to vector addition, yielding the following testable consequence. The full telescoping derivation and its approximation residual are given in Appendix C.

Hypothesis 1 (Flat Additive LRH–Hierarchy Compatibility)**.** *For every nondegenerate ancestor–descendant subpath from $r = w_q$ to $t = w_k$, hierarchical concept contributions compose by vector addition under the flat additive LRH geometry, and the resulting normalized reconstruction residual is small:*

$$\widehat{p}_k = \mathcal{C}^{\mathrm{flat}}_{q:k}(p_r) = p_r + \sum_{i=q+1}^{k} \mathbf{c}_{i|i-1}, \quad \mathcal{E}_{\mathrm{rec}}(r,t) := \frac{\|\widehat{p}_k - p_t\|_2}{\|p_t - p_r\|_2} = \frac{\left\| \sum_{i=q+1}^{k} \mathbf{c}_{i|i-1} - (p_t - p_r) \right\|_2}{\|p_t - p_r\|_2} \approx 0. \tag{7}$$

3 EMPIRICAL INSTANTIATION OF THE FLAT LRH

3.1 SETUP

Models. We evaluate four frozen pretrained vision encoders: ConvNeXt-B (Liu et al., 2022), ViT-B/16 (Kolesnikov et al., 2021), DINOv2 ViT-B/14 (Oquab et al., 2024), and CLIP ViT-B/32 (Radford et al., 2021). For each encoder, we extract its final image representation and train a linear ImageNet-1K classifier as a fixed readout for the intervention-based evaluations. The encoder parameters remain frozen throughout all experiments.

Data. We use the ImageNet-1K training and validation sets (Deng et al., 2009). WordNet (Miller, 1995) provides the semantic attributes and their hierarchy: for each WordNet concept w, its support consists of images whose ImageNet leaf labels are descendants of the corresponding synset. A proof that this construction satisfies the definitions of valid semantic attributes and hierarchical relations is provided in Appendix D.

To better accommodate population-level representation estimation and the extraction of visual hierarchical concepts, we curate the original WordNet structure associated with ImageNet-1K. The

resulting hierarchy contains 246 parent–child concept relations, 272 unique internal semantic nodes, and 214 semantic paths. Full curation details are provided in Appendix D.

3.2 Estimating Semantic Nodes and Hierarchical Concepts

Let $f(x) \in \mathbb{R}^d$ denote the frozen representation of image x. We work in the centered coordinate system $z(x) = f(x) - \bar{\mu}$, where $\bar{\mu}$ is the leaf-balanced mean over the 1,000 ImageNet classes. The ambient representation space is equipped with the standard Euclidean inner product $\langle a, b\rangle = a^\top b$.

Semantic node. For a WordNet concept w, let $\mathcal{D}(w)$ denote the set of its descendant ImageNet leaf classes. We estimate its population-level representation using leaf-balanced weighting:

$$\mu_w = \frac{1}{|\mathcal{D}(w)|} \sum_{\ell \in \mathcal{D}(w)} \frac{1}{N_\ell} \sum_{n: y_n = \ell} z(x_n),$$

where ℓ indexes an ImageNet leaf class and N_ℓ is the number of training images from that class. We set $H^w = \mu_w$ as the empirical population-level representation of the semantic attribute w. Under the flat realization in Section 2.4, the corresponding manifold point is therefore $p_w = \pi_{\mathcal{M}}(H^w) = \Pi_{\mathcal{S}}\mu_w$.

Hierarchical concepts. For each retained edge $w_{i-1} \to w_i$, we estimate the conditional concept direction by discriminating the target child w_i from its siblings within the parent w_{i-1}. We use linear discriminant analysis (LDA), with a shared within-leaf covariance $\Sigma_{i-1}^{\text{sib}}$ estimated over the retained children $\mathcal{C}(w_{i-1})$ using leaf-balanced weighting and Ledoit–Wolf shrinkage. Writing $\mu_i = \mu_{w_i}$, the pairwise LDA direction comparing the target child w_i with a sibling s is

$$v_{i,s} = \operatorname{norm}\big[(\Sigma_{i-1}^{\text{sib}})^\dagger(\mu_i - \mu_s)\big].$$

Under the shared Euclidean metric, the inverse Riesz map is the identity, so the LDA discriminant covector has the same coordinates as its corresponding representation-space direction.

We aggregate the pairwise directions against the other siblings as $v_{i|i-1} = \operatorname{norm}(\sum_{s \in \mathcal{C}(w_{i-1}),\, s \neq w_i} \omega_{s|i} v_{i,s})$, where $\omega_{s|i}$ is normalized proportionally to $|\mathcal{D}(s)|$ over the non-target siblings. This yields the empirical estimate of the conditional direction $v_{w_i|w_{i-1}}$ defined in Section 2.4.

To estimate its population-level strength, we form the leaf-balanced centroid of the retained children, $\mu_{i-1}^{\text{sib}} = \sum_{s \in \mathcal{C}(w_{i-1})} \lambda_s \mu_s$, where λ_s is normalized proportionally to $|\mathcal{D}(s)|$ over all retained children. We then set

$$\alpha_{i|i-1} := \left\langle \mu_i - \mu_{i-1}^{\text{sib}},\, v_{i|i-1} \right\rangle, \qquad \mathbf{c}_{i|i-1} := \alpha_{i|i-1} v_{i|i-1}.$$

Thus, $v_{i|i-1}$ specifies the conditional semantic direction, while $\alpha_{i|i-1}$ specifies its population-level strength. Under the flat Euclidean realization, $u_{i|i-1} = \mathbf{c}_{i|i-1}$.

