SEMANTIC-FRAME IDENTIFIABILITY AND RECOVERY IN GEOMETRIC REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate distance prediction does not necessarily identify which direction represents hierarchical depth, abstraction, or uncertainty. Geometric models can preserve every pairwise distance while changing a fixed-origin semantic readout, leaving downstream decisions unsupported by predictive accuracy alone. We formalize this distinction as semantic-frame identifiability. For observationally equivalent states with different semantic targets, minimax estimation risk is at least half the target orbit diameter. We derive recovery conditions for sparse, noisy asymmetric evidence and introduce FramePosterior: a decision layer that keeps the native predictor frozen, conditions candidate frames on declared anchors, and marginalizes task-specific answers rather than committing to one root. Exact-isometry audits separate prediction preservation, semantic sensitivity, and actual decision loss. An ImageNet-100 radial-depth readout loses 12.68 percentage points of accuracy, while 32 noisy landmarks recover 89.88% of its depth-MAE gap. On full WordNet, posterior voting improves ordinal accuracy over matched MAP inference by 1.48 percentage points with 16 anchors at 20% corruption. A topology-informed routing atlas increases delivery from 69.90% to 77.92% over a matched no-revisit control. Stable visual readouts delimit the claim: frame ambiguity need not degrade every task. The resulting framework distinguishes what metric observations identify from what semantic evidence must supply, enabling targeted recovery without modifying native predictions.

1 INTRODUCTION

Geometric representations are useful not only for predicting relations, but also for making decisions about hierarchy and uncertainty. In hyperbolic models, distance describes similarity, whereas distance to an origin is often interpreted as hierarchical depth or confidence (Nickel & Kiela, 2017; Chami et al., 2019; Yang et al., 2023; Sinha et al., 2024; Kim et al., 2026). These interpretations answer different questions. Knowing that two concepts are close does not, by itself, tell us which is more general. Even the metric of a three-node path admits either endpoint as its root, reversing all nontrivial depth comparisons without changing a single pairwise distance. A good relation predictor can therefore leave a downstream semantic decision undetermined.

The missing ingredient is a *semantic frame*: the task-relevant reference or orientation that makes a coordinate readout meaningful. This could be a declared taxonomy root, an abstract-to-specific direction, or the reference used by a radial uncertainty score. Under a common hyperbolic isometry, pairwise distances and distance-decoder predictions are unchanged, but distances to a fixed origin can change. Consequently, metric fit alone cannot certify these interpretations. Figure 1 illustrates this distinction: observationally equivalent representations can support incompatible semantic answers. When a system uses those answers to estimate depth or select uncertain samples, the ambiguity can affect its decisions even though ordinary predictive validation detects no change. It need not harm every task: an intrinsic readout or a correctly transported reference can remain stable.

Centering and hierarchy-aware training address different parts of this problem. A data-dependent, equivariant center supplies a consistent geometric reference; identifying it with the intended semantic root requires task-specific evidence. Directed supervision supplies orientation information that symmetric distance observations leave unresolved. Geometric regularity and semantic evidence are therefore distinct resources. We ask: *which semantic decisions can the available observations iden-*

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.

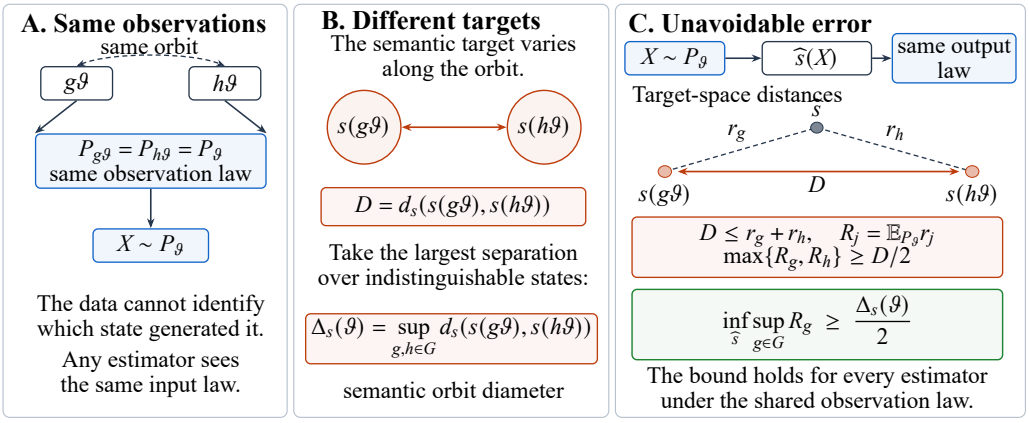

Figure 1: Orbit lower bound. Observationally equivalent states may have distinct semantic targets. If their target distance is D, at least one expected estimation risk is at least $D/2$; the same bound applies to decision disagreement.

tify, and how much asymmetric evidence is needed to resolve the remaining ambiguity? Crucially, two possible roots need not be distinguished if they give the same answers to the requested queries. This decision-level formulation avoids demanding unnecessary coordinate recovery.

We first quantify this limitation through a group-orbit lower bound and characterize recovery from sparse, noisy anchors. We then introduce *FramePosterior*, a semantic decision layer that leaves the trained native predictor frozen. Rather than committing prematurely to one root, it weights compatible frames using declared evidence and marginalizes the requested decision over rooted charts; insufficient agreement motivates another query or abstention. Figure 2 shows the resulting separation of geometry and semantics on WordNet. This design retains successful embeddings while making their downstream interpretation evidence-dependent and uncertainty-aware.

Our controlled audits separate coordinate sensitivity from actual task loss. Exact interventions preserve native predictions, yet the ImageNet-100 radial-depth readout loses 12.68 percentage points of accuracy; 32 noisy landmarks recover 89.88% of its depth-MAE gap. WordNet compares posterior voting with a single MAP frame under matched evidence, while CIFAR and transported-center controls identify stable readouts. Together, these studies locate the practical effect in decisions that depend on an unresolved semantic frame.

This paper formalizes *semantic-frame identifiability*: which observations determine a frame, up to which symmetry, and for which decisions? Our contributions are:

1. **Decision-level identifiability.** We formulate semantic frames relative to an explicit observation law and query family, and quantify unavoidable decision error by an orbit lower bound.
2. **Evidence requirements.** We derive noisy-anchor recovery bounds and a finite-DAG decision-resolving condition, connecting anchor selection to resolving sets, set cover, and equivalence-class determination under their stated assumptions.
3. **FramePosterior.** We isolate frame inference from the frozen native predictor, derive conditional decision-risk and posterior-approximation bounds, and compare marginalization with MAP under the same anchors and likelihood.
4. **Mechanism and application evidence.** Exact interventions audit published radial readouts; matched recovery studies establish sparse semantic calibration and ordinal gains, alongside stable tasks and destination-conditioned routing results.

Positioning relative to prior work. Hyperbolic embeddings and graph networks use negative curvature to represent hierarchical metrics (Nickel & Kiela, 2017; Chami et al., 2019). HIE uses origin-relative norms as hierarchy information and recenters representations around a data-dependent center (Yang et al., 2023); HypStructure adds tree and centering losses to visual embeddings (Sinha

et al., 2024). Entailment and relation-specific hyperbolic cones encode directed order directly (Ganea et al., 2018; Bai et al., 2021). Our audits complement these methods: equivariant centering can be stable under coordinate changes, while directed supervision can break symmetry, but neither fact alone establishes that an externally specified frame is identified for a given decision. Work on mixed curvature, geometry selection, and robust hyperbolic learning chooses or protects the metric family (Gu et al., 2019; Fu et al., 2021; Lee et al., 2024b;a; Malik et al., 2025); we study ambiguity within a fixed geometry.

The broader identifiability literature shows that latent factors may be unrecoverable without additional assumptions or observations (Locatello et al., 2019; Khemakhem et al., 2020); class–pose decomposition uses symmetry supervision to separate invariant and equivariant factors (Marchetti et al., 2023). We specialize this perspective to semantic decisions attached to geometric representations. Our bounds and anchor conditions are indexed by decision equivalence, so they require recovery only of distinctions relevant to the requested task, not the full latent coordinates.

2 Semantic-Frame Identifiability and Evidence Requirements

Let $\vartheta \in \Theta$ parameterize a representation and observation law P_ϑ, let a group $\mathcal{G}$ act on Θ, and let $s(\vartheta)$ be a semantic target in a pseudometric space $(\mathcal{S}, d_s)$. Observational invariance means $P_{g\vartheta} = P_\vartheta$ for every $g \in \mathcal{G}$. For a query family $\mathcal{Q}$, decision disagreement is the pseudometric $d_{\mathcal{Q}}(s,t) = \Pr_{q\sim\nu_{\mathcal{Q}}}[a_s(q) \neq a_t(q)]$, with a fixed query distribution $\nu_{\mathcal{Q}}$. Define $\Delta_s(\vartheta) = \sup_{g,h\in\mathcal{G}} d_s(s(g\vartheta), s(h\vartheta))$.

Theorem 1 (Orbit lower bound)**.** *Under observational invariance, every possibly randomized estimator satisfies*

$$\inf_{\widehat{s}} \sup_{g\in\mathcal{G}} \mathbb{E}_{X\sim P_\vartheta} d_s(\widehat{s}(X), s(g\vartheta)) \geq \tfrac{1}{2}\Delta_s(\vartheta). \tag{1}$$

The same bound holds for $d_{\mathcal{Q}}$.

Proof sketch. Observationally equivalent states induce the same estimator output law. The triangle inequality bounds their target separation by the sum of their two risks. Taking the orbit supremum yields the result; full proofs are in Appendix B.

Thus a semantic functional is identifiable only if it is constant on each class of states having the same observation law; a group orbit can be a subset of this class. The two-point comparison in Figure 1 illustrates the resulting lower bound. For $Z = (z_i)$ in $\mathbb{H}^d_c$, any distance objective $\mathcal{L}(Z) = F(D(Z))$ and distance decoder satisfy $D(TZ) = D(Z)$ under a common isometry T. The fixed-origin readout instead obeys $d_c(o, Tz) = d_c(T^{-1}o, z)$ and can change along the orbit. An equivariant Fréchet center moves with T, but identifying it with an externally named semantic root requires additional evidence. The observation-law argument also applies to Euclidean axes and spherical poles (Appendix A).

Finite-sample recovery from noisy anchors. For a finite root family $\mathcal{R}$ and a fixed anchor design $A = (a_1, \ldots, a_m)$, let $q_A(r) = (d(r, a_1), \ldots, d(r, a_m))$ and observe $y = q_A(r_\star) + \xi$, with independent mean-zero σ-sub-Gaussian coordinates. Least squares has error bound

$$\Pr(\widehat{r} \neq r_\star) \leq \sum_{s\neq r_\star} \exp[-\Delta_A^2(r_\star, s)/(8\sigma^2)], \qquad \Delta_A(r,s) = \|q_A(r) - q_A(s)\|_2. \tag{2}$$

The bound follows from a sub-Gaussian tail and a union bound over competing signatures (Appendix A). Decision error sums only over candidates disagreeing with $r_\star$ on $\mathcal{Q}$. For nearest-signature decoding under bounded coordinate noise, $\|\xi\|_\infty \leq \eta$, a minimum L_∞ signature gap greater than 2η guarantees exact root recovery. Decision recovery needs this separation only between candidate frames that disagree on the requested decisions.

For Gaussian noise, $D_{\mathrm{KL}}(P_r \| P_s) = \Delta_A^2(r,s)/(2\sigma^2)$. If $M \geq 2$ decision-distinct candidates have pairwise squared gaps at most B^2, Fano's inequality gives

$$\Pr(\text{decision error}) \geq 1 - \frac{B^2/(2\sigma^2) + \log 2}{\log M} \tag{3}$$

under the uniform prior on these candidates. With one canonical hypothesis per decision class, uniform squared separation $m\gamma^2$, and a regular packing with comparable upper separation, the upper

bound and Fano/two-point lower bounds yield $m = \Theta(\sigma^2\gamma^{-2}[\log N_{\mathcal{Q}} + \log(1/\delta)])$ for error $\delta < 1/4$. Here $N_{\mathcal{Q}}$ counts decision classes. The rate holds under the canonical-family and packing assumptions; the root union bound otherwise retains within-class nuisance variation (Corollary A.2 in Appendix A).

Decision-resolving evidence for finite DAG frames. Let $\mathfrak{F}$ be admissible frames sharing an observed undirected skeleton and $U_{\mathcal{Q}} = \{\{f, f'\} : f \not\sim_{\mathcal{Q}} f'\}$ the pairs disagreeing on requested decisions. Anchor a separates $S_a = \{\{f, f'\} \in U_{\mathcal{Q}} : q_a(f) \neq q_a(f')\}$. A noiseless anchor set identifies all decisions iff its S_a cover $U_{\mathcal{Q}}$: an uncovered pair has identical observations and incompatible answers; a complete cover maps each observed signature to one decision class. Classical greedy set cover therefore uses at most $1 + \log|U_{\mathcal{Q}}|$ times the optimal anchor count. If each anchor has at most B outcomes, an adaptive query tree of depth m has at most B^m leaves, so $m \geq \lceil \log_B N_{\mathcal{Q}} \rceil$ is necessary. For a finite self-certifying noise model, prior-weighted cross-class edge cutting reduces to EC^2 and inherits its logarithmic expected-cost guarantee (Golovin et al., 2010). This reduction connects decision-resolving evidence to established acquisition guarantees. We evaluate continuous and entropy-based acquisition rules empirically.

3 FramePosterior: Evidence-Conditioned Semantic Decisions

The theory suggests separating what geometry already determines from what additional evidence must resolve. FramePosterior takes a frozen representation, a specified semantic query and loss, and asymmetric anchors $A = \{(a_j, y_j)\}$, such as noisy depths at named nodes. It changes the semantic readout, not the encoder or native decoder. Its three steps are to score candidate frames against the anchors, aggregate their task-specific answers, and check whether the remaining disagreement permits action.

Formally, attach the evidence to the intrinsic quotient $U = [Z]_{\mathcal{G}}$ and infer $p(F \mid U, A)$. Implementations represent this conceptual quotient through declared invariants and existing frozen outputs. For query q, define $\mathcal{R}_p(a) = \mathbb{E}_{F\sim p} L(a; F, q)$ and take

$$a_p(q) \in \arg\min_a \mathcal{R}_p(a), \tag{4}$$

$$\mathcal{R}_p(a_{\widetilde{p}}) - \mathcal{R}_p(a_p) \leq 2L_{\max}\,\mathrm{TV}(p, \widetilde{p}), \tag{5}$$

for an approximate posterior $\widetilde{p}$ and $0 \leq L \leq L_{\max}$. The bound follows by comparing each action's risk under the two posteriors (Appendix A). Native outputs are unchanged because their input and predictor are frozen. The Bayes guarantee concerns the declared frame model and loss. For exact posterior truncation to K, $\mathrm{TV}(p, p_K) = 1 - p(K)$, so omitted mass directly controls the approximation term. The WordNet study instantiates this separation between geometry and semantics: an exact Möbius map reverses fixed-origin radial order while preserving pairwise distances, and noisy anchors update semantic decisions without changing the native predictor (Figure 2).

Step 1: evidence-conditioned frame inference. A root candidate defines $h_r(v) = d(r, v)$, using the declared graph metric or frozen embedding metric. With uniform prior, our graph experiments use the robust Gibbs posterior

$$w_r = \frac{\exp[-\sum_j \rho_H(h_r(a_j) - y_j)/T]}{\sum_s \exp[-\sum_j \rho_H(h_s(a_j) - y_j)/T]}, \qquad \rho_H(t) = \begin{cases} t^2/2, & |t| \leq 1, \\ |t| - 1/2, & |t| > 1. \end{cases} \tag{6}$$

The analytic Gaussian case replaces $\rho_H(t)/T$ by $t^2/(2\sigma^2)$. Summing independent anchor losses permits concentration. Experiments use $T = 1$ and top-64 charts on WordNet, top-128 on the Disease acquisition task; retained weights are renormalized.

