SEPARATING ANGULAR ALIGNMENT FROM RADIAL ORDERING IN HYPERBOLIC VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperbolic vision–language models are well suited to modeling the hierarchical structure of multimodal data, because hyperbolic volume grows exponentially with radius, enabling low-distortion representations of tree-like taxonomies from general to specific concepts. However, existing pretraining objectives for hyperbolic vision–language models entangle semantic alignment with entailment modeling, forcing radial coordinates to simultaneously encode semantic matching and hierarchical structure, leading to conflicting geometric constraints. In this paper, we introduce AURORA, a hyperbolic vision–language framework that decouples the objectives through angular alignment and a novel *Radial–Angular Order Cone* (RAOC) for entailment modeling. Angular alignment captures semantic similarity while preserving the radial structure for representing hierarchies. Unlike conventional entailment cones, RAOC determines angular allowance from relative parent–child radial gaps. It is incorporated into hierarchical entailment contrastive training to capture partial-order relations. We further introduce a norm regularization term that prevents undesirable radial degeneration for stable training. We train models on millions of image–text pairs and evaluate them across zero-shot classification, image–text retrieval, hierarchical reasoning, and compositional understanding. Our model consistently outperforms strong Euclidean and hyperbolic vision–language baselines, demonstrating the effectiveness of decoupling semantic alignment from hierarchical entailment modeling.

1 INTRODUCTION

Hyperbolic vision–language models, such as MERU (Desai et al., 2023) and HyCoCLIP (Pal et al., 2025), embed images and text in hyperbolic spaces (negatively curved Riemannian manifolds) (Ratcliffe, 1994) to represent hierarchical relations among visual and linguistic concepts. In the Lorentz model (Cannon et al., 1997), a model of hyperbolic space, distance from the origin serves as a proxy for semantic specificity, with more general concepts represented at smaller radii (Nickel & Kiela, 2017). Under the hyperbolic entailment convention, parent–child relations form a partial order (Vendrov et al., 2016): abstract parents entail more specific children, whose embeddings lie within their parents' entailment cones (Ganea et al., 2018). For example, the text "*a dog*" entails the more specific description "*a dog running on grass*", and an image depicting "*a dog running on grass*" is compatible with both descriptions. Modeling this partial order allows matching visual and textual concepts to occupy different levels of the hierarchy, reflecting their differences in specificity. Hyperbolic volume grows exponentially with radius (Krioukov et al., 2010), accommodating the increasing number of child nodes with depth. Hyperbolic space also enables low-distortion tree embeddings (Sarkar, 2012; Sala et al., 2018), making it naturally suitable for representing hierarchies.

Learning the hierarchical structures of data requires compatible objectives for semantic alignment and hierarchical entailment. However, existing hyperbolic pretraining objectives impose conflicting constraints on radial information: semantic alignment favors similar radii for matched representations, while hierarchical entailment uses radial separation to distinguish parents from children. Moreover, entailment supervision based on positive pairs encourages specified parent–child relations while leaving unpaired relations unpenalized. As a result, semantically unrelated representations may still form spurious hierarchical relations. For example, in Figure 1, the text "*apple*" could incorrectly entail an image depicting "*a rocket successfully lifting off*", with no entailment penalty for this mismatch.

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.