3.3 Validating Population Representations and Hierarchical Concepts

Population-level representations. We first test whether the empirical population representation $H^w = \mu_w$ retains the semantics of w under the model's fixed ImageNet readout. Because μ_w is expressed in the centered coordinate system, we evaluate $W(\mu_w + \bar{\mu}) + b$ and count a top-k prediction as correct when it contains an ImageNet leaf label in $\mathcal{D}(w)$. This validation precedes the geometry-dependent realization and involves no classifier retraining. As shown in Table 1(a), the top-1 descendant hit rate ranges from 90.8% to 96.7%, while the top-5 rate is at least 99.6% across all four backbones.

Hierarchical concepts. We perform two complementary validations of the estimated hierarchical concept directions. First, we test whether each LDA direction discriminates the corresponding conditional attribute on held-out data. Directions are estimated on the training split, and validation images within $\mathcal{X}(w_{i-1})$ are scored by $\langle z(x), v_{i|i-1}\rangle$. Table 1(b) reports AUC and average precision (AP), together with random-direction and within-parent label-permutation controls. The LDA directions transfer strongly to held-out data across all four backbones, whereas both controls remain near chance in AUC.

Table 1: Semantic validity of population-level representations and hierarchical concept directions across backbones. (a) Fixed-readout descendant prediction from 272 empirical internal-node means $H^w = \mu_w$. (b) Held-out conditional-attribute discrimination averaged over 246 hierarchical concept directions, with random-direction and within-parent label-permutation controls; entries are AUC/AP.

(a) Population-level representations

Model	Top-1	Top-5
ConvNeXt-B	96.69%	100.00%
ViT-B/16	90.81%	99.63%
DINOv2-B/14	91.18%	99.63%
CLIP-B/32	91.18%	100.00%

(b) Hierarchical concepts

Model	LDA	Random	Label perm.
ConvNeXt-B	.98/.97	.49/.33	.48/.36
ViT-B/16	.98/.96	.50/.33	.48/.36
DINOv2-B/14	.98/.96	.50/.32	.48/.37
CLIP-B/32	.97/.92	.49/.32	.49/.36

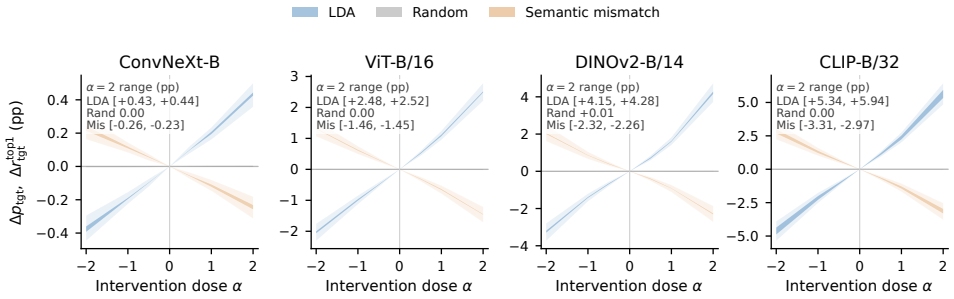

Figure 2: Hierarchical-concept intervention validity. Relative to $\alpha_{\text{int}} = 0$, Δp_{tgt} denotes the change in target-child conditional probability and $\Delta r^{\text{top1}}_{\text{tgt}}$ the change in its parent-conditioned top-1 rate among sibling children, both in percentage points (pp). At each dose, the dark band spans their means, and the light band spans their 95% confidence envelope across concepts. Top-left annotations report the corresponding mean-response range at highest standardized dose.

Second, we test the intervention criterion in Definition 2. After converting the LDA discriminant covector into the corresponding representation-space direction $v_{i|i-1}$, we intervene as $h \longmapsto h + \alpha_{\text{int}} v_{i|i-1}$. Under the fixed ImageNet readout, positive α_{int} should increase the target child's conditional probability and parent-conditioned top-1 rate, whereas negative α_{int} should decrease them. Random and same-parent mismatched directions serve as controls. Figure 2 shows a consistent signed dose response for the LDA directions across all four backbones, while random directions remain near zero and mismatched directions produce responses in the opposite direction. Together, these results support $v_{i|i-1}$ as a valid hierarchical concept direction under Definition 2. More details are provided in Appendix E.

3.4 Flat Residual Reconstruction under the LRH

We next evaluate the residual reconstruction error $\mathcal{E}_{\text{rec}}$ from Hypothesis 1 using the realized semantic-node points p_w and the hierarchical concept vectors $\mathbf{c}_{i|i-1}$ estimated above. The reconstruction is performed entirely within the flat concept manifold $\mathcal{S} = \text{span}\{v_{i|i-1}\}$. As a supplementary diagnostic, Appendix F reports the fraction of the population-level representations retained by this concept manifold.

Figure 3 reports all 460 retained ancestor–descendant pairs. The mean residual reconstruction errors are 0.77, 0.80, 0.76, and 0.84 for ConvNeXt-B, ViT-B/16, DINOv2-B/14, and CLIP-B/32, respectively. The reconstructed endpoints thus remain within the parent-to-target displacement scale but far from the zero-residual ideal of exact flat compatibility. The error generally increases with path length. The small exception is the CLIP-B/32 length-four group, whose mean is slightly lower than that of the length-three group; these groups contain different node pairs, and their uncertainty intervals overlap. Finally, a common rescaling of every path sum is minimized at $\alpha_{\text{rec}} = 1$ for all four backbones, so the residual cannot be substantially reduced by globally adjusting the magnitude of the estimated concept vectors.