Step 2: task-loss-aware decision marginalization. For squared depth loss, the Bayes field is $\sum_r w_r h_r(v)$; absolute loss instead uses a posterior median. Binary ordinal decisions use

$$v_p(u, v) = \sum_r w_r \,\mathrm{sign}[h_r(u) - h_r(v)], \qquad \widehat{a}(u, v) = \mathrm{sign}\, v_p(u, v). \tag{7}$$

Chart ties contribute zero; $v_p = 0$ is unresolved, and a margin threshold on $|v_p|$ provides selective predictions. MAP uses one maximizing root with the same observations and losses. The separate

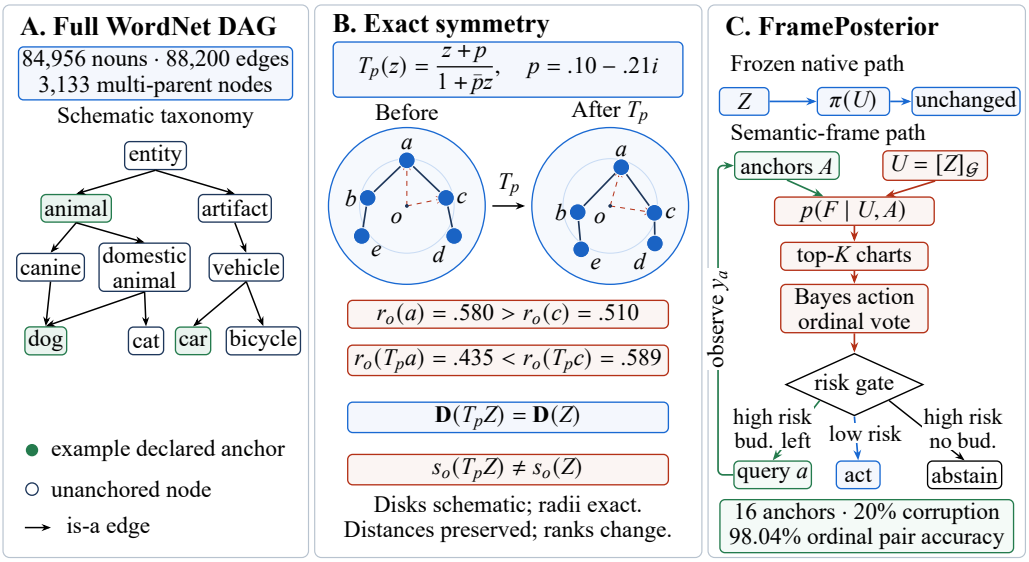

Figure 2: WordNet illustration of FramePosterior. (A) Benchmark scale and a schematic taxonomy excerpt. (B) The exact Möbius map preserves pairwise distances while reversing the fixed-origin radial order; disk coordinates are schematic and the displayed radii are computed from the map. (C) Anchors update semantic-frame decisions while the native predictor remains frozen. With 16 anchors and 20% corruption, ordinal pair accuracy is 98.04%.

minimum-loss atlas uniformly averages tied loss minimizers, rather than the full Gibbs posterior. Multiple operational regions receive separate charts. The benefit is decision-specific: uncertain root identity need not prevent a stable depth comparison when the supported charts agree on its answer.

Step 3: uncertainty-aware action and abstention. For graph-distance charts, let $R \sim p(\cdot \mid U, A)$, $p(C_\alpha \mid U, A) \geq 1 - \alpha$, $\widehat{h}_{C_\alpha} = \mathbb{E}[h_R \mid R \in C_\alpha, U, A]$, and $\mathrm{diam}_h(C) = \sup_{r,s \in C} \|h_r - h_s\|_\infty$. With graph diameter D_G,

$$\mathbb{E}[\|h_R - \widehat{h}_{C_\alpha}\|_\infty \mid U, A] \leq \mathrm{diam}_h(C_\alpha) + \alpha D_G. \tag{8}$$

The bound splits the expectation inside and outside the credible set (Appendix A). If realized sup-norm error is at most ϵ, a true pair margin above 2ϵ keeps its sign. The bound measures risk under the specified frame and observation model. Its external reliability depends on that model's adequacy. For binary chart decisions, low posterior disagreement supports action; otherwise the system queries within budget or abstains.

Evidence acquisition and task-specific implementations. The formal decision-cover policy cuts unresolved cross-class edges under the finite model above. Our Disease implementation instead chooses the largest predictive anchor-outcome entropy over a fixed candidate pool; we call it active querying (AQ) and evaluate it as a heuristic. ImageNet uses radial coverage (RDC) of stored source calibration radii and resampled monotone charts. This calibration ensemble approximates decision-level recovery through monotone readouts. Each implementation specifies its frame family, evidence, inference rule, and decision loss. At inference: acquire the budgeted evidence, update frame weights, materialize retained charts, compute the query action and its disagreement, then act, query, or abstain.

4 Experimental Evaluation of Identifiability and Recovery

The evaluation follows the theoretical distinction between metric prediction and semantic decision. **First, can semantic readouts change while native predictions remain fixed?** Exact-isometry audits isolate this mechanism on HIE, HypStructure, UNCHA, and CIFAR-100. **Second, does**

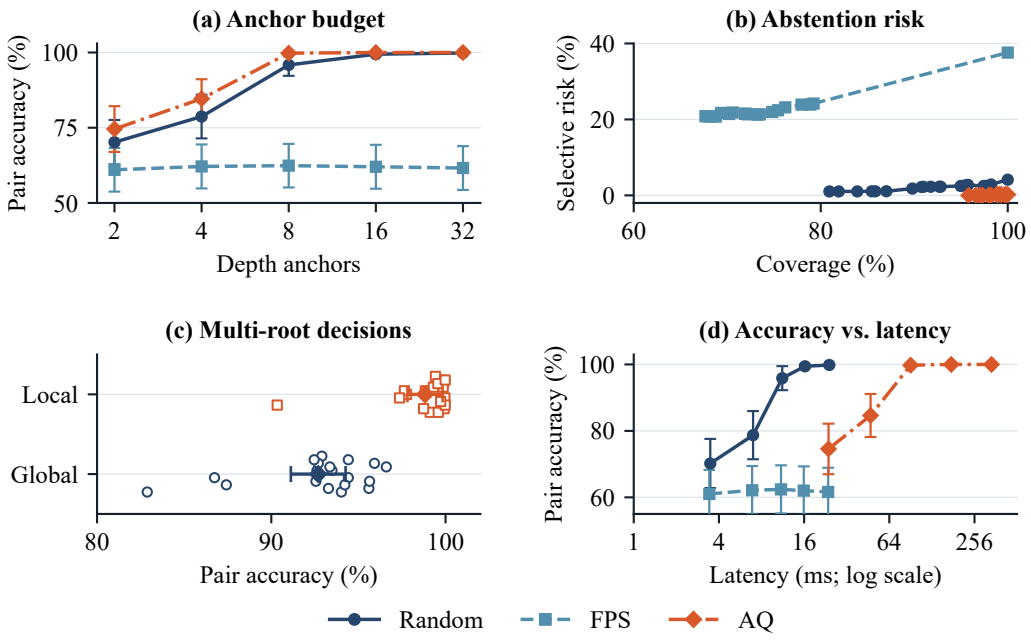

Figure 3: Disease diagnostics on the HGCN benchmark (Chami et al., 2019) at 20% anchor corruption. (a,b,d) show accuracy versus budget, selective risk versus coverage, and accuracy versus acquisition-plus-inference time over 40 root trials; zero-coverage endpoints are omitted in (b). (c) compares local and global roots over 20 trials (6.13-pp mean gain). Whiskers show 95% t intervals; FPS is farthest-point sampling and AQ is predictive-entropy active querying.

sensitivity cause task loss? Radial-depth accuracy and uncertainty-scaled NLL measure consequences beyond selection-set turnover; stable readouts provide negative controls. **Third, can limited evidence restore useful decisions?** Disease and WordNet compare matched anchor budgets and decision rules, and ImageNet tests sparse semantic calibration. Destination-conditioned routing separately tests changing roots, utility, and state overhead with additional declared topology information. Each intervention is applied to a trained, frozen model; repeated transformations are paired within a seed and are not counted as independent model runs. The primary GPU server ran Ubuntu 22.04.4 with an Intel Xeon Silver 4210 (40 logical CPUs), 94 GiB RAM, and two NVIDIA RTX 4090 GPUs. Appendix E lists the software versions, information budgets, seeds, hashes, confidence intervals, and protocols.

Task-specific baselines and evidence budgets are summarized in Table 6 in Appendix E; the main text reports the decisive matched comparisons.

Intervals for trained models cluster by seed; ImageNet transformations and noise draws are paired within seed. UNCHA intervals span ten fixed directions on one checkpoint, not checkpoint variation. Routing pairs the same 100 destinations after averaging five seeds; WordNet redraws corruption trials. Each task is evaluated against its prespecified decision thresholds.

4.1 Prediction Preservation, Semantic Sensitivity, and Task Loss

HIE: geometric centering versus semantic orientation. We implement HIE's published centering and hierarchy regularization on the official HGCN pipeline (Yang et al., 2023). Across three HGCN and three HGCN–HIE runs, 100 common isometries per run preserve native decoder outputs to at most 7.2×10^{-13}, while fixed-origin depth decisions flip by as much as 32.2%. The HIE center transforms equivariantly, so distances to the transported center remain invariant. Re-rooting relative to an independently declared semantic root gives 48.6%–48.9% mean pair accuracy. Equivariant

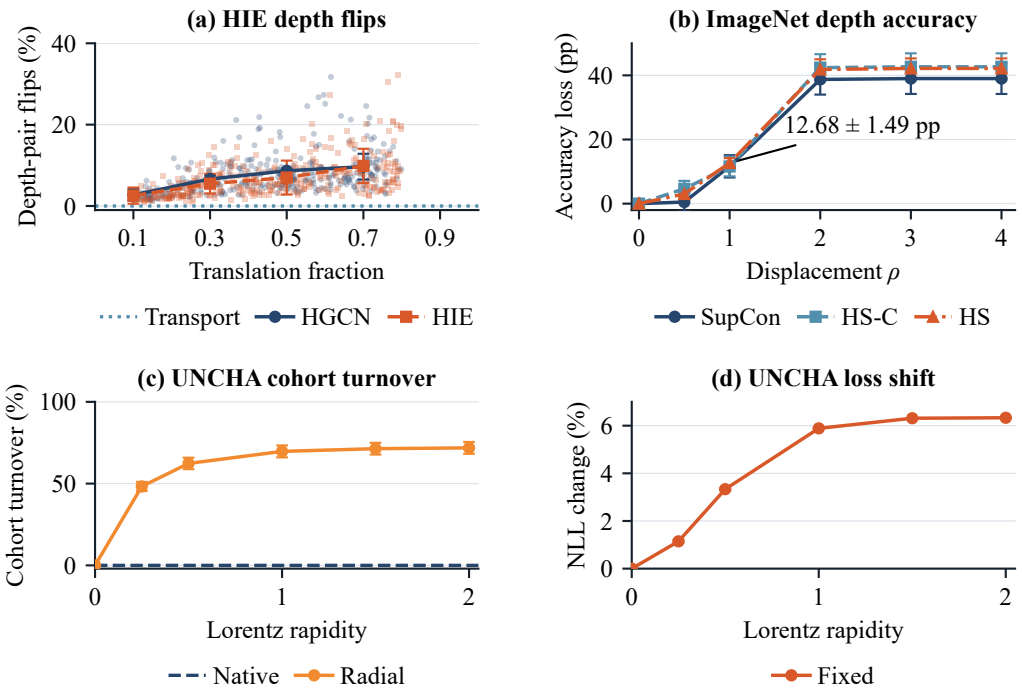

Figure 4: Exact intervention audits. (a) HGCN/HIE (Chami et al., 2019; Yang et al., 2023) fixed-origin flips and invariant transported readout (95% t intervals over three seeds). (b) ImageNet-100 HypStructure (Sinha et al., 2024) depth-accuracy loss at $\rho = 1$ (12.68 ± 1.49 pp; ten full-model seeds and five per control). (c,d) UNCHA (Kim et al., 2026) uncertainty-set turnover and scaled-NLL change despite unchanged class predictions; intervals span ten directions on one checkpoint.

canonicalization establishes coordinate stability, while task-specific orientation requires asymmetric evidence.

We also run the public HIE configuration for five seeds with its training loop and optimizer unchanged, adding a separate output hook to save the validation-selected embedding for external readout. Its fixed-origin radial set changes by 6.71%–7.21% on average, below the 10% mean-effect threshold. This audit reproduces radial sensitivity in public code, with an effect below the prespecified practical threshold.

ImageNet-100: radial-depth loss and sparse recovery. We train the public HypStructure ImageNet-100 configuration for 20 epochs and freeze each encoder (Sinha et al., 2024; Deng et al., 2009). Full HypStructure uses ten independent seeds; Flat SupCon and a no-center control use five each. A monotone readout is fitted on 107 taxonomy calibration nodes and evaluated on 60 held-out nodes from the 167-node decision quotient. It maps prototype radius to directed WordNet depth and is frozen before intervention. At unit boost rapidity ($\rho = 1$), the transported-frame oracle and native classifier are invariant, while within-one-level depth accuracy drops 12.68 ± 1.49 percentage points and depth MAE rises 0.614 levels (95% t intervals over ten training seeds; Figures 4b and 5a). Thirty-two active landmarks with Gaussian noise at 5% of the source-radius standard deviation recover 89.88% of the MAE gap under seed-wise averaging (Figure 5b); eight recover 80.69%, versus 58.81% for random landmarks. At 32 landmarks, ensemble MAE is 1.340 versus 1.343 for the single isotonic chart, locating the visual gain in sparse calibration. Rooted ancestor retrieval does not deteriorate, so the measured effect is specific to the radial- depth decision.

UNCHA: radial uncertainty and downstream loss. UNCHA defines hyperbolic uncertainty by $u(z) = \mathrm{softplus}(-\|z_{\mathrm{space}}\|_2)$ and uses $\exp(-u/2)$ to scale contrastive logits (Kim et al., 2026). We load the released ViT-S/16 checkpoint, encode all 5,000 ImageNet-100 validation images with

Table 1: Matched decision evidence. WordNet compares MAP and posterior voting at 16 anchors over 50 paired corruption trials; gains show paired 95% t intervals. Disease averages 100 roots. Root accuracy is the shared MAP diagnostic; native AUC is unchanged.

Setting	Anchors/noise	Rule	Root	Pair	Gain [95% CI], pp
Disease	8 / 20%	Min.-loss atlas	3.0%	99.33%	–
Disease	8 / 40%	Min.-loss atlas	2.0%	99.16%	–
WordNet	16 / 20%	MAP	100%	96.56%	–
WordNet	16 / 20%	Posterior vote	100%	98.04%	**+1.48 [1.45, 1.51]**
WordNet	16 / 40%	MAP	98%	96.24%	–
WordNet	16 / 40%	Posterior vote	98%	97.74%	**+1.50 [1.07, 1.93]**

deterministic 70% central crops and class prompts, and apply the same exact Lorentz boost to image and text embeddings. Ten pre-specified directions at unit rapidity yield maximum pair-distance error 4.43×10^{-14} and zero class-decision flips, yet 69.70% (95% interval $[61.54, 77.81]$) of the top uncertainty quintile changes and held-out uncertainty-scaled NLL increases 5.89% $[5.86, 5.93]$. Inverse frame transport restores uncertainty and NLL to numerical precision. This checkpoint-level audit tests one published radial mechanism rather than the full UNCHA benchmark.

CIFAR-100: sensitivity without substantial OOD loss. On CIFAR-100 (Krizhevsky, 2009), three matched ResNet-34 conditions, each trained with three seeds, reach 76.47–76.87% test fine-class prototype accuracy. Across 100 exact isometries, native prototype predictions never flip, while the fixed-origin 80%-coverage selection set changes by 18.89–21.99%; a transported equivariant center has zero measured selection-set flips. The worst OOD AUROC loss is only 0.56–0.74 percentage points, below the five-point gate. This negative control separates selection-set sensitivity from substantial OOD loss. The full protocol and additional results are in Appendix D.7.

4.2 Semantic Recovery from Sparse and Noisy Evidence

Disease-LP contains a 2,665-node tree from the HGCN benchmark (Chami et al., 2019). In the 40-root acquisition study, each anchor depth is independently perturbed by ± 1 hop with probability 20% and clipped at zero. AQ with eight anchors obtains 99.78% pair accuracy versus 95.87% for random acquisition under the same budget. A separate 100-root minimum-loss-atlas study tests exact-root versus decision recovery (Table 1). On the full Open English WordNet noun DAG (Fellbaum, 1998) (84,956 nodes, 88,200 edges, 3,133 multi-parent nodes), posterior ordinal voting reaches 98.04% pair accuracy with 16 anchors at 20% corruption; with 40% corruption it obtains 97.74% at 16 anchors and 98.01% at 32. The corruption probability has the same ± 1-hop meaning as on Disease. Candidate charts and signatures use the undirected skeleton, while the evaluation target is directed shortest-path depth. Even the correct root can therefore produce incorrect chart orders through undirected shortcuts; posterior voting improves this structural approximation, rather than root identity. Queries are fixed unequal-depth pairs; a predicted tie counts as an error. At 20% corruption, MAP and posterior predict ties on 2.36% and 0.007% of queries, respectively. Landmark signatures use $O(m|V|)$ storage without an all-pairs matrix.