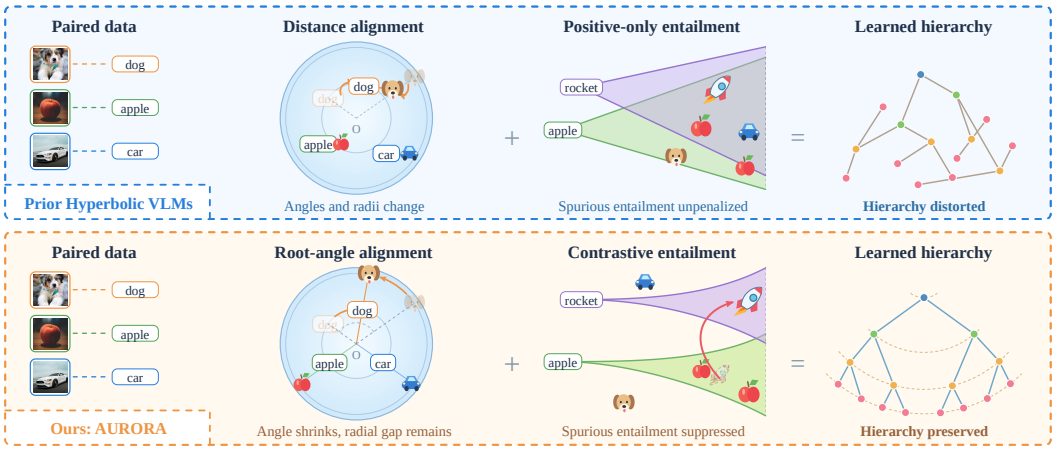

Figure 1: **Motivation. Top:** Prior hyperbolic VLMs pull paired embeddings together, disrupting radial ordering and entailment. The apple image can even lie closer to the Lorentz origin than "apple," while spurious relations distort the hierarchy. **Bottom:** AURORA preserves radial separation and cleaner entailment: the rocket returns to its own cone, incorrect relations are reduced, and a coherent hierarchy emerges.

We analyze the geometric compatibility of Lorentz-distance alignment and conventional entailment cones to clarify the optimization conflict, identifying two limitations. Firstly, minimizing Lorentz distance between a matched parent–child pair favors reducing the radial gap that models their difference in specificity. Conversely, an unmatched child can remain inside a spurious parent's cone even when moved arbitrarily far outward along its radial direction. These cases reveal a mismatch between semantic discrimination and hierarchical entailment. Secondly, conventional cone apertures (Le et al., 2019) are inversely coupled to the norm of the parent representation. As this norm approaches the admissible lower limit, the cone becomes less restrictive for partial-order relations.

In this paper, we propose AURORA, a pretraining framework that separates **A**ng**U**lar alignment from **R**adial **O**rdering in hype**R**bolic vision–l**A**nguage models. AURORA learns semantic alignment and hierarchical entailment through distinct, complementary objectives. For semantic matching, we perform angular alignment using cosine similarity between the spatial directions of Lorentz embeddings, allowing their radial positions to model specificity. For hierarchical entailment, we introduce the *Radial–Angular Order Cone*, which controls angular allowance through the parent–child radial gap, replacing the dependence on the parent's norm in conventional cones. We prove that the resulting geometric entailment relation defines a partial order with nested descendant regions, ensuring that a child's descendants are also descendants of its parent. We further introduce *hierarchical entailment contrastive* training, using negative pairs to suppress spurious entailment between unpaired examples. Finally, we present a norm regularization term that prevents undesirable radial degeneration, stabilizing the radial scale during training.

Our contributions are summarized as follows:

- We introduce AURORA, a hyperbolic vision–language pretraining framework that decouples angular semantic alignment from radial modeling of hierarchical entailment.
- We define the *Radial–Angular Order Cone* (RAOC), prove that its geometric entailment relation defines a partial order, and develop hierarchical entailment contrastive (HEC) training on RAOC to better reinforce hierarchical relations.
- We train hyperbolic vision–language models with AURORA and demonstrate consistent improvements in zero-shot classification, retrieval, hierarchical reasoning, and compositional understanding over competitive Euclidean and hyperbolic baselines.

2 PRELIMINARIES

Hyperbolic space is a Riemannian manifold with constant negative curvature. We adopt the d-dimensional Lorentz model with curvature $-c$, where $c > 0$ denotes the curvature magnitude (Nickel & Kiela, 2018). The Lorentz model represents hyperbolic space as the upper sheet of a two-sheeted hyperboloid in $(d+1)$-dimensional Minkowski space.