These results show measurable but incomplete reconstruction ability. Because WordNet is a finite semantic inventory and the concept directions are empirical estimates, we read this not as a univer-

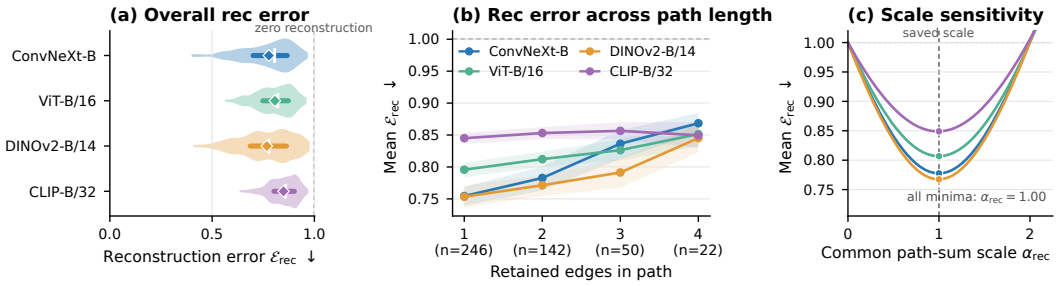

Figure 3: Reconstruction of internal-node displacements within the subspace spanned by the 246 retained hierarchical concept directions. (a) Distributions over all 460 ancestor–descendant pairs; violin width shows pair density, thick colored segments show interquartile ranges, white ticks show medians, and diamonds show means. (b) Mean error by path length; parentheses give the numbers of node pairs and light bands are 95% nonparametric path-bootstrap intervals. (c) Scale sensitivity when the complete path sum, regardless of its length, is multiplied by one common α_{rec}. Here, α_{rec} multiplies each $\mathbf{c}_{i|i-1}$ in the path sum and is different from the intervention dose α_{int}.

sal rejection of LRH but as evidence that the flat Euclidean realization evaluated here has limited capacity to compose these hierarchical concepts.

4 Poincaré Instantiation of Residual Reconstruction

Building on the flat-LRH results, we instantiate the same residual reconstruction framework in a Poincaré ball, whose volume grows exponentially with radius and so represents tree-like relationships efficiently (Ganea et al., 2018; He et al., 2024). The semantic hierarchy and hierarchical concepts are unchanged from Section 3.4: we learn a geometric realization in which each internal semantic node is mapped to a point in the Poincaré ball and each ambient concept vector $\mathbf{c}_{i|i-1}$ is realized as a tangent vector at its parent, so only the geometric realization and composition operations differ from the flat setting.

Formally, let

$$\mathbb{B}_c^m = \{p \in \mathbb{R}^m : c\|p\|_2^2 < 1\}, \qquad g_p^c = \left(\frac{2}{1 - c\|p\|_2^2}\right)^2 I$$

be an m-dimensional Poincaré ball of curvature $-c$, where $c > 0$ and g_p^c denotes the Riemannian metric tensor at p. We learn an encoder $E_\theta : \mathbb{R}^d \to \mathbb{B}_c^m$ and a decoder $D_\phi : \mathbb{B}_c^m \to \mathbb{R}^d$, together with the curvature c. The resulting realization of Definition 4 is

$$(\mathcal{M}, g) = (\mathbb{B}_c^m, g^c), \qquad \pi_{\mathcal{M}} = E_\theta, \qquad p_w = E_\theta(H^w) = E_\theta(\mu_w).$$

For each hierarchical edge, its fixed ambient concept vector $\mathbf{c}_{i|i-1}$ is realized as a parent-local tangent vector $u_{i|i-1} \in T_{p_{w_{i-1}}} \mathbb{B}_c^m$ by matching its image under the decoder Jacobian to $\mathbf{c}_{i|i-1}$. The precise tangent-coordinate optimization, chart-fitting constraints, and implementation details are given in Appendix G.

With this realization, the general path-composition map in Equation (3) is instantiated using Poincaré parallel transport and exponential maps. Writing $d_{\mathbb{H}}$ for the corresponding geodesic distance, the residual reconstruction error becomes

$$\mathcal{E}_{\mathbb{H}}(r, t) := \frac{d_{\mathbb{H}}\big(\mathcal{C}_{q:k}^{\mathbb{H}}(p_r), p_t\big)}{d_{\mathbb{H}}(p_r, p_t)} = \frac{d_{\mathbb{H}}(\widehat{p}_k, p_t)}{d_{\mathbb{H}}(p_r, p_t)}. \tag{8}$$

We evaluate the 214 paths containing at least two retained edges. Direct parent–child edges are used to fit the geometric realization but are not themselves used to evaluate multi-edge composition. In particular, the multi-edge paths used for evaluation are excluded from training and model selection. All fitted chart parameters, curvature, and LDA-derived concept vectors are frozen before evaluation, which uses independently estimated ImageNet-validation node representations.

Permutation control. To test whether reconstruction depends on the semantic correspondence between hierarchical edges and concept vectors, we freeze the trained encoder, decoder, curvature, and realized node points $\{p_w\}$, and permute only the assignment of the 246 hierarchical concept vectors to WordNet edges. Each reassigned concept vector is then realized as a tangent vector at its new parent using Equation (25) and composed using the same Poincaré operations. The flat control applies the identical global permutation before vector addition. Thus, the real and permuted conditions differ only in the edge–concept correspondence, with no retraining of the geometric model. This tests whether reconstruction depends on the semantic correspondence rather than the mere collection of concept vectors.

Results. Table 2 reports the mean relative path-reconstruction error, restricted to the 214 multi-edge paths, under the flat and Poincaré realizations (this flat baseline is recomputed on the multi-edge subset and differs slightly from the all-460-pair mean in Section 3.4). Sequential Poincaré composition produces lower intrinsic residual error for all four backbones: mean errors range from 0.516 to 0.600, versus 0.741–0.823 for the corresponding flat evaluation.

Table 2: Mean relative path-reconstruction error over 214 ImageNet-val multi-edge paths (lower is better). Flat uses Eq. (7); Poincaré uses Eq. (8) and is averaged over three chart initializations.

Model	Flat	Poincaré	Flat perm.	Poincaré perm.
ConvNeXt-B	.749	**.600**	1.682	1.655
ViT-B/16	.768	**.516**	1.627	1.863
DINOv2-B/14	.741	**.559**	1.472	1.413
CLIP-B/32	.823	**.571**	1.518	1.622

Because the two residuals are computed in different carrier geometries, their numerical values should not by themselves be interpreted as a direct cross-geometry comparison. We therefore also perform a common-space comparison by decoding the reconstructed Poincaré endpoints back into the original representation space, as described in Appendix G. The resulting comparison remains favorable to the Poincaré realization for all four backbones.