Table 1 shows both decision-level effects: Disease reaches 99.33% pair accuracy with only 3% root recovery, while WordNet posterior voting improves over MAP at the same budget despite identical root accuracy. Native link AUC is unchanged.

On Disease, AQ rises from 74.57% pair accuracy at two anchors to 99.78% at eight, while farthest-point selection stays near 62%. In 20 multi-root tasks, local roots reach $98.82 \pm 2.16\%$ vs. $92.69 \pm 3.36\%$ for one global root (16 anchors plus region ID). WordNet posterior trials take under 0.45 seconds through 64 charts and 1.57 seconds at 256. Directional entailment cones reach .748 edge AUC but near-chance global radial depth, separating local order, root identity, and scalar depth.

4.3 Destination-Conditioned Routing: Utility and Resource Costs

Routing is an identity-sensitive stress test because the semantic root changes with each destination. We freeze official Airport HGCN embeddings from five seeds on the 3,158-node largest compo-

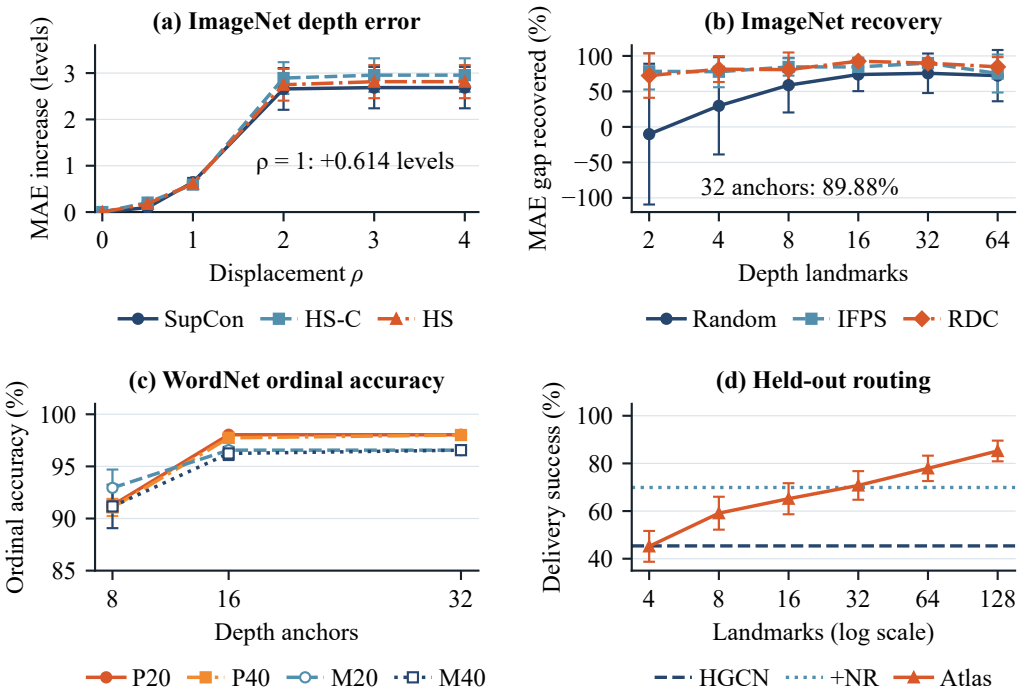

Figure 5: Sparse recovery and downstream decisions. (a,b) ImageNet-100 HypStructure (Sinha et al., 2024) depth MAE increases 0.614 levels at $\rho = 1$; 32 noisy landmarks recover 89.88% of the gap. (c) WordNet noun DAG (Fellbaum, 1998) compares posterior (P) and MAP (M) at 20/40% corruption. (d) Routing compares atlas sizes with the no-revisit control (+NR; HGCN backbone (Chami et al., 2019)). Whiskers are 95% intervals over ImageNet seeds, 50 WordNet trials, or 100 destinations averaged over five frozen seeds.

nent, hold out the same 100 destinations for all seeds, and evaluate 50 sources per destination. Re-centering at r_t gives $h_{r_t}(v) = d(z_{r_t}, z_v)$, so re-centering alone leaves native greedy forwarding unchanged. The atlas additionally stores graph shortest-path signatures and scores neighbors by $b_m(v, r_t) = \max_{j \le m} |d_G(v, a_j) - d_G(r_t, a_j)|$, with a $10^{-3} h_{r_t}(v)$ secondary term and no revisit. The triangle inequality gives $b_m(v, r_t) \le d_G(v, r_t)$. Thus the delivery gain uses declared topology and destination evidence; no target next-hop labels are supplied.

Atlas-64 delivers $77.92 \pm 5.33\%$ vs. 69.90% for HGCN+no-revisit and 45.38% for native HGCN, with unchanged AUC. Its 8.02-pp gain over +no-revisit has a paired 95% interval $[3.26, 12.79]$ over 100 destinations. It stores 0.385 MiB, stretches successful paths 1.699, and leaves 22.08% undelivered. State scales as $2m|V|$ bytes; neighbor scoring takes 6.6–7.4 μs. Atlas-128 is post-hoc; anonymous destinations remain identity-limited.

5 Conclusion

Metric prediction and semantic orientation require distinct information. FramePosterior uses declared evidence to resolve task-relevant frame ambiguity while preserving native predictions. The recovery studies establish sparse calibration and ordinal gains; stable visual readouts identify cases in which ambiguity has little practical effect. The guarantees apply to the specified frame and observation models, and the routing results concern topology-level forwarding.

AI USE STATEMENT

Generative AI tools (OpenAI Codex) were used for manuscript editing, code development and revision, and research assistance. The authors reviewed and verified AI-assisted content and take full responsibility for the research claims, accuracy, attribution, and all content of this submission.

REPRODUCIBILITY STATEMENT

Appendices A and B state assumptions and full proofs. Appendices C, D, and E specify implementations, data processing, evidence budgets, seeds, and statistical units; Appendices F and G report additional diagnostics. Exact interventions are evaluated on frozen models, and repeated transformations are not treated as independent training runs. Numerical replication is assessed separately from scientific-conclusion replication.

REFERENCES

Yushi Bai, Rex Ying, Hongyu Ren, and Jure Leskovec. Modeling heterogeneous hierarchies with relation-specific hyperbolic cones. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12316–12327, 2021.

Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

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.

Christiane Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Xingcheng Fu, Jianxin Li, Jia Wu, Qingyun Sun, Cheng Ji, Senzhang Wang, Jiajun Tan, Hao Peng, and Philip S. Yu. ACE-HGNN: Adaptive curvature exploration hyperbolic graph neural network. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 2021.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1646–1655. PMLR, 2018.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, volume 23, pp. 766–774, 2010.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019.

Tejaswi Kasarla, Max van Spengler, and Pascal Mettes. Balanced hyperbolic embeddings are natural out-of-distribution detectors. *International Journal of Computer Vision*, 134(202), 2026. doi: 10.1007/s11263-026-02775-6.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Hayeon Kim, Ji Ha Jang, Junghun James Kim, and Se Young Chun. Uncertainty-guided compositional alignment with part-to-whole semantic representativeness in hyperbolic vision-language

models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36861–36870, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kim_Uncertainty-guided_Compositional_Alignment_with_Part-to-Whole_Semantic_Representativeness_in_Hyperbolic_Vision-Language_CVPR_2026_paper.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Eric W. Lee, Bo Xiong, Carl Yang, and Joyce C. Ho. Hypmix: Hyperbolic representation learning for graphs with mixed hierarchical and non-hierarchical structures. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024a.

See Hian Lee, Feng Ji, and Wee Peng Tay. Node-specific space selection via localized geometric hyperbolicity in graph neural networks. *Transactions on Machine Learning Research*, 2024b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schoelkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Nikita Malik, Rahul Gupta, and Sandeep Kumar. Hyperdefender: A robust framework for hyperbolic gnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025. doi: 10.1609/aaai.v39i18.34135.

Giovanni Luca Marchetti, Gustaf Tegnér, Anastasiia Varava, and Danica Kragic. Equivariant representation learning via class-pose decomposition. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4745–4756. PMLR, 2023. URL https://proceedings.mlr.press/v206/marchetti23b.html.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Menglin Yang, Min Zhou, Rex Ying, Yankai Chen, and Irwin King. Hyperbolic representation learning: Revisiting and advancing. In *International Conference on Machine Learning*, pp. 39639–39659, 2023.

A Identifiability, Information, and Active Symmetry Breaking

A.1 Non-Identifiability under Observational Symmetry

For $\vartheta \in \Theta$, define its semantic orbit and diameter

$$\mathsf{Orb}_s(\vartheta) = \{s(g\vartheta) : g \in \mathcal{G}\}, \qquad \Delta_s(\vartheta) = \sup_{g,h\in\mathcal{G}} d_s(s(g\vartheta), s(h\vartheta)). \tag{A.1}$$

Theorem A.1 (Semantic-frame orbit bound)**.** *Suppose $P_{g\vartheta} = P_\vartheta$ for every $g \in \mathcal{G}$, and d_s is a pseudometric. Then for every (possibly randomized) estimator $\widehat{s}(X)$,*

$$\inf_{\widehat{s}} \sup_{g\in\mathcal{G}} \mathbb{E}_{X\sim P_\vartheta} d_s(\widehat{s}(X), s(g\vartheta)) \geq \frac{1}{2}\Delta_s(\vartheta). \tag{A.2}$$

The same statement holds with d_s replaced by the decision pseudometric $d_{\mathcal{Q}}$. Consequently, statistical identifiability of a semantic functional requires it to be constant on every operational-equivalence class $[\vartheta]_{\mathcal{O}} = \{\vartheta' : P_{\vartheta'} = P_\vartheta\}$. Constancy is also sufficient for the functional to be well-defined on the quotient statistical model. A group orbit can be a strict subset of this class, so invariance under $\mathcal{G}$ supplies a readily checkable necessary condition, not an assertion that $\mathcal{G}$ exhausts all non-identifiabilities.

The bound covers any semantic frame that varies along an observational symmetry orbit, including Euclidean axes, spherical poles, and graph-role orientations. Appendix B provides the complete two-point proof and the remaining derivations.

For example, an orthogonal transformation of a Euclidean embedding leaves all pair distances unchanged while rotating a named concept axis; a rotation of a spherical embedding leaves all pairwise angles unchanged while moving a declared pole. Theorem A.1 applies to both cases. An architecture can choose a canonical frame; identifying that convention with the application's semantic frame requires evidence that distinguishes the corresponding states.

Corollary A.1 (Hyperbolic fixed-origin semantics)**.** *If $\mathcal{L}(Z) = F(D(Z))$, every $T \in \mathrm{Isom}(\mathbb{H}^d_c)$ satisfies $\mathcal{L}(TZ) = \mathcal{L}(Z)$. Any statistic not constant on this orbit—including radius from a fixed coordinate origin—is not identifiable. A Möbius translation can change its hierarchy ranking while leaving every distance-decoder score exactly unchanged.*

An isometry-equivariant Fréchet center moves with the representation, making center-relative distance intrinsic. Its correspondence with an externally declared semantic root is a separate identification problem, evaluated in the HIE audit.

A.2 Root and Decision Recovery under Noisy Anchors

For anchors $A = (a_1, \ldots, a_m)$ define the metric signature $q_A(r) = (d(r, a_1), \ldots, d(r, a_m))$.

Definition A.1 (Root and decision equivalence)**.** *Roots are observationally equivalent, $r \sim_A s$, when they induce the same anchor-observation law; without noise this is $q_A(r) = q_A(s)$. For query family $\mathcal{Q}$, they are decision equivalent, $r \sim_{\mathcal{Q}} s$, when all decisions induced by $h_r(v) = d(r, v)$ and h_s agree on $\mathcal{Q}$.*

A large root-equivalence class can have small decision diameter, whereas endpoint delivery distinguishes every destination.

Proposition A.1 (Metric-dimension condition)**.** *Under noiseless depths, a root in $\mathcal{R}$ is identifiable iff q_A is injective on $\mathcal{R}$; a universally identifying anchor set is a resolving set. Under bounded noise $\|\xi\|_\infty \leq \eta$, nearest-signature recovery is exact if $\min_{r\neq s} \|q_A(r) - q_A(s)\|_\infty > 2\eta$.*

Let $y = q_A(r_\star) + \xi$, with independent mean-zero σ-sub-Gaussian coordinates, and $\Delta_A(r, s) = \|q_A(r) - q_A(s)\|_2$.

Theorem A.2 (Finite-sample noisy-anchor recovery)**.** *The least-squares estimator satisfies*

$$\Pr(\widehat{r} \neq r_\star) \leq \sum_{s\neq r_\star} \exp\{-\Delta_A^2(r_\star, s)/(8\sigma^2)\}. \tag{A.3}$$

Table 2: Identifiability across geometric representations. The listed symmetry-breaking evidence is additional to the invariant observable.

Space	Invariant observable	Non-intrinsic frame	Symmetry breaker
Hyperbolic	pair distances / symmetric links	root, radial depth	known depth / order
Euclidean	distance or Gram orbit	named concept axis	signed concept labels
Spherical	pairwise angles	pole / north–south order	pole landmark
Graph	undirected skeleton	role or flow direction	time / causal / flow edge
Vision	class similarity metric	abstract–specific direction	class hierarchy edge

If every incorrect root has average squared anchor gap at least γ^2, then $m \geq 8\sigma^2\gamma^{-2}[\log(|\mathcal{R}| - 1) + \log(1/\delta)]$ suffices for error at most δ. Decision error obeys the same bound after summing only over $s \not\sim_{\mathcal{Q}} r_\star$.

The matching obstruction depends on signature separation, not embedding dimension.

Theorem A.3 (Unresolved-root lower bound)**.** *Under Gaussian noise, suppose $S \subseteq \mathcal{R}$ contains $M \geq 2$ roots with $\max_{r,s\in S} \Delta_A^2(r,s) \leq B^2$. For a uniformly random root in S, every estimator obeys*

$$\Pr(\widehat{R} \neq R) \geq 1 - \frac{B^2/(2\sigma^2) + \log 2}{\log M}. \tag{A.4}$$

The same lower bound applies to a declared decision if roots in S are decision-inequivalent.

This follows from the Gaussian KL divergence and Fano's inequality (Cover & Thomas, 2006); Appendix B gives complete proofs.

The preceding upper and lower bounds become rate-matching after separating root identity from the number of decisions that must be distinguished. Let $N_{\mathcal{Q}} = |\mathcal{R}/\sim_{\mathcal{Q}}|$. Consider a finite canonical family containing one representative of each decision class and Gaussian observations $Y_j = \mu_R(a_j) + \xi_j$. For a randomized anchor design ν, define

$$d_\nu^2(r,s) = \mathbb{E}_{a\sim\nu}[\mu_r(a) - \mu_s(a)]^2. \tag{A.5}$$

Corollary A.2 (Decision sample complexity under a regular design)**.** *Assume the anchor designs A_m obey $\Delta_{A_m}^2(r,s) \geq m\gamma^2$ for all decision-inequivalent pairs, and that the family contains a packing of $M \asymp N_{\mathcal{Q}}$ hypotheses with $\Delta_{A_m}^2(r,s) \leq mC^2\gamma^2$ for a constant C. Assume also that the packing contains a pair separated by at most $mC_0^2\gamma^2$. Then, up to constants depending only on C, C_0, the minimax number of anchors needed for decision error at most $\delta < 1/4$ is*

$$m = \Theta\left(\frac{\sigma^2}{\gamma^2}[\log N_{\mathcal{Q}} + \log(1/\delta)]\right). \tag{A.6}$$

An iid design from ν is covered whenever these empirical inequalities hold with high probability. Without uniform separation and regular packing no universal $\Theta(\log N_{\mathcal{Q}})$ claim is possible: geometry enters through the signature separation and nuisance variation within a decision class must be controlled.