For vectors $\mathbf{x} = [x_0; \tilde{\mathbf{x}}]$ and $\mathbf{y} = [y_0; \tilde{\mathbf{y}}]$ in $\mathbb{R}^{d+1}$, $x_0, y_0 \in \mathbb{R}$ are the time components and $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathbb{R}^d$ are the spatial components. Their Lorentz inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = -x_0 y_0 + \tilde{\mathbf{x}}^\top \tilde{\mathbf{y}}. \tag{1}$$

The Lorentz model is defined as

$$\mathbb{H}^d_c = \left\{ \mathbf{x} \in \mathbb{R}^{d+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -\frac{1}{c},\ x_0 > 0 \right\}, \tag{2}$$

where $x_0 > 0$ selects the upper sheet. Its origin is $O = [1/\sqrt{c}; \mathbf{0}_d]$, and the Lorentz distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{H}^d_c$ is

$$d_c(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{c}} \operatorname{arcosh}(-c \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}). \tag{3}$$

Following prior hyperbolic vision–language models (Desai et al., 2023; Pal et al., 2025), we treat the projected encoder features $\mathbf{z} \in \mathbb{R}^d$ as tangent vectors $[0; \mathbf{z}] \in T_O \mathbb{H}^d_c$ at the origin. The exponential map converts these tangent vectors into hyperbolic embeddings:

$$\mathbf{x} = \exp^c_O([0; \mathbf{z}]) = \frac{1}{\sqrt{c}} \left[\cosh(\sqrt{c}\|\mathbf{z}\|);\ \sinh(\sqrt{c}\|\mathbf{z}\|) \frac{\mathbf{z}}{\|\mathbf{z}\|} \right], \tag{4}$$

with the zero vector mapped to the origin, $\exp^c_O(\mathbf{0}_{d+1}) = O$.

For a non-origin embedding $\mathbf{x}$, we use the radial coordinate $\rho_x = \sqrt{c}\, d_c(O, \mathbf{x})$ and the unit spatial direction $\mathbf{u}_x = \tilde{\mathbf{x}}/\|\tilde{\mathbf{x}}\|$. At fixed curvature, larger ρ_x corresponds to greater distance from O. Their connection to the Lorentz coordinates is provided in Appendix A.1.

3 Compatibility of Alignment and Entailment

Distance alignment couples direction and radius. For non-origin embeddings $p, q \in \mathbb{H}^d_c$, let $\rho_x = \sqrt{c}\, d_c(O, x)$ denote radial depth and $\theta_{pq} = \angle pOq$ their angular separation at the origin. Writing $\delta_{pq} = \sqrt{c}\, d_c(p, q)$, the hyperbolic law of cosines gives

$$\begin{aligned} \cosh \delta_{pq} &= \cosh \rho_p \cosh \rho_q - \sinh \rho_p \sinh \rho_q \cos \theta_{pq} \\ &= \underbrace{\cosh(\rho_q - \rho_p)}_{\text{radial-gap contribution}} + \underbrace{\sinh \rho_p \sinh \rho_q (1 - \cos \theta_{pq})}_{\text{radius-weighted angular contribution}}. \end{aligned} \tag{5}$$

The first term depends only on the radial gap, and the second depends on angular separation and is weighted by both radii. Distance alignment therefore acts on radial depth as well as semantic direction. For a parent–child pair with $\rho_q > \rho_p$, holding the parent and both directions fixed, decreasing the child's radius toward the parent's decreases their distance. Even when $\theta_{pq} = 0$, the distance remains $(\rho_q - \rho_p)/\sqrt{c}$. Thus, distance minimization continues to penalize the radial separation that encodes different hierarchical levels. The derivation is given in Appendix B.1.