Permuting the hierarchical concept assignments increases the Poincaré error from 0.516–0.600 to 1.413–1.863, indicating that the low reconstruction error depends on the correspondence between hierarchical relations and their associated concept vectors rather than merely on the multiset of vectors. The fitted curvature and additional geometric diagnostics are reported in Appendix G.

5 Conclusion

We introduced residual reconstruction, a geometry-agnostic framework for testing whether hierarchical concepts compose consistently within a candidate representation geometry: semantic nodes are represented as points, parent–child refinements as local tangent vectors, and descendants are reconstructed through geometry-specific path composition. In Euclidean space this reduces to additive path reconstruction, giving a direct empirical test of compatibility between hierarchical concepts and the flat additive geometry considered by the LRH.

Across four vision backbones, the flat additive realization produced substantial residual reconstruction error that generally increased with path length, remaining far from the near-zero ideal of flat compatibility, while the Poincaré instantiation yielded lower intrinsic error for all four backbones on the same multi-edge paths – an improvement that persisted after decoding back into the common embedding space. Together, these results indicate that the hierarchical concept structures studied here are better captured by the tested Poincaré realization than by the flat additive one.

Our experiments consider a finite WordNet-derived hierarchy, one construction of population-level semantic representations and hierarchical concept directions, and one alternative non-Euclidean geometry, leaving broader hierarchies, alternative concept estimators, and other candidate geometries to future work. More generally, residual reconstruction offers a geometry-agnostic criterion for identifying when a candidate geometry – flat or otherwise – fails to support semantic composition.

AI USE STATEMENT

In this work, we used generative AI tools for literature retrieval, manuscript organization, language polishing, LaTeX formatting, and reference formatting. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bussmann25a.html.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. Probing bert in hyperbolic spaces. In *International Conference on Learning Representations*, 2021.

Valérie Costa, Thomas Fel, Ekdeep S Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. *Advances in Neural Information Processing Systems*, 38:34381–34424, 2026.

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

Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36:54805–54818, 2023a.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023b.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*, pp. 1646–1655. PMLR, 2018.

Yuan He, Zhangdie Yuan, Jiaoyan Chen, and Ian Horrocks. Language models as hierarchy encoders. *Advances in Neural Information Processing Systems*, 37:14690–14711, 2024.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

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 Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

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

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Nghia Nguyen, Tianjiao Ding, and René Vidal. Hierarchical concept embedding & pursuit for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2907–2917, June 2026.

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

Matthew Lyle Olson, Musashi Hinck, Neale Ratzlaff, Changbai Li, Phillip Howard, Vasudev Lal, and Shao-Yen Tseng. Analyzing hierarchical structure in vision models with sparse autoencoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4835–4839, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* , 2024. doi: 10.48550/arxiv.2304.07193. URL https://hal.science/hal-04376640.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. In *Conference on Language Modeling*, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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*, volume 2025, pp. 76441–76463, 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. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Adam Stein, Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. Towards compositionality in concept learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46530–46555. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/stein24b.html.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second International Conference on Machine Learning*, 2025.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20554–20565. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ecb287ff763c169694f682af52c1f309-Paper.pdf.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73918–73956. PMLR, 2025.

A Related Work

Concept-based interpretation and linear representations. Concept-based interpretability seeks to describe model representations and predictions using human-recognizable semantic variables rather than individual input features. Early approaches associated hidden units or representation directions with semantic concepts. Network Dissection measures alignment between individual visual units and a labeled concept vocabulary (Bau et al., 2017), whereas TCAV represents a user-defined concept by a linear direction and evaluates prediction sensitivity along that direction (Kim et al., 2018). Concept bottleneck models instead explicitly predict human-specified concepts before the task output, thereby supporting direct concept interventions (Koh et al., 2020). Subsequent work has developed automatic concept dictionaries using matrix factorization (Fel et al., 2023a;b), sparse autoencoders (Huben et al., 2024; Thasarathan et al., 2025; Rajamanoharan et al., 2024), and objectives promoting completeness or compositionality (Yeh et al., 2020; Stein et al., 2024). These methods differ in how concepts are obtained, but commonly operationalize a concept as a direction or coordinate whose activation can be read out or manipulated. The LRH makes this linear structure explicit and connects concept directions to probing, intervention, and representation geometry (Park et al., 2024); related evidence has also been reported for vision–language embedding spaces (Papadimitriou et al., 2025). Our work does not propose another concept-extraction objective. Instead, it asks whether already identified hierarchical concept contributions compose consistently under a candidate geometry.

Hierarchical and compositional concept structure. Most concept methods initially treat concepts as a flat dictionary, although semantic attributes naturally occur at multiple levels of abstraction. Recent work has begun to model this structure directly. Matryoshka sparse autoencoders learn nested dictionaries intended to retain both broad and fine-grained features (Bussmann et al., 2025); hierarchical sparse autoencoders and related analyses study multilevel structure in vision representations (Zaigrajew et al., 2025; Olson et al., 2025). Hierarchical Concept Embedding and Pursuit introduces an explicitly hierarchical representation for interpretable image classification (Nguyen et al., 2026), while hierarchical Matching Pursuit extracts coarse-to-fine sparse representations (Costa et al., 2026). Complementarily, compositional concept learning asks whether several concepts jointly explain a representation or prediction (Stein et al., 2024). Our setting connects these themes: each edge is a conditional refinement of its parent, and compatibility is evaluated by composing such refinements along an ancestor–descendant path rather than by judging individual concept quality alone.