A.3 Decision-Resolving Conditions for Multi-Root DAG Frames

Let $\mathfrak{F}$ be a finite set of admissible DAG frames $f = (R_f, \preceq_f)$ sharing the same observed undirected skeleton. An anchor query $a \in \mathcal{A}$ returns noiseless signature $q_a(f)$ (a depth, root-membership bit, or ordinal relation), and a query family $\mathcal{Q}$ induces frame decision equivalence $\sim_{\mathcal{Q}}$. Define

$$\dim_{\mathcal{Q}}(\mathfrak{F}) = \min_{A\subseteq\mathcal{A}} \{|A| : q_A(f) = q_A(f') \Rightarrow f \sim_{\mathcal{Q}} f'\}. \tag{A.7}$$

Theorem A.4 (Decision-resolving characterization for DAGs)**.** *A noiseless anchor set A identifies every requested DAG decision iff it separates every pair (f, f') with $f \not\sim_{\mathcal{Q}} f'$. Equivalently, let $U_{\mathcal{Q}} = \{\{f, f'\} : f \not\sim_{\mathcal{Q}} f'\}$ and $S_a = \{\{f, f'\} \in U_{\mathcal{Q}} : q_a(f) \neq q_a(f')\}$. Then a decision-identifying anchor set is exactly a set cover of $U_{\mathcal{Q}}$ by $\{S_a : a \in \mathcal{A}\}$. Greedily choosing the anchor that separates the most currently unresolved pairs uses at most $[1 + \log|U_{\mathcal{Q}}|]\dim_{\mathcal{Q}}(\mathfrak{F})$ anchors.*

The theorem provides both a necessary-and-sufficient condition and a constructive approximation guarantee. It also formalizes why a root set alone is generally insufficient for a multi-parent DAG: two candidate partial orders can share the same roots and undirected distances yet disagree on a parent–child query, placing their pair in $U_{\mathcal{Q}}$ until an orientation-sensitive anchor is acquired.

If every anchor takes at most B values, the characterization also yields an information lower bound independent of the set-cover relaxation.

Proposition A.2 (Decision counting lower bound)**.** *Any deterministic adaptive or nonadaptive noiseless strategy that resolves all $N_{\mathcal{Q}} = |\mathfrak{F}/\sim_{\mathcal{Q}}|$ decision classes needs*

$$m \geq \lceil \log_B N_{\mathcal{Q}} \rceil \tag{A.8}$$

queries in the worst case. More generally, if an adaptive strategy's query tree has branching numbers $B_1, \ldots, B_m$, then $\prod_{j=1}^{m} B_j \geq N_{\mathcal{Q}}$ is necessary.

Adaptive noisy acquisition. For noisy queries we augment a frame by its finite noise realization, partition the augmented hypotheses by $\sim_{\mathcal{Q}}$, and connect every pair in different decision classes with weight equal to the product of their prior probabilities. An observed anchor cuts an edge when it rules out at least one endpoint. This is decision-aware: distinguishing two frames in the same decision class earns no utility.

Theorem A.5 (Decision-edge-cut acquisition)**.** *Suppose the finite frame/noise model is self-certifying, acquisition costs are positive, every anchor outcome is deterministic conditional on an augmented hypothesis, and complete acquisition identifies its decision class. Then the total cross-class edge weight cut is strongly adaptive monotone and adaptive submodular. Let $p_{\min} > 0$ be the smallest augmented-hypothesis prior mass. The adaptive policy that maximizes expected cut weight per unit cost has expected identification cost*

$$c(\pi_{\mathrm{FP}}) \leq [2\log(1/p_{\min}) + 1]c(\pi^\star), \tag{A.9}$$

where $\pi^\star$ is the optimal policy for identifying the decision class. Edge weights use the original prior masses on augmented hypotheses; a normalized posterior-disagreement objective or predictive-entropy heuristic is not automatically the same greedy policy.

The result follows by reducing semantic-frame decisions to equivalence-class determination and applying EC^2 (Golovin et al., 2010); Appendix B records the reduction. For a fixed budget before complete identification, generic adaptive submodularity still yields the standard $(1 - 1/e)$ maximization guarantee (Golovin & Krause, 2011). Continuous neural noise models or approximate posteriors need not satisfy these premises; there we report the policy as a heuristic and do not inherit Equation (A.9) automatically.

A.4 Quotient–Frame Decomposition and Prediction Preservation

The orbit bound suggests a representation with two statistically different parts: an intrinsic quotient variable $U = [Z]_{\mathcal{G}}$, and a frame variable F supported on decision-equivalence classes. Let $p(F \mid U, A)$ be any posterior formed from declared asymmetric evidence A, and let $\pi(U)$ be a native predictor that is invariant to $\mathcal{G}$.

Proposition A.3 (Native preservation and semantic Bayes optimality)**.** *Suppose FramePosterior freezes U and π, and changes only the semantic action. Then every native prediction and native empirical risk are exactly unchanged. For a query q and loss $0 \leq L(a; F, q) \leq L_{\max}$, the posterior Bayes action*

$$a_p(q) \in \arg\min_a \mathbb{E}_{F\sim p}[L(a; F, q)] \tag{A.10}$$

minimizes conditional semantic risk. If an approximate posterior $\widetilde{p}$ is used, its action satisfies

$$\mathcal{R}_p(a_{\widetilde{p}}) - \mathcal{R}_p(a_p) \leq 2L_{\max}\,\mathrm{TV}(p, \widetilde{p}). \tag{A.11}$$

Thus approximation error affects the semantic head but cannot silently alter the native task.

This is the constructive counterpart of Theorem A.1: invariant evidence is consumed only by the quotient predictor, whereas non-invariant decisions are exposed through an auditable frame posterior. A single canonical center is the degenerate point-posterior case; multiple rooted charts retain ambiguity when the data do not justify collapsing it.

A.5 Conditional Decision Risk and Application-Specific Limits

Theorem A.6 (Decision risk from a credible set)**.** *Let C_α have posterior mass at least $1-\alpha$, let $\mathrm{diam}_h(C) = \sup_{r,s\in C} \|h_r - h_s\|_\infty$, and let D_G be graph diameter. The posterior mean truncated to C_α satisfies*

$$\mathbb{E}[\|h_R - \widehat{h}_{C_\alpha}\|_\infty \mid A] \le \mathrm{diam}_h(C_\alpha) + \alpha D_G. \tag{A.12}$$

If realized depth error is at most ϵ, every pair with true margin $> 2\epsilon$ is ordered correctly.

Theorem A.7 (Positive semantic error at zero distortion)**.** *For the same unrooted metric under equiprobable semantic roots r, s, every metric-only estimator obeys*

$$\inf_{\widehat{h}} \sup_{R\in\{r,s\}} |V|^{-1}\|\widehat{h} - h_R\|_1 \ge (2|V|)^{-1}\|h_r - h_s\|_1. \tag{A.13}$$

The bound can be positive at zero pairwise-distance distortion; pair-order error is at least half the fraction of queries reversed between h_r and h_s.

Proposition A.4 (Anonymous delivery bound)**.** *If a routing policy observes only (G, A) and succeeds only by terminating at anonymous destination R, its delivery probability cannot exceed the best root-identification probability from the same observations.*

Thus extra directional evidence is optional only when the supported roots are decision-equivalent; it is information-theoretically necessary for identity-sensitive tasks.

B Complete Proofs and Reduction Details

Proof of Theorem A.1. Fix $g, h \in \mathcal{G}$. Because $P_{g\vartheta} = P_{h\vartheta} = P_\vartheta$, the distribution of $\widehat{s}(X)$ is identical under the two states. Pointwise, the triangle inequality gives

$$d_s(s(g\vartheta), s(h\vartheta)) \le d_s(s(g\vartheta), \widehat{s}(X)) + d_s(\widehat{s}(X), s(h\vartheta)).$$

After taking expectation, at least one of the two risks is at least half their semantic separation. Taking the supremum over g, h and then the infimum over estimators yields Equation (A.2); approximation handles a non-attained supremum. The proof uses only the triangle inequality, so it also holds for $d_{\mathcal{Q}}$. Constancy on every operational class is necessary by this argument and sufficient for s to be a well-defined functional of the probability law (the standard definition of statistical identifiability). The group orbit need not equal the full operational class. □

Proof of Corollary A.1. An isometry preserves every entry of $D(Z)$ and hence $F(D(Z))$. In contrast, $d_c(o, Tz) = d_c(T^{-1}o, z)$, which varies with T unless the statistic is constant along the orbit. Apply Theorem A.1. □

Proof of Proposition A.1. Equal signatures give identical noiseless observations. Conversely, an injective signature has an inverse on $\mathcal{R}$. Under bounded noise, observations lie in an ℓ_∞ ball of radius η around the true signature; the stated separation makes these balls disjoint. □

Proof of Theorem A.2. For $u = q_A(s) - q_A(r_\star)$, candidate s beats $r_\star$ only if $2\langle \xi, u\rangle \ge \|u\|_2^2$. The inner product is $\sigma\|u\|_2$-sub-Gaussian, so the event probability is at most $\exp\{-\|u\|_2^2/(8\sigma^2)\}$. A union bound proves Equation (A.3); substituting $\|u\|_2^2 \ge m\gamma^2$ gives the sample complexity. A decision error requires a decision-inequivalent competitor, which gives the restricted sum. □

Proof of Theorem A.3. The Gaussian observation laws obey $D_{\mathrm{KL}}(P_r\|P_s) = \Delta_A^2(r,s)/(2\sigma^2) \le B^2/(2\sigma^2)$. Fano's inequality gives Equation (A.4). If every root in S induces a distinct declared decision, a correct decision identifies the corresponding member of S, reducing decision estimation to root estimation. □

Proof of Corollary A.2. The upper bound follows from Theorem A.2 and the uniform lower separation: the error is at most $(N_{\mathcal{Q}} - 1)\exp[-m\gamma^2/(8\sigma^2)]$. For the lower bound, apply Theorem A.3 to the regular packing, whose pairwise KL divergences are at most $mC^2\gamma^2/(2\sigma^2)$; constant error therefore requires $m = \Omega(\sigma^2\gamma^{-2}\log N_{\mathcal{Q}})$. Applying the Gaussian two-point testing bound to the declared close pair gives error at least a constant multiple of $\exp[-mC_0^2\gamma^2/(2\sigma^2)]$, and hence $m = \Omega(\sigma^2\gamma^{-2}\log(1/\delta))$. Taking the larger of these lower bounds, and noting that max and sum differ by at most a factor two, matches the upper rate. □

Proof of Theorem A.4. If a decision-inequivalent pair has the same joint signature, the two frames produce identical noiseless observations but require different decisions, so identification is impossible. Conversely, if every such pair is separated, each observed signature maps to a single $\sim_{\mathcal{Q}}$ class and all requested decisions are determined. Query a separates exactly the pairs in S_a, so an identifying query set is precisely a cover of $U_{\mathcal{Q}}$. The final statement is the classical greedy set-cover bound applied to this universe. □

Proof of Proposition A.2. A depth-m deterministic adaptive query tree with at most B outcomes per query has at most B^m leaves. Distinct decision classes must terminate at distinct leaves, hence $B^m \geq N_{\mathcal{Q}}$. Replacing the common branching factor by the product of level-wise branching numbers proves the second statement. A nonadaptive signature table is the special case in which all queries are fixed before observing outcomes. □

Proof of Theorem A.5. Create one hypothesis for each supported pair (f, θ) of semantic frame and noise realization. By assumption, its outcome for every anchor is deterministic. Put two hypotheses in the same equivalence class exactly when they induce the same requested semantic decisions, and connect only hypotheses from different classes. An acquisition policy identifies the semantic decision iff its surviving version space lies in one class, which occurs iff it cuts all cross-class edges. This is precisely the equivalence-class determination construction of Golovin et al. (2010). Its edge-cut objective is strongly adaptive monotone and adaptive submodular, and its adaptive greedy expected cost is at most $[2\log(1/p_{\min}) + 1]$ times the optimal expected cost. Mapping the selected test back to its anchor proves the claim. □

Proof of Proposition A.3. The native statement follows because neither its input U nor predictor π is modified. For the semantic statement, a_p is a conditional risk minimizer by definition. Let $\mathcal{R}_p(a) = \mathbb{E}_p L(a; F, q)$. For any fixed action, $|\mathcal{R}_p(a) - \mathcal{R}_{\widetilde{p}}(a)| \leq L_{\max}\,\mathrm{TV}(p, \widetilde{p})$. Insert the optimality inequality $\mathcal{R}_{\widetilde{p}}(a_{\widetilde{p}}) \leq \mathcal{R}_{\widetilde{p}}(a_p)$ and apply this bound twice to obtain Equation (A.11). □

Proof of Theorem A.6. On $R \in C_\alpha$, every chart in the truncated mean is within $\mathrm{diam}_h(C_\alpha)$; outside it, loss is at most D_G. Taking expectation proves Equation (A.12). If each estimated depth differs by at most ϵ, a true pair margin greater than 2ϵ cannot change sign. □

Proof of Theorem A.7. The metric observation and therefore $\widehat{h}$ are identical under both roots. By the triangle inequality, $\|h_r - h_s\|_1 \leq \|h_r - \widehat{h}\|_1 + \|\widehat{h} - h_s\|_1$; at least one term is at least half the left-hand side. Apply the same two-hypothesis argument to each reversed query for the ordinal statement. □

Proof of Proposition A.4. The terminal node of any successful policy is itself a root estimator. A policy with delivery above the optimal identification probability would contradict optimality of that estimator. □

C FramePosterior: Formulation and Implementation Details

Theorem A.1 motivates a separation between intrinsic geometry and semantic orientation. FramePosterior represents these components as an intrinsic orbit and an evidence-conditioned posterior over frames,

$$Z \longmapsto \big([Z]_{\mathcal{G}},\ p(F \mid A, [Z]_{\mathcal{G}})\big), \tag{C.1}$$

where A contains a declared, budgeted source of symmetry-breaking evidence. Native prediction uses only $[Z]_{\mathcal{G}}$; semantic decisions marginalize F. We call the resulting construction *FramePosterior*. Its formal target is the decision quotient $\mathfrak{F}/\sim_{\mathcal{Q}}$: frames need to be distinguished only when their task answers differ. Implementations retain finite candidate frames and aggregate their decisions; the declared information budget specifies which asymmetric observations they may use. The graph implementation below specializes it to roots and multiple rooted charts, while the visual implementation uses superclass/subclass frames. Proposition A.3 guarantees that the frozen native predictor is untouched.

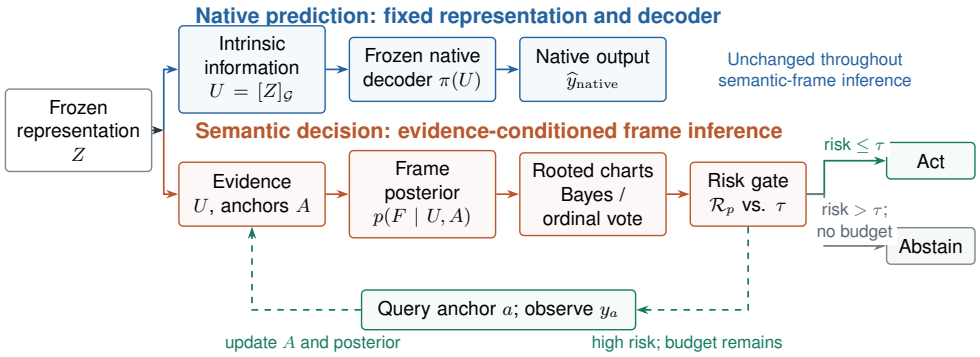

Figure 6: FramePosterior freezes the native predictor and updates an auditable semantic-frame posterior. A budgeted query loop returns a Bayes action when supported, and otherwise queries or abstains.

Posterior equivariance. For transformed inputs and anchors, the posterior must satisfy

$$p(gF \mid gA, [gZ]_{\mathcal{G}}) = p(F \mid A, [Z]_{\mathcal{G}}). \tag{C.2}$$

For discrete candidates these probabilities are masses; for continuous frames, equivariance means equality of pushforward posterior measures, rather than equality of arbitrary densities. When $A = \varnothing$, candidates on the same observational orbit receive the same probability unless the prior explicitly declares asymmetric knowledge. This constraint makes the source of orientation information explicit. For a finite candidate set $\mathfrak{F}$, we use

$$p_\phi(f \mid A, Z) \propto p_0(f) \exp\{-E_A(f)/T - E_\phi(f; \mathrm{Inv}(Z))\}, \tag{C.3}$$

where $\mathrm{Inv}(Z)$ contains pairwise distances or other declared invariants. E_ϕ can share statistical strength across frames, but is constrained to be constant on unresolved frame orbits. Reported graph updates use $E_\phi = 0$. The learned-distance experiments compare intrinsic representations; they do not train this optional amortized frame-energy term.