Conventional cones depend on more than pairwise distance. Conventional hyperbolic entailment cones admit descendants within an angular region extending outward from the parent (Ganea et al., 2018; Desai et al., 2023). Their membership is not determined by distance: two distinct embeddings at the same radial depth can be arbitrarily close yet remain outside each other's cones, whereas a child on the parent's outward radial ray remains inside its cone at arbitrarily large distances. Consequently, bringing a matched pair closer does not ensure entailment, and pushing an unrelated pair farther apart does not ensure its exclusion from the descendant region. With positive-only entailment supervision, such erroneous relations receive no direct entailment penalty. Cone selectivity also depends on the parent's absolute depth: for a fixed cone-width setting, the half-aperture satisfies $\sin \omega(p) \propto 1/\sinh \rho_p$. As the parent moves inward toward its admissible lower radius, $\omega(p)$ approaches $\pi/2$, admitting the entire outward hemisphere of geodesic directions at the boundary. Parent contraction can therefore weaken angular discrimination. The distance counterexamples and aperture analysis are provided in Appendix B.1.

Together, these observations motivate semantic alignment based on direction alone, an entailment relation defined through parent–child radial gaps and angular differences, and direct entailment supervision for both positive and negative pairs.

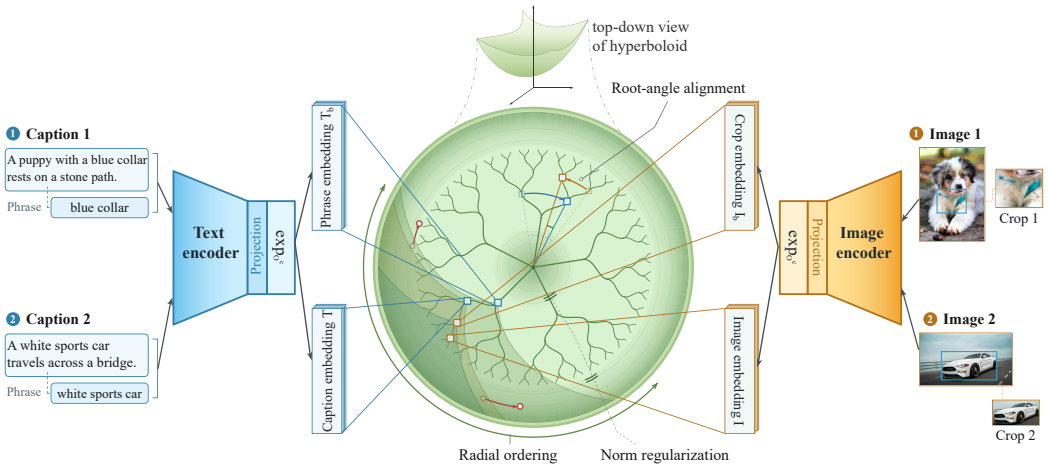

Figure 2: **Overview of AURORA.** Root-angle alignment brings related visual and textual semantics together, while HEC training via RAOC organizes concepts radially from general to specific.

4 AURORA

4.1 FORMULATION

Following the training setup of HyCoCLIP (Pal et al., 2025), we use image–caption pairs (I, T) with object-level annotations. A bounding box in I specifies an object crop I^b, and the corresponding phrase T^b describes its content. Depicted in Figure 2, AURORA uses a visual encoder f_v and a text encoder f_t, followed by projections P_v and P_t, respectively. Images and crops share the visual branch, while captions and phrases share the text branch. For an input $X \in \{I, T, I^b, T^b\}$, let f and P denote the corresponding encoder and projector. We obtain the projected feature $z_X = P(f(X))$ and its hyperbolic representation $x_X = \exp_O^c(0, z_X)$.

Hierarchy supervision uses the relations types $\mathcal{R} = \{T \to I,\ T^b \to I^b,\ I^b \to I,\ T^b \to T\}$, with arrows pointing from general to specific. Text is treated as more general than its visual counterpart, which contains additional instance details. Likewise, crops and phrases are treated as more general than full images and captions, which add scene context. We align matching visual and textual representations through root-angle alignment (Sec. 4.2) and learn the hierarchical relations in $\mathcal{R}$ through RAOC and HEC training (Secs. 4.3–4.4). Norm regularization controls their radial scale.