Geometry of hierarchical representations. Hyperbolic spaces have long been used to represent tree-like and taxonomic structure because their volume growth can accommodate rapidly expanding hierarchies. Poincaré embeddings demonstrated this principle for symbolic hierarchical data (Nickel & Kiela, 2017), and entailment cones further modeled asymmetric hierarchical relations (Ganea et al., 2018). Hyperbolic probing and representation analyses have subsequently identified hierarchical organization in language-model representations (Chen et al., 2021; He et al., 2024). More recently, Park et al. (2025) studied categorical and hierarchical concepts through their geometric relationships. These works primarily learn hierarchical embeddings or probe whether hierarchy is encoded. In contrast, residual reconstruction holds the semantic hierarchy and its ambient concept vectors fixed, realizes them under a candidate geometry, and tests whether local parent–child refinements reconstruct observed descendants. This provides a common composition-based criterion whose Euclidean instance directly tests the flat additive LRH and whose Poincaré instance evaluates a geometry adapted to hierarchical structure.

B Conditional Path Composition from an Observed Ancestor

Consider a root-to-node path $w_0 \to w_1 \to \cdots \to w_k$, and let $r = w_q$ and $t = w_k$. For any contiguous subpath from w_a to w_b, let $\mathcal{C}^{\mathcal{M}}_{a:b}$ denote the composition map obtained by successively applying the realized edge concepts along that subpath. Because path composition is recursive, these maps satisfy

$$\mathcal{C}^{\mathcal{M}}_{0:k} = \mathcal{C}^{\mathcal{M}}_{q:k} \circ \mathcal{C}^{\mathcal{M}}_{0:q}. \tag{9}$$

In the exact form of the compositional-manifold hypothesis, the global reconstructions of r and t from the common base point p_0 satisfy

$$p_r = \mathcal{C}^{\mathcal{M}}_{0:q}(p_0), \qquad p_t = \mathcal{C}^{\mathcal{M}}_{0:k}(p_0). \tag{10}$$

Substituting Equation (9) into Equation (10) gives

$$p_t = \mathcal{C}^{\mathcal{M}}_{q:k}\big(\mathcal{C}^{\mathcal{M}}_{0:q}(p_0)\big) = \mathcal{C}^{\mathcal{M}}_{q:k}(p_r). \tag{11}$$

Thus, the prefix from p_0 to p_r can be represented by the observed ancestor point, and conditional reconstruction can be evaluated by applying only the suffix map $\mathcal{C}^{\mathcal{M}}_{q:k}$ from p_r.

C LRH Compatibility as the Euclidean Residual-Reconstruction Instance

We derive Hypothesis 1 entirely within the flat concept manifold. Let $\mathcal{S}$ be the concept subspace in Equation (6). From the outset, take

$$(\mathcal{M}, g) = \big(\mathcal{S}, \langle\cdot,\cdot\rangle\big), \qquad \pi_{\mathcal{M}} = \Pi_{\mathcal{S}}, \qquad u_{i|i-1} = \mathbf{c}_{i|i-1}. \tag{12}$$

The orthogonal projection constitutes the realization map from the ambient representation space to the flat concept manifold.

To express the LRH directly in this geometry, define

$$b_{\mathcal{S}} := \Pi_{\mathcal{S}} b, \qquad \epsilon_{\mathcal{S}}(x) := \Pi_{\mathcal{S}} \epsilon(x). \tag{13}$$

Under the compatibility correspondence in Section 2.4, the columns of V are the directions $v_{t|r}$ and therefore already belong to $\mathcal{S}$. Hence $\Pi_{\mathcal{S}} V = V$, so no separately projected concept matrix is required. For a valid semantic attribute w, also define

$$\bar{a}^w := \mathbb{E}[a(X) \mid X \in \mathcal{X}(w)], \qquad \bar{\epsilon}^w_{\mathcal{S}} := \mathbb{E}[\epsilon_{\mathcal{S}}(X) \mid X \in \mathcal{X}(w)]. \tag{14}$$

By linearity of conditional expectation and of $\Pi_{\mathcal{S}}$, the population-level point p_w defined above satisfies

$$p_w = b_{\mathcal{S}} + V\bar{a}^w + \bar{\epsilon}^w_{\mathcal{S}}. \tag{15}$$

Thus, for a direct edge $w_{i-1} \to w_i$, the displacement intrinsic to the flat concept manifold is

$$p_{w_i} - p_{w_{i-1}} = V\big(\bar{a}^{w_i} - \bar{a}^{w_{i-1}}\big) + \bar{\epsilon}^{w_i}_{\mathcal{S}} - \bar{\epsilon}^{w_{i-1}}_{\mathcal{S}}. \tag{16}$$

Definitions 2 and 3 establish that $\mathbf{c}_{i|i-1}$ is a valid concept vector for the conditional semantic attribute on this edge. They do not by themselves imply that this vector reconstructs the displacement between the corresponding population points. We make the discrepancy explicit by defining

$$\delta_i := \mathbf{c}_{i|i-1} - \big(p_{w_i} - p_{w_{i-1}}\big). \tag{17}$$

Using Equation (16), this residual can also be written as

$$\delta_i = \mathbf{c}_{i|i-1} - V\big(\bar{a}^{w_i} - \bar{a}^{w_{i-1}}\big) - \bar{\epsilon}^{w_i}_{\mathcal{S}} + \bar{\epsilon}^{w_{i-1}}_{\mathcal{S}}. \tag{18}$$

It therefore measures, within $\mathcal{S}$, the mismatch between the population-level LRH increment and the proposed hierarchical concept vector.