Residual frame group. For a decision family $\mathcal{Q}$, partition admissible frames by agreement on all requested actions and infer on $\mathfrak{F}/\sim_{\mathcal{Q}}$. This is a quotient set. A representation as group cosets $\mathcal{G}/\mathcal{H}_{\mathcal{Q}}$ additionally requires a compatible group action and an actual subgroup stabilizer; agreement at one fixed embedding alone does not establish those properties. Depth-only queries can ignore some sibling permutations, whereas exact endpoint routing separates destination identities. This decision partition permits reliable actions despite ambiguous roots.

Bayes decisions and abstention. For action loss $L(a; f, q)$, FramePosterior returns

$$\hat{a}(q) = \arg\min_a \mathbb{E}_{f \sim p(f \mid A, Z)} L(a; f, q). \tag{C.4}$$

It abstains when the posterior decision diameter, Bayes regret bound, or credible-set disagreement exceeds a calibrated threshold. This targets decision identification rather than forcing an arbitrary point estimate of a statistically unidentified frame.

Algorithm 1 records each acquired observation against the information budget. With no asymmetric evidence, the prior retains unresolved decision classes. The algorithm describes a credible-set variant. Reported graph runs instead use the stated top-K truncations and posterior voting margins; their selective curves evaluate that implementation. Posterior support and abstention remain visible to downstream users.

C.1 Rooted-Chart Inference and Task-Specific Decision Rules

Our design principle is to leave the intrinsic representation and its native decoder unchanged, and attach semantic direction only through declared asymmetric evidence. Let $A = \{(a_j, y_j)\}_{j=1}^m$ contain noisy depths of landmark nodes. Every candidate root $r \in \mathcal{R}$ defines a rooted chart

$$\psi_r(v) = d(r, v), \tag{C.5}$$

Algorithm 1 FramePosterior with a declared acquisition policy

Require: frozen representation Z, invariant native predictor π, candidate frames $\mathfrak{F}$, prior p_0, budget b, query family $\mathcal{Q}$, thresholds α, ϵ, η

$U \leftarrow \mathrm{Inv}(Z)$; $p \leftarrow p_0$; $A \leftarrow \varnothing$
for $t = 1, \ldots, b$ **do**
  form decision classes and unresolved pair mass under p
  if credible-set decision diameter is below ϵ **then**
    break
  end if
  choose $a_t \leftarrow \arg\min_a \mathcal{U}_t(a)$ using Eq. (C.10)
  observe Y_{a_t}; append (a_t, Y_{a_t}) to A and update $p(F \mid U, A)$
end for
construct only charts in the smallest $(1-\alpha)$ credible decision set
if posterior actions disagree above η **then**
  return $(\pi(U), \text{ABSTAIN})$
end if
return native output $\pi(U)$ and semantic Bayes action from Eq. (C.4)

where d can be a graph shortest-path distance or the intrinsic distance between learned hyperbolic representations. With Huber loss ρ and a robust Gibbs update, the root posterior is

$$p(r \mid A) = \frac{p_0(r) \exp\{-\sum_j \rho(\psi_r(a_j) - y_j)/T\}}{\sum_{s \in \mathcal{R}} p_0(s) \exp\{-\sum_j \rho(\psi_s(a_j) - y_j)/T\}}. \tag{C.6}$$

Summing independent anchor losses allows evidence to concentrate the posterior. The temperature T sets the observation-noise scale.

Decision marginalization. Exact root identity is often unnecessary. For squared metric-depth loss we use the Bayes field

$$\widehat{h}(v) = \mathbb{E}_{r \sim p(r|A)}[\psi_r(v)]. \tag{C.7}$$

For ordinal queries, we instead marginalize the decision, $\Pr[\psi_r(u) < \psi_r(v) \mid A]$, avoiding artifacts caused by averaging tied depths. We retain the smallest posterior credible set C_α with mass at least $1 - \alpha$ and the likelihood-equivalence atlas

$$C_\beta = \{r : \mathcal{L}_{\text{anchor}}(r) \le \min_s \mathcal{L}_{\text{anchor}}(s) + \beta\}. \tag{C.8}$$

The separate Disease minimum-loss-atlas experiment sets $\beta = 0$ and uniformly averages only tied minimum-loss charts; the full posterior is retained for uncertainty. This prevents a long low-probability tail from spuriously breaking same-depth ties. The method abstains when its decision diameter is too large or its posterior ordering margin is too small. This separates unresolved root identity from stable task-specific comparisons.

Multiple charts. A single radius must simultaneously preserve branching adjacency and encode one global depth, which produced the empirical plateau in Section D. Our atlas keeps one chart per supported root and combines decisions only at the output. This does not move embeddings and therefore preserves a frozen distance decoder exactly. In the known-root regime the posterior is a point mass and the atlas reduces to ordinary destination- or taxonomy-rooted distance.

Scalable posterior and direct ordinal vote. An all-pairs candidate distance matrix is unnecessary. We retain m farthest-point landmark signatures and evaluate candidate likelihoods in $O(m(|V| + |E|))$ time and $O(m|V|)$ memory on a graph. Only the K supported candidates are materialized as charts. For a query (u, v) we predict directly from

$$v_p(u, v \mid A) = \sum_{r \in C} p(r \mid A) \operatorname{sign}[\psi_r(u) - \psi_r(v)], \tag{C.9}$$

rather than ranking posterior-mean depths. Retained weights are renormalized; chart ties contribute zero, and a zero signed vote is unresolved. This matches the ordinal rule in Equation (7). This

distinction matters in a multi-parent DAG, where undirected metric shortcuts need not preserve directed taxonomy depth. For learned embeddings, anchor-to-candidate distances can be batched in $O(m|V|d)$; the native representation and decoder remain frozen, and a structural chart is used when the learned metric does not identify direction.

Anchor acquisition. Random anchors are compared with topology-covering landmarks selected greedily by farthest-point sampling. This minimizes a covering-radius surrogate without using the unknown root. The implemented Disease active-querying variant selects the largest predictive outcome entropy over its candidate anchor pool. Both are label-budgeted and use the same observations as single-root baselines.

When anchor outcomes are noiseless and finite, the formal *decision-cover greedy* policy maintains all frame pairs not yet separated and chooses the anchor covering the largest number of unresolved pairs. It is the algorithm induced directly by Theorem A.4, has its logarithmic approximation guarantee, and differs from farthest-point sampling whenever geometric coverage and decision coverage disagree.

For a candidate anchor a, define its expected unresolved decision-pair mass after acquisition,

$$\mathcal{U}_t(a) = \mathbb{E}_{Y_a \sim p_t} \sum_{f \not\sim_{\mathcal{Q}} f'} p_{t+1}(f \mid Y_a) p_{t+1}(f' \mid Y_a). \tag{C.10}$$

This surrogate chooses $a_{t+1} = \arg\min_a \mathcal{U}_t(a)$ for unit-cost anchors. Three acquisition rules have distinct guarantees: noiseless cross-class covering uses Theorem A.4; finite noisy EC^2 maximizes the expected cut of edges weighted by their original augmented-hypothesis prior masses; the reported Disease AQ maximizes predictive outcome entropy. Equation (C.10) is a posterior-concentration heuristic. We assess the heuristics through equal-budget comparisons with random and farthest-point sampling; the EC^2 guarantee applies to the prior-weighted edge-cut rule.

Hierarchical-vision specialization. For a label taxonomy, a frame specifies which prototype is the semantic root and which directed coarse-to-fine relations downstream decisions should respect. Let $z(x)$ be a frozen image representation, p_y the equivariant hyperbolic gyromidpoint prototype for fine class y, and p_c the corresponding prototype of superclass c. Nearest-prototype classification uses only $d(z(x), p_y)$ and is therefore retained as the native task. The semantic action for an edge $c \to y$ is instead the ordinal claim that c precedes y in a declared frame. A fixed-origin implementation tests $d(o, p_c) < d(o, p_y)$; an equivariant geometric-center control replaces o by the center transformed jointly with all prototypes; FramePosterior evaluates the same order after conditioning on a budgeted set of directed superclass/subclass landmarks. Common isometries are applied to images and prototypes together, so any change in native distance or prototype classification is an implementation error, whereas a change in an undeclared fixed-origin order is the predicted intervention. The hierarchy graph permits the same decision-cover acquisition as a taxonomy DAG. This construction makes the visual experiment a test of the theorem rather than an unrelated downstream benchmark.

For a multi-level taxonomy, nodes inducing the same set of observed leaf descendants belong to one decision-equivalence class and are collapsed before evaluation. A frozen radial-depth readout $h(d(o, p_v))$ is calibrated only in the source frame. A semantic landmark supplies its canonical source radius together with the transformed prototype. FramePosterior may then operate directly on the decision quotient: each resampled monotone chart maps observed radii back to the canonical radial coordinate, and the frozen depth decoder is applied before averaging chart depth predictions. The ImageNet implementation uses isotonic calibration with an anchored resampling ensemble, not an exact posterior over ambient isometries. Radial coverage selection uses stored source calibration radii; its budget counts noisy landmark values supplied for recovery. This identifies the depth decisions needed by the task without claiming recovery of an unobservable ambient root coordinate. The calibration and evaluation protocol is given in Appendix D.8.

Dynamic roots. For routing, the semantic root is the current destination. If its identity is known, a Poincaré chart is re-rooted by the Möbius translation $z_v \mapsto (-z_r) \oplus_c z_v$; its radius is exactly $d_c(z_r, z_v)$ and all native pair distances remain unchanged. If the destination is anonymous, Equation (C.6) applies, but exact delivery is attempted only when the root is identifiable; otherwise the method may provide prefix decisions or request another landmark. For reliable local forwarding,

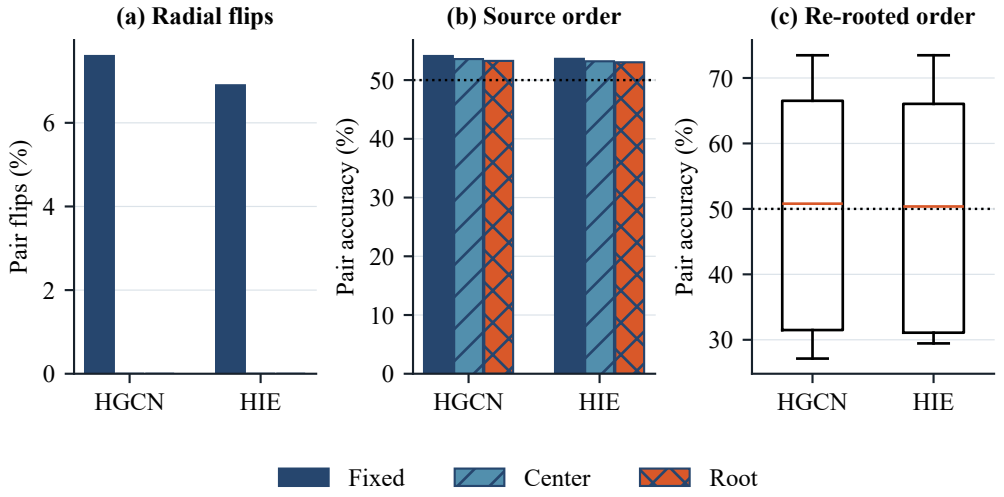

Figure 7: HIE audit: fixed-origin rankings change under isometry, while distance to its transported center is invariant but near chance for the declared root and compatible re-rootings. HIE denotes HGCN+HIE; dotted lines mark chance.

a compact m-root atlas stores each node's landmark-distance signature. The target advertises its signature, and a router orders neighbors by the ALT lower bound $\max_j |d(v, a_j) - d(r_t, a_j)|$, with the implemented additive secondary term $10^{-3} d(z_v, z_{r_t})$. A no-revisit safety rule prevents the two-node loops responsible for many failures of unconstrained greedy descent. This adds $O(m|V|)$ offline state and $O(m \deg(v))$ work per hop without modifying HGCN.

D Extended Experimental Protocols and Results

This section presents the intervention results, recovery diagnostics, and secondary outcomes. Appendix E consolidates datasets, information budgets, statistical units, software versions, and checkpoint provenance; Appendix F collects posterior sensitivity analyses, and Appendix G reports routing overhead.

D.1 Exact-Isometry Intervention Protocol

We apply exact Poincaré isometries to official HGCN checkpoints. On Airport (five seeds, 600 transforms), maximum link-AUC change is 1.1×10^{-16}, while an extreme transform replaces a median 10.1% and as much as 61.0% of the nodes in the lowest-radius decile. On Disease (three seeds, 360 transforms), the median and worst replacement are 22.7% and 68.4%. An explicit semantic anchor is invariant to numerical precision. Thus native prediction cannot choose among the conflicting origin-based rankings.

D.2 Published-Method Audit: Geometric Centering and Semantic Roots

HIE and HGCN (Yang et al., 2023; Chami et al., 2019) provide a direct test because HIE estimates a hyperbolic embedding center, aligns that center with the origin, and uses distance from it for hierarchical stretching. Its coordinate-equivariant center provides a geometric reference. The audit tests separately whether this reference identifies a declared semantic root.

The three-seed HGCN/HIE protocol and public-code replication settings are specified in Appendix E. We jointly transform all nodes and separately re-interpret the unrooted tree under 100 sampled semantic roots. This separates three questions: (i) whether the native metric changes, (ii) whether a score is coordinate stable, and (iii) whether that stable score identifies the requested semantic depth.

The intervention preserves native distances to at worst 7.2×10^{-13}. Fixed-origin pair order changes by 6.9–7.6% on average over all transformation magnitudes, exceeds 10% on average for the strongest HGCN transformations, and reaches 32.2% in the worst run. In contrast, the HIE-center and labelled-root distance rankings have zero measured coordinate flips. Nevertheless, HIE-center depth accuracy is 53.58% for HGCN and 53.20% after HIE training; over 100 admissible re-rootings it averages 48.90% and 48.64%. HIE raises native link AUC from 71.58% to 72.57%. Eight declared landmarks raise pair-order accuracy from 53.20% to 99.33% without changing the frozen AUC, recovering 98.6% of the available semantic gap. Data-dependent centering removes coordinate variation in the center-relative readout, while semantic-root identification still requires asymmetric evidence.

Author-code replication. We also run the authors' public HIE configuration (repository commit `131f97c255796868dc24992f2b651937fc0e46be`) on Disease node classification for five fixed seeds. The official HGCN baseline obtains $88.66 \pm 1.77\%$ test accuracy (mean$\pm$sample standard deviation), whereas `hire_tangent` obtains $89.92 \pm 7.83\%$. Four HIE runs lie between 92.52% and 94.49%, but one reaches only 75.98%; hence the mean improvement is not a stable across-seed claim. Because the public entry point does not save validation-selected embeddings, an external audit wrapper imports the official model and optimizer and changes only serialization/readout. Across the same five seeds and 100 paired isometries per seed, pairwise distance error is at most 3.1×10^{-11} and an intrinsic 5-NN decision has zero flips, while membership in the fixed-origin top radial quintile changes by 7.21% for HGCN and 6.71% for HIE on average, below the pre-registered 10% mean practical-effect gate. The public-code audit reproduces radial sensitivity with a smaller effect and substantial training variance.

D.3 Limitations of Single-Radius Depth Representations

The 2,665-node Disease benchmark is the HGCN link-prediction tree (Chami et al., 2019). Across three official link-prediction HGCNs, origin radius, an HIE-style weighted center, and distance to the declared root all obtain depth Kendall $\tau \approx 0.04$. Supplying the correct root after training is insufficient because symmetric link supervision did not encode directed depth. Oracle-root fine-tuning with pair-distance distillation raises τ to 0.525 while changing mean AUC by only $+0.081$ percentage points; direct ranking loss does not break this plateau and can substantially damage AUC. This motivates separate rooted charts rather than forcing all branching and depth information into one radius.

On the full noun DAG, stronger baselines sharpen the diagnosis. Uniform and degree-weighted HIE centers obtain 49.06% and 49.08% depth-pair accuracy. Entailment cones, which receive directed parent–child supervision, learn the local partial order (directed AUC $.748 \pm .008$ and F1 $.686 \pm .013$), yet their scalar radial depth remains $49.95 \pm .09\%$. Local relation, global root, and scalar depth are therefore distinct targets; Appendix E reports the full information-budget audit.