4.2 ROOT-ANGLE ALIGNMENT

For two non-origin hyperbolic representations $x, y \in \mathbb{H}_c^d \setminus \{O\}$ from either modality, let $\theta_{xy} = \angle xOy$ be their root angle. Write z_x and z_y for the projected features that generate x and y through the origin exponential map. This map preserves their directions, giving $u_x = z_x/\|z_x\|$ and $u_y = z_y/\|z_y\|$. We measure semantic similarity by $s_\angle(x, y) = \cos\theta_{xy} = u_x^\top u_y$. This score depends only on direction, leaving radial differences available for hierarchical ordering.

We optimize $s_\angle$ with a pairwise sigmoid loss. For embedding batches $A = \{a_i\}_{i=1}^N$ and $B = \{b_j\}_{j=1}^N$, set $y_{ij} = +1$ for matching pairs $(i = j)$ and $y_{ij} = -1$ otherwise. With $\sigma(t) = (1+e^{-t})^{-1}$,

$$\mathcal{L}_\sigma(A, B) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\sigma(y_{ij}\left[a_m s_\angle(a_i, b_j) + b_m\right]), \tag{6}$$

where $a_m > 0$ and b_m are a learned scale and bias. The complete alignment objective is

$$\mathcal{L}_{\text{align}} = \frac{1}{4}\left[\mathcal{L}_\sigma(I, T) + \mathcal{L}_\sigma(I^b, T^b) + \mathcal{L}_\sigma(I^b, T) + \mathcal{L}_\sigma(T^b, I)\right]. \tag{7}$$

The first two terms align corresponding visual and textual content, while crop–caption and phrase–image matching connect object-level semantics to the full scene. All similarity-based downstream evaluations use $s_\angle$.

4.3 RADIAL–ANGULAR ORDER CONE

Root-angle alignment measures semantic agreement, while hierarchical entailment additionally requires an order from general to specific. For a candidate parent p and descendant q in $\mathbb{H}^d_c \setminus \{O\}$, let $\rho_x = \sqrt{c}\, d_c(O, x)$ denote scaled radial depth. We define the signed radial gap $g(p, q) = \rho_q - \rho_p$ and root angle $\theta_{pq} = \angle pOq$. The *Radial–Angular Order Cone* (RAOC) defines the admissible descendants of p:

$$\mathcal{C}_\beta(p) = \left\{ q \in \mathbb{H}^d_c \setminus \{O\} : g(p, q) \geq \frac{\theta_{pq}}{\beta} \right\}, \qquad \beta > 0. \tag{8}$$

The required radial gap grows with angular separation: a larger directional difference requires the child to lie farther outward from its parent (Figure 3). The fixed hyperparameter β controls the angular allowance at each gap, with larger values giving a looser constraint. RAOC therefore ties angular selectivity to the relative parent–child radial gap rather than the absolute parent radius. It defines a partial order with nested descendant regions, so a child's descendants are also descendants of its parent. The formal statement and proof are given in Appendix C.

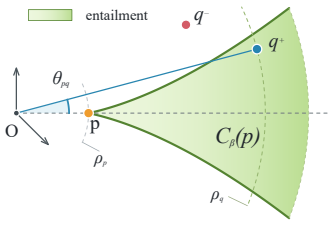

Figure 3: **RAOC.** Greater angular separation requires a larger radial gap.

For optimization, we formulate the RAOC constraint in terms of two bounded margins. The radial margin measures radial precedence, while the angular margin measures the admissible angular allowance minus the observed angle. We define

$$m_r(p, q) = \mathrm{clip}\left(\frac{g(p, q)}{\sigma_r}, -1, 1 \right), \qquad m_a(p, q) = \frac{\mathrm{clip}(\beta g(p, q), 0, \pi) - \theta_{pq}}{\pi}. \tag{9}$$

Here $\sigma_r > 0$ sets the radial margin scale, and $\mathrm{clip}(t, a, b) = \min\{b, \max\{a, t\}\}$. Both margins lie in $[-1, 1]$. For negative gaps, the angular allowance is clipped to zero, while $m_r < 0$ identifies the incorrect radial order even when $\theta_{pq} = 0$.