Summing Equation (17) along $w_0 \succ w_1 \succ \cdots \succ w_k$ telescopes the population-point differences and gives the exact identity

$$p_{w_k} = p_{w_0} + \sum_{i=1}^{k} \mathbf{c}_{i|i-1} - \rho_{0:k}, \qquad \rho_{0:k} := \sum_{i=1}^{k} \delta_i. \tag{19}$$

When $\rho_{0:k}$ is small, this is the root-to-node case of Hypothesis 1. For an ancestor $r = w_q$ and descendant $t = w_k$, summing only the suffix similarly gives

$$\sum_{i=q+1}^{k} \mathbf{c}_{i|i-1} - (p_t - p_r) = \rho_{q:k}, \qquad \rho_{q:k} := \sum_{i=q+1}^{k} \delta_i. \tag{20}$$

Its small-residual form is the additive reconstruction tested in Equation (7).

It remains to connect this telescoping relation to path composition. Because $\mathcal{S}$ is Euclidean, parallel transport is the identity and $\mathrm{Exp}_p(u) = p + u$. Starting from the observed ancestor point p_r, the recursive rule in Equation (3) therefore yields

$$\widehat{p}_k = p_r + \sum_{i=q+1}^{k} \mathbf{c}_{i|i-1}. \tag{21}$$

Combining Equations (20) and (21) gives

$$d_{\mathcal{M}}(\widehat{p}_k, p_t) = \|\rho_{q:k}\|_2, \qquad d_{\mathcal{M}}(p_r, p_t) = \|p_t - p_r\|_2. \tag{22}$$

Hence the general residual in Definition 5 reduces to

$$\mathcal{E}_{\mathcal{M}}(r,t) = \frac{\|\rho_{q:k}\|_2}{\|p_t - p_r\|_2} = \frac{\left\|\sum_{i=q+1}^{k} \mathbf{c}_{i|i-1} - (p_t - p_r)\right\|_2}{\|p_t - p_r\|_2}. \tag{23}$$

Hypothesis 1 is the prediction that this normalized accumulated residual is near zero.

D Proof of Validity of WordNet-Based Concept and Hierarchy Instantiation

In our instantiation, semantic attributes are defined as WordNet synsets that are not ImageNet-1K leaf classes. For each semantic attribute w, let $D(w)$ denote the set of ImageNet-1K leaf classes descending from its WordNet synset. The satisfaction relation is instantiated as

$$x \models w \iff y(x) \in D(w),$$

where $y(x)$ denotes the ImageNet-1K label of input x. The support of a semantic attribute is therefore

$$\mathcal{X}(w) = \{x \in \mathcal{X} : x \models w\} = \{x \in \mathcal{X} : y(x) \in D(w)\}.$$

Validity under Definition 1. A semantic attribute w is valid if its support is non-empty:

$$\mathcal{X}(w) \neq \varnothing.$$

In our construction, only WordNet synsets with at least one descendant ImageNet-1K class are retained. Therefore, for every retained synset w,

$$\exists x \in \mathcal{X}, \quad y(x) \in D(w).$$

Hence,

$$\mathcal{X}(w) \neq \varnothing,$$

and w satisfies Definition 1.

Hierarchical relations. We next verify that the WordNet hierarchy used in our experiments satisfies the hierarchical relation defined in Section 2.1. Consider two WordNet synsets t and r, where t is a descendant synset of r in the WordNet hierarchy. By the transitivity of the WordNet taxonomy,

$$D(t) \subseteq D(r).$$

For any input $x \in \mathcal{X}(t)$, the satisfaction relation gives

$$x \in \mathcal{X}(t) \iff y(x) \in D(t).$$

Since $D(t) \subseteq D(r)$, we have $y(x) \in D(r)$, and therefore $x \in \mathcal{X}(r)$. Thus,

$$\mathcal{X}(t) \subseteq \mathcal{X}(r).$$

Furthermore, for each parent–child relation retained in our construction, the parent concept contains at least one ImageNet-1K leaf class that does not belong to the descendant set of the child concept:

$$D(r) \setminus D(t) \neq \varnothing.$$

Thus, there exists an input $x \in \mathcal{X}$ with $y(x) \in D(r) \setminus D(t)$. Consequently,

$$x \in \mathcal{X}(r) \quad \text{and} \quad x \notin \mathcal{X}(t),$$

which gives

$$\mathcal{X}(t) \subsetneq \mathcal{X}(r).$$

Therefore, according to the definition of hierarchical concepts,

$$t \prec r.$$

Hence, the retained WordNet-based hierarchy instantiated on ImageNet-1K satisfies the hierarchical relation assumed in the paper.

Hierarchy curation. The full WordNet hierarchy contains $1{,}000$ ImageNet leaf classes and 860 internal semantic nodes. To make this hierarchy suitable for population-level representation estimation and visual hierarchical concept extraction, we first remove internal nodes with only one child or more than 150 descendant ImageNet-1K leaf classes, reconnecting their children to the original parent. For each retained parent, its direct internal children form a sibling group, and each target child is treated as a conditional attribute relative to the remaining siblings.

To avoid highly overlapping child concepts under the same parent, for every pair of children a and b we compute the pairwise containment ratio from their covered ImageNet-1K leaf-class sets:

$$\mathrm{overlap}(a, b) := \frac{|D(a) \cap D(b)|}{\min\{|D(a)|, |D(b)|\}}. \tag{24}$$

We regard a pair as highly overlapping when $\mathrm{overlap}(a, b) \geq 0.5$. We then iteratively remove the child involved in the largest number of highly overlapping pairs until the containment ratio between every pair of retained children is below 0.5. Finally, we retain only sibling groups containing at least two valid children. The curated hierarchy contains 75 parent groups, 246 parent–child concept relations, 272 unique internal semantic nodes, and 214 semantic paths, each containing at least two hierarchical concepts.