D.4 Root-Posterior Recovery on Disease and WordNet

We observe noisy landmark depths, evaluate every candidate root with Equation (C.6), and form a minimum-loss equivalence atlas. Anchors are selected without the root by farthest-point coverage. Disease uses 100 independently sampled semantic roots. We first use the 1,188-synset mammal taxonomy and then the full Open English WordNet noun DAG (Fellbaum, 1998): 84,956 synsets, 88,200 directed hypernym edges, and 3,133 multi-parent nodes. The full experiment uses landmark signatures rather than an all-pairs matrix and runs 50 independent corruption trials per setting.

The corresponding main-paper comparison is consolidated in Table 1. For additional context, a no-anchor medoid on Disease obtains 0.5% root accuracy and Kendall $\tau = .587$, while the mammal taxonomy reaches 100% root accuracy and $\tau = .993$ with eight anchors at 20% noise. On Disease, eight anchors are only 0.30% of nodes. At 20% corruption, exact MAP-root accuracy is 3%, yet Kendall is .970 and pair-order accuracy is 99.33%. This is the central root/decision separation. Four anchors are insufficient ($\tau = .615$ with high variance), supporting a coverage-dependent recovery. On WordNet the root is nearly uniquely resolved: the MAP root nearly matches the named root, while its undirected-distance chart can disagree with directed depth through metric shortcuts. A full diffuse posterior mean is weaker because it breaks true same-depth ties; it is retained as an

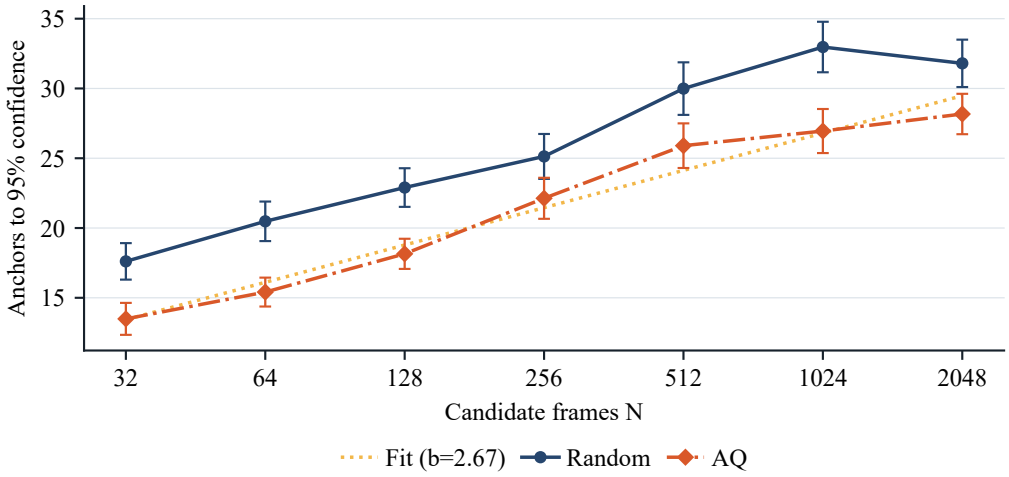

Figure 8: Query complexity in a noisy binary-signature model. Active-query budget grows approximately logarithmically with candidate-frame count and remains below random querying. AQ is active querying; the dotted fit has slope $b = 2.67$ per doubling.

ablation. On full WordNet, direct ordinal voting improves over a MAP chart at both 20% and 40% corruption despite identical exact-root accuracy (Table 1). Disease and WordNet therefore instantiate complementary non-identifiable and identifiable regimes. The posterior is stable for temperatures 0.1–1 and top-$K = 16$–64, while 16 anchors form the coverage knee; Figure 16 reports all ablations.

D.5 Evidence Budgets, Active Acquisition, and Local Semantic Roots

The Disease acquisition study fixes topology and native representation, samples 40 semantic roots, and uses 5,000 held-out ordinal decisions per trial under 20% anchor corruption. Predictive-entropy AQ, random acquisition, and farthest-point sampling share the same Huber/Gibbs inference rule. Figure 3 reports the budget, risk–coverage, local-root, and acquisition-time comparisons.

To test finite-frame dependence in isolation, we use a controlled binary noisy-anchor model analogous to the finite-hypothesis setting. Its Bernoulli corruption channel differs from the Gaussian channel of the sample-complexity corollary. Across 100 independent problems for each $N \in \{32, \dots, 2048\}$ and 15% bit corruption, the active budget needed for 95% posterior confidence follows an affine fit in $\log_2 N$ with slope 2.67 anchors per doubling and is lower than random querying at every size (Figure 8). In this controlled binary channel, active querying shows logarithmic candidate dependence; the experiment does not establish monotone frame growth for arbitrary graph metrics.

Finally, we construct four topology-defined regions with independently sampled semantic roots and reveal four noisy anchors per region. A single global-root posterior reaches $92.69 \pm 3.36\%$ local pair accuracy, whereas region-conditioned rooted charts reach $98.82 \pm 2.16\%$ (mean$\pm$sample standard deviation over 20 tasks). This intervention directly tests the multiple-chart design under region-specific semantic frames while holding the native model fixed.

D.6 Proximity Prediction and Directional Recovery in Learned Representations

We train matched 32-dimensional Poincaré and Euclidean embeddings on the full noun DAG with a symmetric negative-sampling link objective, then freeze them before root inference. Across three seeds, held-out redundant-parent link AUC is .736 for Poincaré and .726 for Euclidean. Yet even distance from the *oracle* semantic-root embedding obtains only 51.47% and 49.10% directed depth-pair accuracy, respectively; scalable posteriors using only learned pair distances remain near chance.

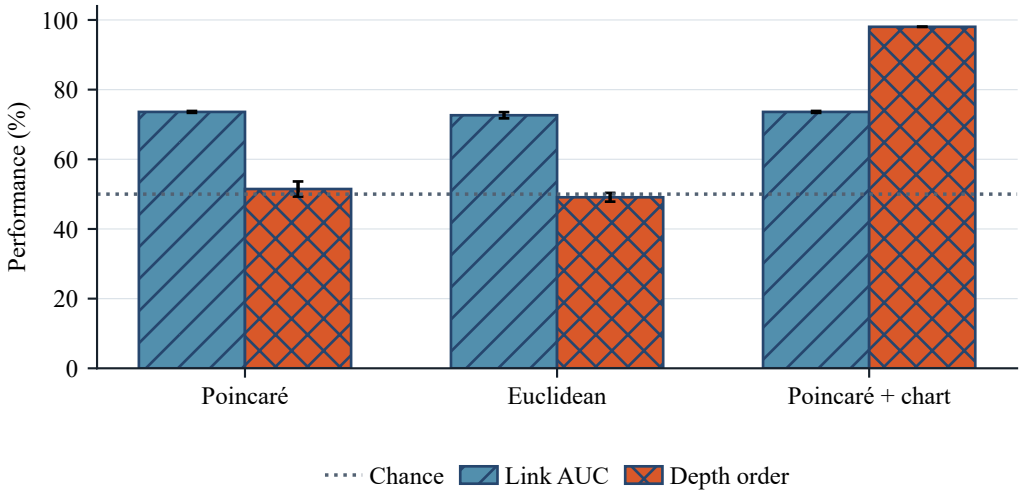

Figure 9: Matched WordNet representations recover proximity but not directed depth; asymmetric evidence restores direction while preserving the native decoder.

In contrast, attaching the structural ordinal atlas reaches 98.04% without changing native link predictions. Pairwise proximity alone leaves directed-depth accuracy near chance, whereas asymmetric structural evidence recovers the requested ordering.

D.7 Pre-Specified Visual-Hierarchy Audit on CIFAR-100

CIFAR-100 (Krizhevsky, 2009) supplies a qualitatively different application: 100 visual classes are nested in 20 superclasses. We use the public HypStructure pipeline (Sinha et al., 2024), which adds a hyperbolic tree-distance correlation loss and an explicit centering loss to supervised contrastive learning (Khosla et al., 2020). Pairwise hierarchy distances are symmetric, while the centering term chooses a coordinate convention; the superclass/subclass relation supplies the semantic direction against which that convention can be tested.

Models and controlled intervention. We train three matched ResNet-34 models with the official 500-epoch CIFAR-100 schedule: flat supervised contrastive learning, HypStructure without centering, and full HypStructure. The optimizer, augmentations, feature dimension, batch size, and random seed are identical. After training, every encoder is frozen. We form the official pipeline's equivariant hyperbolic gyromidpoint prototypes for 100 fine classes and 20 superclasses, then apply 100 exact common Poincaré gyrotranslations to test embeddings and all prototypes. For every transform we measure: (i) maximum pair-distance error, (ii) nearest-fine-prototype prediction flips, (iii) fixed-origin superclass-before-subclass flips, and (iv) the same order relative to the jointly transformed data center. This factorial design separates the hierarchy-distance loss (CPCC) from coordinate centering rather than comparing only the complete method to a flat baseline.

Semantic recovery and positive controls. We next reveal $m \in \{1, 2, 4, 8\}$ directed hierarchy landmarks selected either randomly, geometrically by farthest-point coverage, or by decision-cover greedy. FramePosterior predicts the 100 held-out superclass–subclass order relations and abstains when the credible frames disagree. Full directed hierarchy supervision is the upper bound. Entailment Cones and ConE (Ganea et al., 2018; Bai et al., 2021) are positive controls rather than expected failure cases: because their training observations are directed, Theorem A.1 permits them to identify an orientation. We report fine-class accuracy, hierarchy pair accuracy, semantic flip rate, coverage, selective risk, ECE, and native-task change. Posterior temperature, top-K, anchor policy, and abstention threshold are fixed independently of test superclass edges.

Table 3: CIFAR-100 exact-isometry audit (three training seeds; mean $\pm$ 95% t interval).

Training condition	Prototype acc.	Fixed-origin set flip	Center set flip
HypStructure + center	$76.87 \pm 1.58\%$	$18.95 \pm 1.58\%$	0
HypStructure $-$ center	$76.67 \pm 2.31\%$	$18.89 \pm 0.38\%$	0
Flat SupCon	$76.47 \pm 2.63\%$	$21.99 \pm 1.08\%$	0

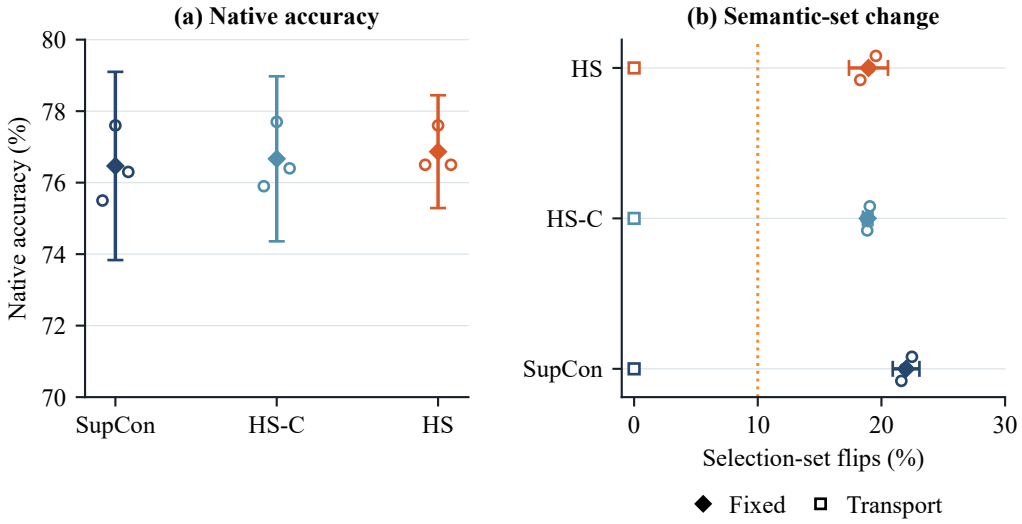

Figure 10: CIFAR-100 accuracy and semantic stability across three matched representations. (a) Fine- class accuracy; native predictions are unchanged over 100 isometries. (b) Change in the 80%-coverage selection set: circles show seed estimates, diamonds and whiskers the fixed-origin mean and 95% t interval, and squares the transported center (zero flips). The dotted line is the preregistered 10% threshold. SupCon is Flat SupCon, HS-C omits centering, and HS is full HypStructure; seeds are the replication units.

Pre-registered interpretation. Before inspecting the formal checkpoints, we fix the following gates. The intervention is valid only if native distance and prototype-decision changes are below 1%. A practically relevant fixed-origin failure requires either more than 10% semantic edge flips or a hierarchy/OOD/ retrieval drop exceeding five points. A coordinate-equivariant center should change by less than one point; if it remains materially below the directed/full-information upper bound, that supports the predicted distinction between choosing a coordinate center and identifying a semantic root. FramePosterior succeeds only if it recovers at least half of this semantic gap while changing native accuracy by less than one point. If the full model is already semantically correct and stable, this test is recorded as a null condition.

Results. All nine 500-epoch runs and frozen-encoder audits completed. Table 3 reports seed means and 95% intervals, with the training seed—not the 100 paired isometries—as the replication unit. Maximum pair-distance error is 5.5×10^{-11} and nearest-prototype prediction flip is exactly zero. Nevertheless, a fixed-origin 80%-coverage set changes by 18.89–21.99% on average. Transporting a data-dependent center gives zero measured selection flips. Thus the visual models reproduce the intrinsic/frame separation across architectures and seeds.

The superclass prototypes remain closer than their fine-class children under every audited random and edge-aligned intervention. The largest observed change in 80%-coverage selective accuracy is 1.75 points, below the five-point threshold. Because these anchors have no nonzero superclass–subclass depth gap, the noisy visual-anchor experiment has no recovery gap to measure. The CIFAR-100 results therefore quantify prediction stability and selective-set turnover; Disease and WordNet provide the recovery tests with nonzero decision gaps.

Table 4: ImageNet-100 radial depth audit at unit boost rapidity ($\rho = 1$). Full HypStructure uses ten seeds and controls five; directions are paired interventions within each seed.

Training condition	Native acc.	Within-one-level drop	Depth-MAE increase
Flat SupCon	89.38%	11.63 pp	0.65
HypStructure − center	89.36%	11.60 pp	0.59
HypStructure + center	$89.38 \pm 0.08\%$	12.68 ± 1.49 pp	0.614

We additionally lock a smaller random/edge-aligned stress family before testing application loss. In this family, fixed-origin 80%-coverage membership changes by only 0.52–0.66%, the mean AURC difference from the transported-root oracle ranges from -0.077 to $+0.048$ percentage points, and hierarchy accuracy drops by at most 0.23 points. With 40% corrupted hierarchy anchors, posterior averaging is substantially more stable than MAP (e.g., 99.96% versus 83.30% held-out edge accuracy at eight anchors), but the fixed and oracle accuracies are both already 100%. This is a null recovery condition for the isometry-induced-gap hypothesis.

Finally, motivated by norm-based hyperbolic OOD detection (Kasarla et al., 2026), we perform a diagnostic rather than a reproduction of that full method: CIFAR-100 test images are ID, CIFAR-10 test images are near-OOD, and radius is the score. Across 100 paired isometries for each of the nine frozen encoders, baseline radial AUROC is 79.31%, 79.30%, and 71.62% for full HypStructure, no-center HypStructure, and Flat SupCon. Mean absolute AUROC changes are 0.34, 0.32, and 1.09 points, while the worst decreases are only 0.61, 0.56, and 0.74 points. The intrinsic nearest-prototype control changes by exactly zero (maximum score error 2.3×10^{-11}). The OOD AUROC change remains below the prespecified five-point threshold.

D.8 ImageNet-100: Radial-Depth Loss and Sparse Calibration

Because CIFAR-100 classification is largely intrinsic, we next test a downstream decision that explicitly uses the disputed frame. We train the public HypStructure ImageNet-100 pipeline (Sinha et al., 2024; Deng et al., 2009) under three matched conditions—Flat SupCon, HypStructure without centering, and full HypStructure. The primary full model uses ten independent training seeds; the two mechanism controls use five seeds each. The authorized subset has 128,241 training and 5,000 validation images. The native representation is the frozen 512-dimensional penultimate feature on which CPCC acts; leaf classification uses intrinsic nearest-prototype distance.