The entailment score is the smaller of the two margins:

$$D_{\mathrm{RAOC}}(p, q) = \min\{m_r(p, q), m_a(p, q)\}. \tag{10}$$

Thus, $D_{\mathrm{RAOC}}(p, q) \geq 0$ requires both radial precedence and angular compatibility and is equivalent to $q \in \mathcal{C}_\beta(p)$. Further details are provided in Appendix D.2.

4.4 HIERARCHY LEARNING VIA CONTRASTIVE ENTAILMENT

Hierarchical Entailment Contrastive (HEC) training reinforces annotated parent–child relations and suppresses spurious relations between nonmatching examples. For each relation type $U \to V \in \mathcal{R}$, x_{U_i} and x_{V_i} form a positive pair, while x_{U_i} and x_{V_j} with $i \neq j$ form a negative pair. Using the labels y_{ij} defined in Sec. 4.2, we optimize

$$\mathcal{L}_{\mathrm{HEC}} = -\frac{1}{|\mathcal{R}|N} \sum_{U \to V \in \mathcal{R}} \sum_{i,j=1}^{N} \log \sigma\big(y_{ij} \left[a_e D_{\mathrm{RAOC}}(x_{U_i}, x_{V_j}) + b_e \right] \big), \tag{11}$$

where $a_e > 0$ and b_e are a learned scale and bias. This objective increases entailment compatibility for matching parent–descendant pairs and decreases it for nonmatching pairs.

Angular alignment leaves radial position unconstrained, while RAOC organizes representations through relative radial gaps. We stabilize their radial scale with the norm penalty $\ell_{\mathrm{norm}}(x) = \widetilde{r}_x^2 - \log \widetilde{r}_x$, where $\widetilde{r}_x = \max\{\|x_s\|, \epsilon\}$ and $\epsilon > 0$ keeps the logarithm finite. The quadratic and negative logarithmic terms penalize large and small spatial norms, respectively. We define $\mathcal{L}_{\mathrm{norm}}$ as the average of this penalty over all image, caption, crop, and phrase embeddings in the minibatch. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{align}} + \lambda_e \mathcal{L}_{\mathrm{HEC}} + \lambda_n \mathcal{L}_{\mathrm{norm}}, \tag{12}$$

where λ_e and λ_n weight entailment learning and norm regularization. Implementation details are provided in Appendix F. Further details on hyperparameter selection and sensitivity analysis are provided in Appendix J.

5 EXPERIMENTS

5.1 TRAINING DETAILS

We pretrain AURORA on GRIT (Peng et al., 2024), comprising 10.3M valid image–caption pairs and 17.8M grounded box–phrase annotations. We mainly report results using ViT-S/16, with complete comparisons for both ViT-S/16 and ViT-B/16 provided in Appendix G. Training schedules, hyperparameters, and implementation details are given in Appendix F. We compare AURORA with CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), MERU (Desai et al., 2023), ATMG (Ramasinghe et al., 2024), HyCoCLIP (Pal et al., 2025), PHyCLIP (Yoshikawa & Matsubara, 2026), and ARGENT (Huynh et al., 2026). For CLIP, SigLIP, and MERU, we include variants pretrained with and without box–phrase pairs to assess the contribution of additional grounded supervision. All comparisons are conducted under the same task-specific evaluation protocols. AURORA uses root-angle similarity $s_{\angle}$ for all similarity-based evaluations.

5.2 EXPERIMENTAL RESULTS

Zero-shot image classification. We evaluate zero-shot transfer on 16 datasets covering general, fine-grained, and specialized recognition tasks. As shown in Table 1, AURORA achieves the highest accuracy on 13 datasets. Its advantage extends to comparisons with box-supervised CLIP, SigLIP, and MERU, suggesting that the improvement cannot be attributed solely to the inclusion of additional grounded training pairs.