E Hierarchical Concept Readout and Intervention Validity Protocols

Held-out readout validity. For each retained edge $w_{i-1} \to w_i$, we estimate $v_{i|i-1}$ using only training features. On the validation split, the positive set contains images in $\mathcal{X}(w_i)$, and the negative set contains images assigned to the other retained children of w_{i-1}. We use $\langle z(x), v_{i|i-1} \rangle$ as the scalar score and report AUC and AP. The random-direction baseline averages 32 isotropic unit directions for every edge. For the label-permutation baseline, we randomly reassign descendant leaf classes among children within each parent while preserving the number of leaves per child, re-estimate LDA on the training split, and evaluate against the original validation labels. We use 20 permutations and macro-average all results over the 246 edges. Because the binary contrasts are generally imbalanced, the AP null need not equal 0.5.

Intervention validity. For the conditional attribute $w_{i|i-1}$, we test Definition 2 using

$$h \longmapsto h + \alpha_{\mathrm{int}} v_{i|i-1}.$$

We evaluate this intervention on held-out images in $\mathcal{X}(w_{i-1})$. The fixed ImageNet linear readout produces leaf logits, which we aggregate within each retained child and normalize over the children of w_{i-1}. We measure changes in the target child's conditional probability, margin, rank, and top-1 rate relative to $\alpha_{\mathrm{int}} = 0$.

The intervention grid is parameterized as $\alpha_{\mathrm{int}} = \beta s_v$, where

$$\beta \in \{-2, -1, -0.5, 0, 0.5, 1, 2\}, \qquad s_v = \mathrm{Std}_{x \in \mathcal{X}_{\mathrm{val}}(w_{i-1})} \left[\langle z(x), v_{i|i-1} \rangle \right].$$

Here, s_v is only a dose-calibration unit used to compare intervention strengths across edges and backbones. It is not the population contribution strength $\alpha_{i|i-1}$ used to construct the vector for the path-composition tests. We compare the real direction with 16 isotropic random directions and with directions belonging to the other children of the same parent.

F Coverage of the WordNet-Defined Concept Space

Residual reconstruction is defined intrinsically on the selected concept manifold, so the orthogonal realization map is already accounted for by the points p_w used in the flat error. Nevertheless, the retained hierarchy is finite, and it is useful to report how much of each original ambient internal-node representation lies outside the resulting concept manifold. We therefore measure, for each of the 272 internal semantic nodes, the norm of the discarded orthogonal component relative to the norm of the original node representation. Table 3 reports the dimension of the resulting subspace and the mean and median of this diagnostic.

Table 3: Relative projection error of the original internal-node representations onto the WordNet-defined concept space. Lower is better.

Model	Subspace dimension	Mean / Median
ConvNeXt-B	238	.847 / .859
ViT-B/16	241	.783 / .790
DINOv2-B/14	240	.727 / .734
CLIP-B/32	240	.725 / .729

The relative projection errors range from (0.725) to (0.847), showing that the retained hierarchical directions capture only part of the original ambient representations. This diagnostic does not alter the residual metric, which evaluates composition directly between the realized points p_w. It instead clarifies the scope of the flat manifold tested here: the reported reconstruction results concern the representation component expressible by the specified hierarchical concept inventory.

G Poincaré Concept-Space Implementation and Ambient-Space Evaluation

For a hierarchical edge $w_{i-1} \to w_i$, we realize its fixed ambient concept vector as a tangent vector at the corresponding parent point by solving

$$u_{i|i-1} := \arg \min_{u \in T_{p_{w_{i-1}}} \mathbb{B}_c^m} \left\| J_{D_\phi}(p_{w_{i-1}})u - \mathbf{c}_{i|i-1} \right\|_2^2 + \lambda \|u\|_2^2. \tag{25}$$

Here, $J_{D_\phi}(p)$ denotes the decoder Jacobian at p, so $u_{i|i-1}$ is the parent-local tangent-space realization of $\mathbf{c}_{i|i-1}$.

The encoder–decoder chart is fitted on ImageNet training statistics subject to three constraints: internal semantic nodes are represented as points on the learned Poincaré manifold, each hierarchical concept is aligned with its local parent–child relation, and the decoder's locally induced metric matches the Poincaré metric. In particular, the latter two constraints target

$$u_{i|i-1} \approx \mathrm{Log}_{p_{w_{i-1}}}(p_{w_i}), \qquad J_{D_\phi}(p)^\top J_{D_\phi}(p) \approx g_p^c.$$

We implement these three constraints through

$$\mathcal{L}_{\text{chart}} = \mathcal{L}_{\text{node}} + \mathcal{L}_{\text{concept}} + \mathcal{L}_{\text{Poincaré}}. \tag{26}$$

Let $\mathcal{R}_{\text{fit}} \subseteq \mathbb{R}^d$ contain the ImageNet-train leaf means and internal-node means used to fit the chart, and let $\mathcal{R}_{\text{edge}}$ denote the 246 retained parent–child edges. For an edge $e = (a \to b)$, write $p_a = E_\theta(\mu_a)$, $p_b = E_\theta(\mu_b)$, $J_a = J_{D_\phi}(p_a)$, and let u_e be the tangent coordinate defined in Equation (25). The node-manifold term is

$$\begin{aligned} \mathcal{L}_{\text{node}} = \lambda_{\text{pt}} \frac{1}{|\mathcal{R}_{\text{fit}}|} \sum_{h \in \mathcal{R}_{\text{fit}}} \frac{\|D_\phi(E_\theta(h)) - h\|_2^2}{\|h\|_2^2 + \epsilon} \\ + \lambda_{\text{ctr}} \frac{1}{m} \left\| \frac{1}{|\mathcal{R}_{\text{ctr}}|} \sum_{h \in \mathcal{R}_{\text{ctr}}} E_\theta(h) \right\|_2^2, \end{aligned} \tag{27}$$

where $\mathcal{R}_{\mathrm{ctr}} \subseteq \mathcal{R}_{\mathrm{fit}}$ is the subset used to center the chart. The parent-local concept term is

$$\begin{aligned}\mathcal{L}_{\mathrm{concept}} = \lambda_{\tan} \frac{1}{|\mathcal{R}_{\mathrm{edge}}|} \sum_{e \in \mathcal{R}_{\mathrm{edge}}} \frac{\|J_a u_e - \mathbf{c}_e\|_2^2}{\|\mathbf{c}_e\|_2^2 + \epsilon} \\ + \lambda_{\mathrm{edge}} \frac{1}{|\mathcal{R}_{\mathrm{edge}}|} \sum_{e=(a \to b) \in \mathcal{R}_{\mathrm{edge}}} \frac{\|u_e - \mathrm{Log}_{p_a}(p_b)\|_2^2}{\|\mathrm{Log}_{p_a}(p_b)\|_2^2 + \epsilon}.\end{aligned} \tag{28}$$