The selected WordNet ancestor closure contains 325 nodes and 338 edges, including 14 multi-parent nodes. Many raw nodes induce the same set of the 100 observed leaves and are therefore indistinguishable. We collapse them before evaluation into a deterministic 167-node, 175-edge decision-equivalence quotient DAG. This removes unary aliases rather than counting identical prototypes as failed directed edges. Full HypStructure attains 97.09% source path-edge accuracy on this quotient, versus 88.99% for Flat SupCon, while native leaf accuracy is 89.38% versus 89.38%.

Frozen depth readout and intervention. Within each taxonomy depth we reserve every third quotient node as a class-disjoint test set and fit a one-dimensional monotone depth decoder on the remaining nodes. The decoder is then frozen. We apply paired exact Poincaré isometries at the complete pre-specified displacement schedule $\rho \in \{0.5, 1, 2, 3, 4\}$, using equal random and edge-aligned directions. No radius or direction is selected using test loss. A fixed-root decoder receives the transformed prototype but not the transported root; the oracle transports both. Every intrinsic distance and checked leaf decision is preserved, and the oracle reproduces source depth MAE to approximately 10^{-15}.

At the primary $\rho = 1$ dose, Table 4 passes both locked application gates. The depth readout changes substantially on the hierarchy-aware encoders despite their better source hierarchy. At $\rho = 2$, within-one-level accuracy drops by 38.73, 42.40, and 41.83 points for Flat SupCon, HypStructure without centering, and full HypStructure, respectively. In contrast, rooted ancestor retrieval remains stable under the same interventions.

Sparse and noisy chart recovery. We next reveal source-frame radii for semantic landmarks chosen only from calibration nodes. Landmark targets receive Gaussian noise equal to 5% of the source-

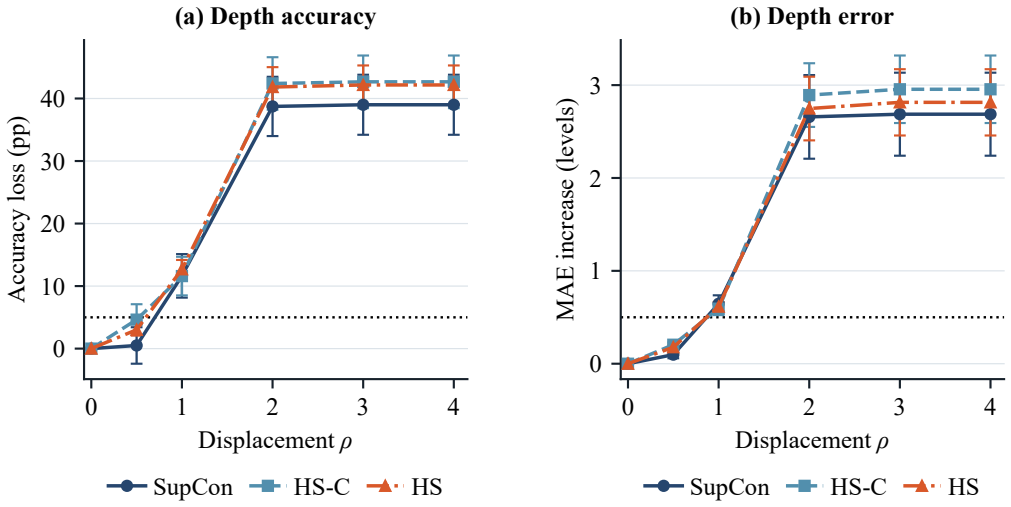

Figure 11: As semantic-frame displacement grows, the frozen radial readout degrades while the transported-root oracle and native classifier remain invariant. Whiskers are seed-clustered 95% intervals (ten full-model and five control seeds); dotted lines mark the 5-pp and 0.5-level gates. SupCon is Flat SupCon, HS-C omits centering, and HS is full HypStructure. Data also appear in main Figures 4b and 5a.

radius standard deviation. We compare random acquisition, intrinsic farthest-point sampling, and source radial coverage (RDC) under identical landmark-count budgets. RDC covers stored source-frame calibration radii; its selection uses source calibration geometry in addition to the queried noisy landmark values. A monotone chart maps observed radii to the canonical gauge; an anchored resampling collection of charts averages depth predictions. This implements decision-level calibration with a resampled monotone ensemble. At $\rho = 1$, fixed-root MAE is 1.906 levels, the transported oracle/source MAE is 1.300, and 32 active landmarks reduce MAE to 1.340, recovering 89.88% of the gap under seed-wise averaging without changing native accuracy. Eight active landmarks recover 80.69%, versus 58.81% for random acquisition. Uncertainty is estimated over ten training seeds, with directions, noise draws, and repeats paired within seed.

The simple chart baseline is already strong: the resampled-chart ensemble improves mean depth MAE by only 0.0035 levels over the MAP chart at the 32-landmark primary setting, and ensemble disagreement does not yield a monotone risk–coverage curve. These results support sparse gauge recovery under the declared chart model.

D.9 UNCHA: Published Radial-Uncertainty Mechanism Audit

The preceding experiment deliberately constructs a radial downstream task. We next ask whether the same failure reaches a radial quantity used inside a published model. UNCHA (Kim et al., 2026) defines hyperbolic uncertainty as $u(z) = \mathrm{softplus}(-\|z_{\mathrm{space}}\|_2)$ and uses $\exp(-u/2)$ to scale its global–local contrastive logits. We use the authors' released ViT-S/16 checkpoint (SHA-256 prefix `7125eb6b`) without retraining. On all 5,000 ImageNet-100 validation images we pair the standard whole-image view with a deterministic central 70% part crop and the corresponding class prompt. We then apply the same exact Lorentz boost to all part and text embeddings. Ten directions and the complete, pre-specified rapidity schedule $\{.25, .5, 1, 1.5, 2\}$ are evaluated; directions are paired interventions, not independent model seeds.

At the primary unit-rapidity intervention, maximum pair-distance error is 4.43×10^{-14} and all 5,000 native class predictions are unchanged. Nevertheless, membership in the highest-uncertainty quintile changes by 69.70% (95% interval $[61.54, 77.81]$), and the published uncertainty-scaled NLL rises from 1.3755 by 5.89% $[5.86, 5.93]$. At rapidity two, the largest semantic-set flip is 80.2% and the largest NLL increase is 6.33%. Applying the inverse boost only to evaluate the semantic frame

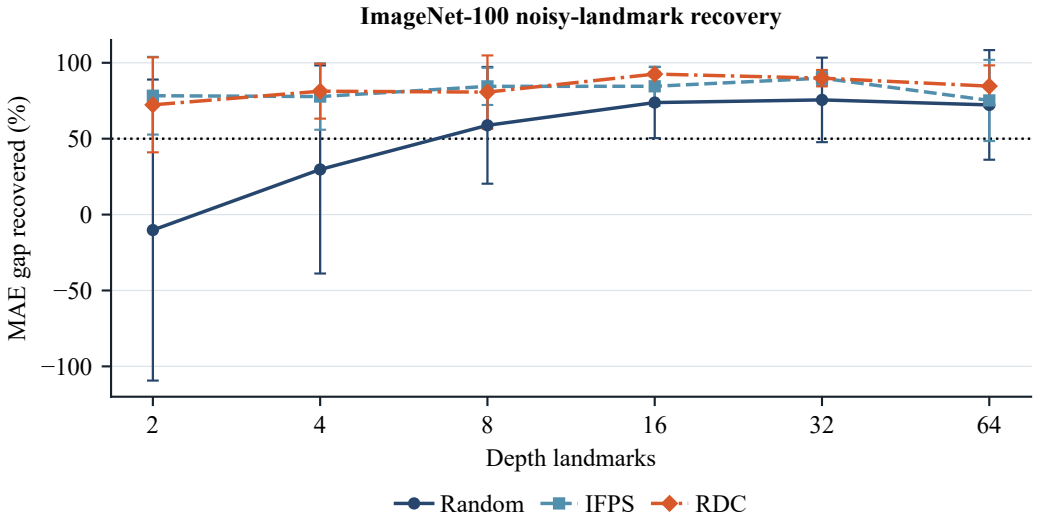

Figure 12: At $\rho = 1$ and 5% landmark noise, sparse landmarks recover most of the ImageNet-100 radial depth-MAE gap. Whiskers are 95% intervals over ten seeds; the dotted line marks 50% recovery. IFPS is intrinsic farthest-point sampling and RDC is source radial coverage. Data also appear in main Figure 5b.

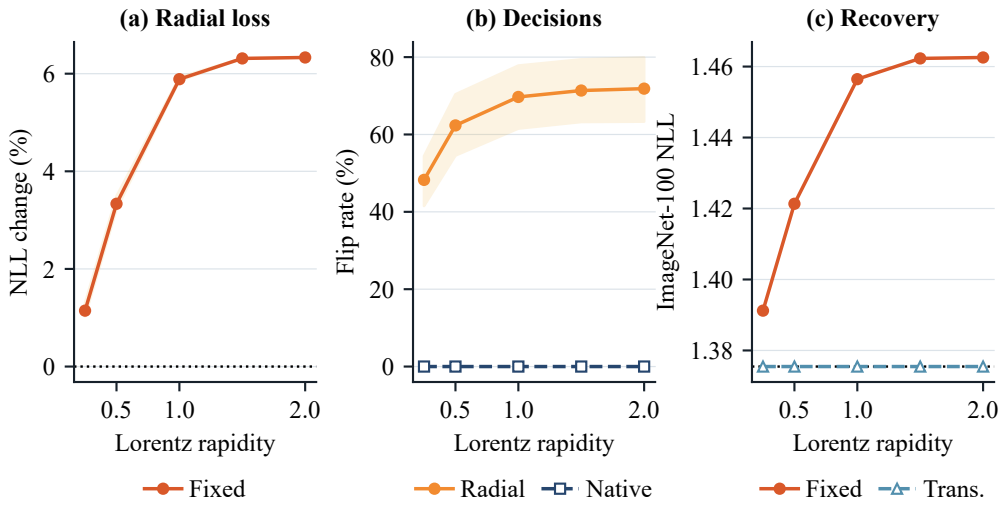

Figure 13: Official UNCHA checkpoint audit. Lorentz boosts preserve class decisions but change fixed-origin radial uncertainty and held-out scaled loss. Transporting the frame (Trans.) restores the source loss; bands show the central 95% over ten sampled directions.

recovers the source uncertainty to 1.6×10^{-16} and the source NLL to numerical precision. Intrinsic classification remains stable while the explicitly radial uncertainty score changes. This checkpoint-level audit tests that mechanism rather than re-evaluating UNCHA's reported benchmarks.

D.10 Destination-Conditioned Routing and Endpoint Identifiability

We first hide the destination identity on the 3,397-node OpenFlights graph and provide only noisy landmark distances. With 16 anchors and 20% corruption, MAP root identification is 15% and exact

Table 5: Held-out routing destinations with frozen HGCN embeddings (mean $\pm$ 95% t interval over 100 destination means, each averaged across five seeds).

Routing score	Delivery	Next-hop agree.	Stretch (success)
Current destination / re-rooted chart	$45.38 \pm 6.39\%$	63.79%	1.117
Re-rooted chart + no revisit	$69.90 \pm 6.67\%$	64.86%	1.619
16-root atlas + no revisit	$65.18 \pm 6.52\%$	54.96%	2.014
32-root atlas + no revisit	$70.76 \pm 6.01\%$	58.15%	1.929
64-root atlas + no revisit	$\mathbf{77.92 \pm 5.33\%}$	65.80%	1.699
Graph shortest-path oracle	100%	100%	1.000

delivery is 15.24%; posterior averaging gives useful partial next-hop agreement but only 3.02% delivery. This negative result matches the anonymous-delivery proposition: decision averaging cannot manufacture endpoint identity.

We then evaluate the realistic known-address setting using five official Airport HGCN seeds on the 3,158-node largest component. We split destinations 512/96/100 for train/validation/test and use 50 sources per held-out test destination. Re-rooting the learned chart at each destination is algebraically identical to intrinsic target distance. Compact 16/32/64-root atlases and the no-revisit rule are fixed before this test.

The 64-root atlas improves delivery over re-rooted HGCN by 32.54 points on identical episodes; averaging across seeds and pairing the 100 destination means gives 95% CI $[27.41, 37.67]$. It exceeds the no-revisit control by 8.02 points $[3.26, 12.79]$. Frozen native AUC changes by exactly zero. A generic learned MLP distance head is a negative baseline: it reaches only 35.88% delivery (60.28% with no revisit), so the topology-informed atlas outperforms this generic distance head. Its higher reliability costs stretch 1.699 and leaves 22.08% undelivered. Its quantized state is 0.385 MiB. A post-hoc scale curve reaches 85.22% delivery with 128 charts and 0.771 MiB state, but the held-fixed 64-chart system remains the primary result; Appendix G reports clustered tests and overhead.

Figure 17 consolidates the routing state–delivery trade-off; the primary WordNet and routing comparisons appear in Figure 5.

E Data, Information Budgets, and Reproducibility

E.1 Task-Specific Comparators and Information Budgets

Information budgets. Every native link model is trained before direction recovery and then frozen. Origin radius, HIE centers, and the matched Euclidean/Poincaré embeddings see only symmetric links. Entailment cones receive additional directed parent–child edges and form a separate information-budget condition. Root posterior methods use the undirected structural metric and explicitly counted noisy anchor depths. WordNet evaluation uses directed taxonomy depth; its chart model is an approximation to that target. ImageNet radial coverage additionally uses stored source calibration radii for selection, and its budget counts noisy landmark values supplied to recovery. The routing atlas sees graph landmark distances and destination addresses but no target labels or shortest-path next hops.

Independent units. WordNet corruption trials independently redraw corrupted anchors. Learned WordNet representations report three training seeds. The primary ImageNet-100 full-HypStructure condition reports ten training seeds, while its two mechanism controls report five. Airport routing uses five official HGCN seeds with the same 100 held-out destinations and 50 sources per destination. We first average each destination's delivery across seeds, then compute 95% t intervals and paired tests over 100 destination means, conditional on these frozen models. The 500 seed/destination records contain paired repetitions of these same destinations. The CIFAR study uses three independently trained seeds for each factorial condition; repeated isometries and anchor corruptions are paired interventions within a seed and are not counted as additional model runs.

Table 6: Task-specific comparisons, information budgets, and headline outcomes. Baselines are matched within each task and do not form a shared leaderboard. Values shown as mean±half-width use 95% t intervals. Detailed MAP–posterior estimates are reported in Table 1; paired transformations are interventions, not independent model replications.

Experiment / target	Matched comparator(s)	Budget / evaluation unit	Key result
HIE / fixed-origin depth	HGCN vs. HGCN+HIE; transported center (Chami et al., 2019; Yang et al., 2023)	3 seeds/condition; 100 paired isometries/seed	Up to 32.2% depth flips; native decoder unchanged
ImageNet-100 / radial depth	HypStructure; no-center and Flat SupCon controls; transported-frame oracle; random vs. active landmarks (Sinha et al., 2024; Khosla et al., 2020)	10 full-model seeds; 5/control; 107 existing fit/selection nodes; 32 noisy landmark queries; 60 test nodes	12.68 ± 1.49 pp accuracy loss; active landmarks recover 89.88% of MAE gap; class predictions fixed
UNCHA / radial uncertainty	Fixed-origin readout vs. inverse-transported reference (Kim et al., 2026)	1 official checkpoint; 5,000 images; 10 paired directions	69.70% uncertainty-set turnover; scaled NLL +5.89%; class decisions fixed
Disease / ordinal decisions (Chami et al., 2019)	Active query vs. random and farthest-point; local vs. global root	40 acquisition roots: 8 anchors, 20% noise; 20 local-root tasks: 16 anchors + region ID	AQ 99.78% vs. random 95.87% pair accuracy; local roots add 6.13 pp
WordNet / directed ordinal depth (Fellbaum, 1998)	MAP vs. posterior ordinal vote with matched anchors	84,956-node DAG; 16 anchors; 50 paired trials at 20/40% noise	Posterior gain +1.48 pp (20%) / +1.50 pp (40%); native AUC unchanged
CIFAR-100 / selective decisions	Centered HypStructure, no-center, Flat SupCon; transported center (Sinha et al., 2024; Khosla et al., 2020)	3 seeds/condition; 100 paired isometries	Class predictions fixed; 18.89–21.99% fixed-origin selection-set turnover
Airport / destination routing	Frozen HGCN, HGCN+no-revisit, landmark atlas (Chami et al., 2019)	5 seeds; 100 shared destinations; 50 sources/destination; 64 charts	Atlas-64 77.92% vs. +NR 69.90% delivery; native AUC unchanged

Compute environment. The primary GPU experiments ran on Ubuntu 22.04.4 LTS with 40 logical CPUs (Intel Xeon Silver 4210 at 2.20 GHz), 94 GiB RAM, and two NVIDIA GeForce RTX 4090 GPUs (24 GiB each). The software stack was Python 3.10.12, PyTorch 2.3.0a0+40ec155e58.nv24.03, and CUDA 12.4 with NVIDIA driver 535.288.01.