Table 1: **Zero-shot image classification on ViT-S/16.** Results are reported as top-1 accuracy (%). **Bold** indicates the best results. Underlining denotes the second best.

ViT-S/16		General datasets						Fine-grained datasets						Specialized datasets			
Model	w/ boxes	ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Food-101	CUB	Cars	Aircraft	Pets	Flowers	DTD	EuroSAT	RESISC45	Country211
CLIP	✗	32.66	74.19	42.22	42.91	69.80	90.09	42.72	12.55	3.93	2.06	43.22	18.03	17.57	25.47	33.64	3.88
CLIP	✓	32.49	73.69	42.15	42.66	70.07	89.83	42.50	12.48	4.01	2.10	43.28	17.79	17.61	26.04	33.37	3.84
SigLIP	✗	34.30	66.57	38.38	44.73	71.42	91.75	38.98	13.33	4.89	2.13	46.65	23.17	18.40	**32.63**	35.42	4.15
SigLIP	✓	34.34	67.07	38.50	44.82	71.55	91.82	39.27	13.33	4.92	2.24	46.56	23.03	18.36	31.42	35.37	4.23
MERU	✗	31.74	76.08	41.24	41.28	67.76	87.24	41.09	11.97	3.80	2.88	42.53	16.75	17.46	27.75	33.88	3.82
MERU	✓	32.01	76.55	41.35	41.21	67.64	87.30	41.24	12.08	3.78	**2.95**	42.54	16.57	17.52	27.53	33.82	3.83
ATMG	✗	33.74	77.21	43.28	43.15	68.08	90.84	44.01	13.27	4.38	1.75	42.53	15.70	17.89	28.69	35.75	4.21
HyCoCLIP	✓	36.38	83.28	47.91	47.88	71.53	93.16	46.73	15.33	5.98	2.45	45.59	20.89	19.47	26.00	38.73	4.09
PHyCLIP	✓	37.06	83.44	48.41	48.88	71.82	93.06	47.26	16.29	6.58	2.57	46.44	21.46	19.13	25.81	39.79	3.95
ARGENT	✓	37.55	83.55	48.62	49.29	72.15	**94.01**	47.70	16.90	6.81	2.48	47.10	21.65	19.67	26.32	40.46	4.19
AURORA	✓	**41.54**	**84.25**	**51.20**	**50.27**	**75.81**	93.19	**50.13**	**17.71**	**8.91**	1.80	**51.41**	**28.99**	**22.87**	30.18	**41.09**	**4.70**

Image–text retrieval and hierarchical classification. We evaluate bidirectional image–text retrieval on COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014), together with hierarchical classification on ImageNet (Deng et al., 2009). The latter uses the WordNet taxonomy to assess how closely predictions agree with ground-truth labels in the hierarchy. Table 2 shows that AURORA achieves the best results in both retrieval directions and across all reported hierarchical metrics. Thus, improved cross-modal matching is accompanied by stronger agreement with the label hierarchy. We follow the evaluation protocols of HyCoCLIP and PHyCLIP, with metric definitions detailed in Appendix H.

Compositional understanding. We evaluate compositional understanding on VL-CheckList-Object (Zhao et al., 2022) and SugarCrepe (Hsieh et al., 2023), where models must distinguish a correct caption from a hard negative containing altered semantic content. VL-CheckList-Object tests sensitivity to object replacement across different object locations and sizes. SugarCrepe extends the evaluation to seven types of perturbations involving objects, attributes, and relations.

As shown in Table 3, AURORA leads on nine of the fourteen reported measures, including the overall SugarCrepe score. The most pronounced gains occur on VL-CheckList-Object, suggesting more reliable identification of objects across variations in location and scale. The best overall per-

formance on SugarCrepe further indicates improved discrimination of compositional mismatches beyond object replacement.

Table 2: **Zero-shot retrieval and hierarchical classification on ViT-S/16.** Retrieval performance is reported as Recall@5/10 (%). Hierarchical metrics are shown to two decimal places, with rankings determined before rounding. **Bold** indicates the best results. Underlining denotes the second best.