Finally, for sampled chart coordinates $\mathcal{Q}$ and all coordinates $\mathcal{Z}$ occurring in the current batch, the Poincaré-geometry term is

$$\begin{aligned}\mathcal{L}_{\mathrm{Poincaré}} = \lambda_{\mathrm{met}} \frac{1}{|\mathcal{Q}|} \sum_{p \in \mathcal{Q}} \frac{\|J_{D_\phi}(p)^\top J_{D_\phi}(p) - g_p^c\|_F^2}{\|g_p^c\|_F^2} \\ + \lambda_{\mathrm{bdry}} \frac{1}{|\mathcal{Z}|} \sum_{p \in \mathcal{Z}} \left[\max\left(0, c\|p\|_2^2 - \rho_{\max}^2\right)\right]^2.\end{aligned} \tag{29}$$

We use $\lambda_{\mathrm{pt}} = 3$, $\lambda_{\mathrm{ctr}} = 10^{-3}$, $\lambda_{\tan} = 1$, $\lambda_{\mathrm{edge}} = 0.5$, $\lambda_{\mathrm{met}} = 0.5$, $\lambda_{\mathrm{bdry}} = 1000$, $\rho_{\max} = 0.9$, and ridge coefficient 10^{-5} in Equation (25). We set $\epsilon = 10^{-12}$ for numerical stability and use 256 training points for $\mathcal{R}_{\mathrm{ctr}}$. The subsidiary terms implement or stabilize the three constraints in Equation (26); they do not introduce additional modeling assumptions.

The encoder is linear, and the decoder consists of a linear map plus a one-hidden-layer tanh residual. The latent and hidden dimensions are 192 and 128, respectively. We normalize all coordinates by the root-mean-square norm of the ImageNet-train leaf means, initialize from PCA, and jointly optimize the chart and curvature. Checkpoints are selected using held-out ImageNet-train leaf means together with node, concept, edge, and metric losses; the 214 multi-edge paths and all ImageNet-val means remain unseen. Reported Poincaré results average three random initializations. Table 4 reports the resulting chart parameters and fit diagnostics.

Table 4: Fitted Poincaré chart parameters and diagnostics, averaged over three seeds. The sectional curvature is $-c$, $R = 1/\sqrt{c}$, and $\rho = \sqrt{c}\|p\|_2$.

Model	c	R	$\rho_{\max}$	Node E–D	Tangent	Edge	Metric
ConvNeXt-B	.26	1.96	.38	.51	.54	.49	.02
ViT-B/16	.23	2.11	.48	.40	.54	.45	.03
DINOv2-B/14	.18	2.38	.42	.41	.49	.46	.02
CLIP-B/32	.15	2.61	.58	.20	.38	.47	.03

Here, $\rho_{\max}$ is the largest normalized radius over the encoded ImageNet-train leaf and internal-node means; it remains below the imposed limit of 0.90. Node E–D is the mean relative encoding–decoding discrepancy $\|D_\phi(E_\theta(\mu_w^{\mathrm{val}})) - \mu_w^{\mathrm{val}}\|_2 / \|\mu_w^{\mathrm{val}}\|_2$ over independently estimated ImageNet-val node means. Tangent and Edge are the relative tangent-containment and one-edge alignment errors, and Metric is the normalized metric-matching loss. Across the three seeds, the standard deviation is at most 0.010 for c, 0.038 for R, 0.014 for $\rho_{\max}$, and 0.013 for the four fit errors.

For the permutation control, the chart fitted with the correct 246 edge–concept correspondences is kept fixed. One global permutation reassigns the same hierarchical concept vectors to the WordNet edges only at evaluation time; the same reassignment is used for flat and Poincaré composition. In the Poincaré realization, each reassigned vector is mapped to a tangent vector at its new parent using Equation (25). No chart is fitted to the permuted correspondence.

For evaluation in the original embedding space, the flat and Poincaré predicted displacements are, respectively,

$$\widehat{\Delta\mu}_{r,t}^{\ \mathrm{flat}} = \sum_{i=q+1}^{k} \mathbf{c}_{i|i-1}, \qquad \widehat{\Delta\mu}_{r,t}^{\ \mathbb{H}} = D_\phi(\widehat{p}_k) - D_\phi(p_r).$$

For $G \in \{\mathrm{flat}, \mathbb{H}\}$, we compute

$$\mathcal{E}^{G}_{\mathrm{amb}}(r,t) = \frac{\|\widehat{\Delta\mu}^{\,G}_{r,t} - (\mu^{\mathrm{val}}_t - \mu^{\mathrm{val}}_r)\|_2}{\|\mu^{\mathrm{val}}_t - \mu^{\mathrm{val}}_r\|_2}. \tag{30}$$

Table 5 compares both methods in these common ambient coordinates.

Table 5: Mean relative reconstruction error in the original embedding space over the same 214 ImageNet-val multi-edge paths (lower is better). Poincaré entries are averaged over three chart initializations.

Model	Flat	Poincaré	Flat perm.	Poincaré perm.
ConvNeXt-B	.949	**.912**	1.174	1.122
ViT-B/16	.939	**.893**	1.225	1.160
DINOv2-B/14	.913	**.872**	1.192	1.149
CLIP-B/32	.933	**.911**	1.248	1.217