Strong direction baselines. We implement uniform- and degree-weighted HIE centers using the Einstein midpoint on the frozen Poincaré representation. Hyperbolic entailment cones use the published aperture and angular violation objective (Ganea et al., 2018), trained for 30 epochs on directed parent–child edges. Table 7 shows that cones learn local directed relations, while their radius still fails as a global root-depth statistic.

HIE reproduction and isometry intervention. Our first link-prediction audit implements Equations (4)–(6), (14), and (28) from Yang et al. (2023) on the official HGCN code snapshot. The native loss is evaluated on the unaligned representation; the auxiliary loss uses the uniform Poincaré gyromidpoint and identity HDO weighting, exactly as stated in the paper. We use 200 epochs, 16 dimensions, Adam with learning rate .01, three seeds (11, 23, 37), and $\lambda = 10^{-3}$ from the reported grid. Larger identity-stretch values drove this older HGCN implementation numerically to the ball boundary; those incomplete runs are treated as a stability diagnostic outside the completed-checkpoint analysis. Every completed checkpoint is subjected to 100 fresh gyrotranslations and the same 100,000 node-pair queries. Re-rooting changes only the declared semantic frame, not the graph, embedding, center, or native decoder.

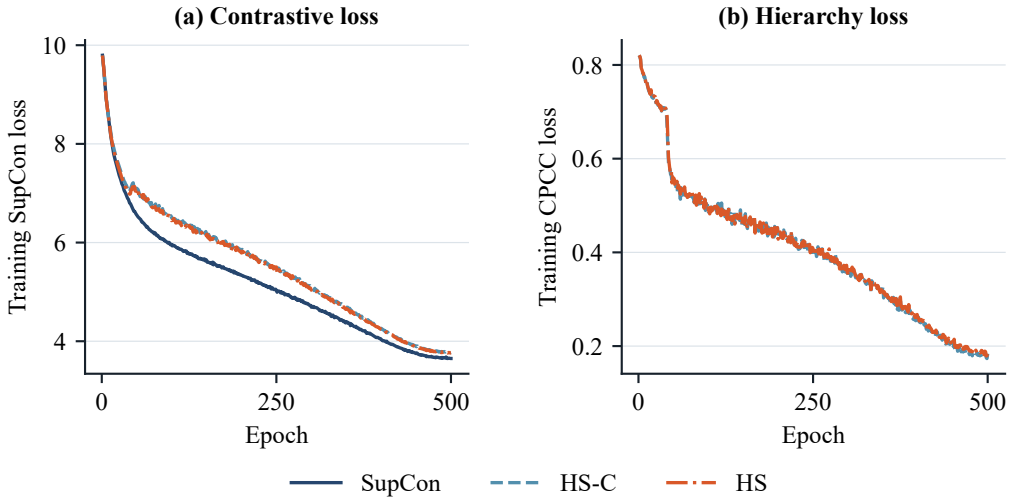

Figure 14: Recorded 500-epoch CIFAR-100 training losses: (a) supervised-contrastive loss and (b) CPCC loss for the two HypStructure variants (Flat SupCon has no CPCC term). SupCon is Flat SupCon, HS-C omits centering, and HS is full HypStructure. One training run per condition is shown; no validation curves or seed aggregates are available.

We separately snapshot the authors' public HIE repository at commit `131f97c255796868dc24992f2b651937fc0e46be`. Without modifying its entry point, we run `train.py` on `disease_nc` with HGCN, dimensions 8, learning rate .01, patience 500, the repository defaults for all remaining settings, and seeds 1234, 2345, 3456, 4567, and 5678. Conditions are `hyp_ireg=0` and `hire_tangent` with $\lambda = .1$. The entry point does not serialize its validation-selected embedding, so a second script imports the same public model, data loader, and optimizer, repeats the unchanged training loop, and saves only that embedding for external readout. We distinguish the original entry-point runs from this instrumented official-code audit. The intrinsic control is 5-NN over Poincaré distances. The frame-dependent decision is membership in the largest-radius node quintile under 100 common isometries per seed.

CIFAR-100 protocol and leakage checks. The three ResNet-34 conditions use the public HypStructure code and its 500-epoch CIFAR-100 schedule. Each final encoder is frozen before extracting all 50,000 training and 10,000 test features. Prototypes and interventions are computed in float64; model inference remains in the checkpoint's native precision. Exact row-hash overlap between train and test features, source file hashes, and every reported aggregate are recomputed by an independent script that imports none of the intervention code. CIFAR-10 OOD data, when used, are restricted to the official 10,000-image test batch (MD5 `40351d587109b95175f43aff81a1287e`) and pass through each frozen CIFAR-100 encoder with the CIFAR-100 training normalization. The independent feature audit finds exactly ten train/test duplicate rows for every encoder. Byte-level hashes of the raw CIFAR-100 arrays identify the same ten test images as exact copies in the official training split (five retain their label and five use a different fine label), so this is a dataset property rather than extraction leakage. They comprise 0.1% of the test set; even an adversarial removal changes a bounded test mean or the ID contribution to AUROC by at most 0.1 percentage points, below the locked one- and five-point gates. The artifact records the duplicate count and input SHA-256 hashes.

Figure 14 shows the recorded per-condition training curves beside the nine-seed task results in Table 3 and Figure 10.

ImageNet-100 and UNCHA protocols. The ImageNet-100 experiment uses 128,241 training and 5,000 validation images in 100 classes. We run the public HypStructure ImageNet configuration for 20 epochs and freeze the selected encoder. Full HypStructure uses seeds 0–9; Flat SupCon and

no-center controls use seeds 0–4. Interventions are evaluated in float64, and uncertainty intervals cluster all directions, noise draws, and repeats within the independent training seed.

For the published radial-method audit, we snapshot the public UNCHA repository and load the authors' ViT-S/16 checkpoint (complete SHA-256 `7125eb6b5b1e0c5ac7a89f2c458ef7ef07dcc2bab7621aa68e853c910f7665b5`). Strict loading differs only by a newly present, unused text-head parameter in the repository snapshot. We encode all ImageNet-100 validation images, deterministic central 70% part crops, and one class prompt per synset. The same analytic Lorentz boost is applied to image and text embeddings in float64 for ten fixed random directions and rapidities $\{.25, .5, 1, 1.5, 2\}$. Native pair distances and predictions are the intrinsic control. The audited radial mechanism is the released implementation's $u(z) = \mathrm{softplus}(-\|z_{\mathrm{space}}\|_2)$ and its $\exp(-u/2)$ logit scaling. The inverse boost evaluated only at readout time is the transported frame oracle. No direction or rapidity is chosen using outcome loss.

Table 7: Full-WordNet direction baselines (mean ± standard deviation over three seeds).

Method	Information	Depth pair	Directed AUC	Link AUC
Origin radius	symmetric links	.4907 ± .0014	–	–
HIE uniform center	symmetric links	.4906 ± .0014	–	–
HIE degree center	symmetric links	.4908 ± .0013	–	–
Entailment cones	directed edges	.4995 ± .0009	.7482 ± .0081	.8345 ± .0124

Table 7 contrasts local directed-relation prediction with global scalar depth under the corresponding information budgets.

F Sensitivity to Posterior and Landmark Design

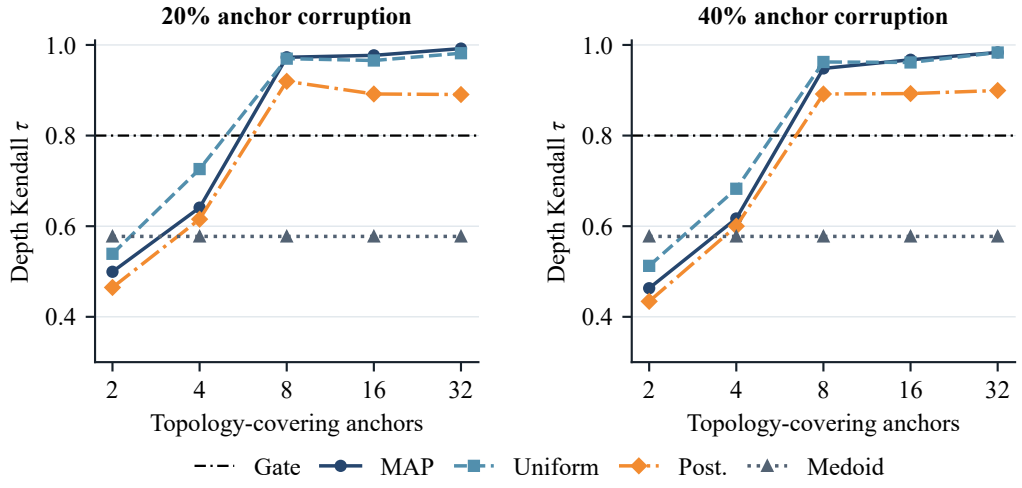

Figure 15: Disease direction recovery with independently corrupted landmark depths. Eight topology-covering anchors pass the preregistered Kendall threshold despite poor exact-root recovery. Post. is posterior multi-chart recovery; Uniform averages charts and Medoid is the structural-medoid baseline. Reference line: Kendall $\tau = 0.8$.

The one-factor-at-a-time base configuration is 16 farthest-point anchors, 20% independently corrupted depths, temperature $T = 1$, and top-$K = 64$ charts. Each setting has 50 trials and 20,000 fixed non-tied ordinal queries. Figure 16 establishes three practical regimes: 16 anchors are the coverage knee; $T \in [0.1, 1]$ and $K \in [16, 64]$ are stable; and farthest/degree landmarks are slightly stronger than random selection. At 60% corruption, all-pair accuracy remains 97.85%, while the margin-selected subset remains 99.10% accurate.

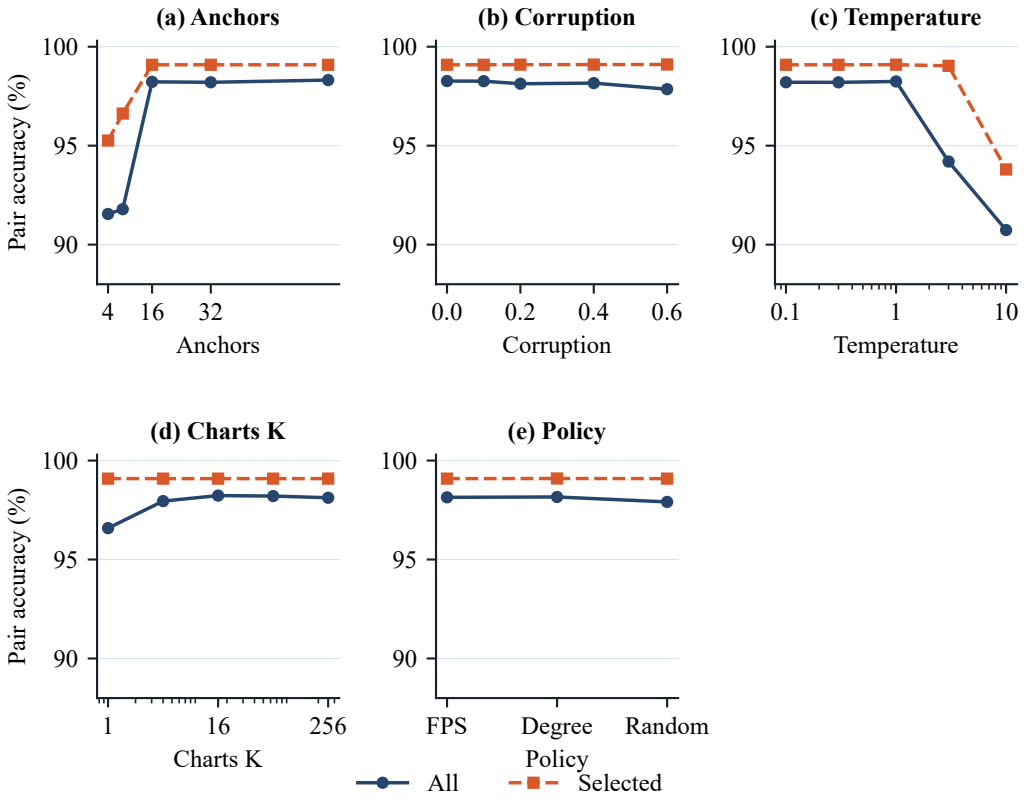

Figure 16: Full-WordNet posterior ablations, varying one factor per panel. All uses every ordinal query; Selected keeps posterior margins of at least 0.5. FPS is farthest-point sampling.

Compute. Median posterior trial time is below 0.45 seconds through 64 charts; 256 charts require 1.57 seconds. The graph signature table is computed once. The root's nearest competing signature is one graph-distance unit away for every tested anchor budget, explaining why bounded corruption, not exact noiseless collisions, controls the transition.

G Routing Scale, State, and Statistical Tests

With $n = 3158$ nodes and m landmarks, quantized atlas state is exactly $2nm$ bytes. Hence state grows linearly rather than quadratically: 0.024, 0.048, 0.096, 0.193, 0.385, and 0.771 MiB for $m = 4, 8, 16, 32, 64, 128$. Median scoring time for one node's neighbor set remains 6.6–7.4 μs because distance-vector subtraction is vectorized at this scale. The pre-registered 64-chart system is the main result; 128 charts are reported only as a post-hoc scale ablation.

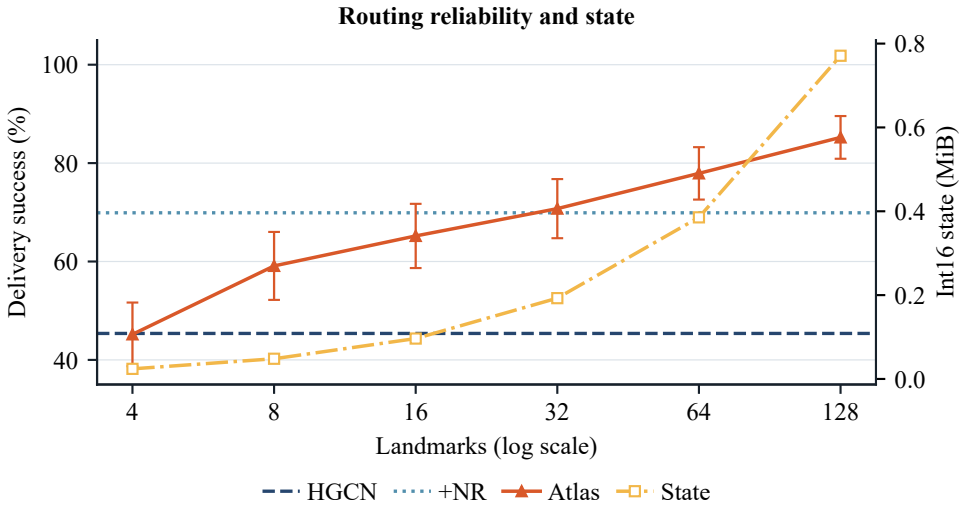

Figure 17: Reliability–state trade-off on held-out destinations (95% intervals over destination clusters). +NR is HGCN with no revisit; state uses the right axis.

Over the 100 destination means, Atlas-128 exceeds HGCN by 39.84 points (95% CI $[34.46, 45.23]$, paired $p < 2 \times 10^{-26}$), HGCN plus no-revisit by 15.32 points ($[10.32, 20.33]$, $p < 3 \times 10^{-8}$), and Atlas-64 by 7.31 points ($[4.47, 10.16]$, $p < 2 \times 10^{-6}$). The primary comparison uses the prespecified 64-chart configuration; these post-hoc results characterize additional state–delivery scaling.

H Failure Modes and Limits of Applicability

Four conditions limit recovery. First, symmetric learned distances alone stay near chance for direction even when link AUC is useful. Second, a single inferred geometric center can place the semantic root near the outer boundary. Third, excessively diffuse posteriors ($T = 10$) average incompatible orders and reduce pair accuracy to 90.74%. Fourth, anonymous exact routing remains bounded by endpoint identification. Each case reflects limited asymmetric information; resolving it requires an additional observation, an informative anchor, or abstention.