ViT-S/16		Text → Image				Image → Text				Hierarchical Classification				
Model	w/ boxes	COCO		Flickr		COCO		Flickr		WordNet				
		R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	TIE↓	LCA↓	$J\uparrow$	$P_H\uparrow$	$R_H\uparrow$
CLIP	✗	50.31	61.52	76.98	84.48	65.75	76.15	87.10	93.00	4.43	2.51	0.72	0.79	0.80
CLIP	✓	50.27	61.47	76.75	84.42	65.49	76.08	87.20	93.17	4.42	2.50	0.72	0.78	0.80
SigLIP	✗	52.32	63.90	77.54	85.24	65.98	76.78	88.60	93.23	4.22	2.47	0.75	0.83	0.82
SigLIP	✓	52.27	63.81	77.21	84.09	66.03	76.84	88.60	93.40	4.23	2.47	0.74	0.83	0.83
MERU	✗	49.79	61.14	76.03	84.34	65.01	75.67	86.80	92.77	4.44	2.52	0.71	0.78	0.79
MERU	✓	49.63	61.08	76.13	84.21	65.16	75.74	86.93	92.77	4.44	2.51	0.71	0.78	0.78
ATMG	✗	50.95	62.36	77.21	84.74	64.96	75.81	87.10	93.17	4.37	2.48	0.73	0.82	0.81
HyCoCLIP	✓	51.35	63.13	77.92	85.52	64.84	75.78	87.30	93.23	3.97	2.33	0.76	0.84	0.84
PHyCLIP	✓	52.24	64.08	78.26	85.84	65.65	76.21	87.20	93.17	3.92	2.31	0.76	0.84	0.84
ARGENT	✓	52.68	64.46	78.59	86.18	65.75	76.46	87.50	93.23	3.91	2.30	0.77	0.84	0.84
AURORA	✓	**55.77**	**66.81**	**81.02**	**88.18**	**69.42**	**79.18**	**89.90**	**95.13**	**3.62**	**2.25**	**0.78**	**0.85**	**0.85**

Table 3: **Hard-negative compositional classification on ViT-S/16.** Models are evaluated on VL-CheckList-Object and SugarCrepe. **Bold** indicates the best results. Underlining denotes the second best.

ViT-S/16		VL-CheckList-Object						SugarCrepe							
Model	w/ boxes	Location			Size			Replace			Swap		Add		
		Center	Mid	Margin	Large	Medium	Small	Obj	Att	Rel	Obj	Att	Obj	Att	Overall
CLIP	✗	85.10	79.35	71.22	84.48	74.61	65.72	86.74	78.00	69.04	59.49	65.27	77.53	70.76	75.71
CLIP	✓	85.24	78.41	70.55	85.32	75.02	64.67	85.94	77.45	**69.94**	59.89	64.41	77.90	70.76	75.69
SigLIP	✗	89.33	83.04	74.91	88.30	78.50	68.82	87.35	78.55	69.49	60.82	65.77	78.08	71.10	76.26
SigLIP	✓	89.83	83.03	74.50	88.13	78.13	69.47	87.01	78.43	69.18	**61.38**	**66.67**	78.03	70.28	76.12
MERU	✗	82.07	76.10	67.93	81.47	71.50	62.40	86.40	77.58	68.73	58.94	64.96	77.17	70.33	75.35
MERU	✓	82.90	75.40	68.40	81.33	71.73	62.40	87.09	77.03	69.11	58.81	65.17	77.19	70.47	75.55
ATMG	✗	84.07	78.40	69.83	83.53	73.60	64.53	87.11	78.34	69.35	60.16	65.62	77.98	71.15	76.10
HyCoCLIP	✓	90.80	86.75	79.95	90.52	82.73	74.23	87.23	78.81	66.64	60.41	64.86	79.44	69.08	75.82
PHyCLIP	✓	91.40	87.14	80.42	90.97	83.06	74.79	88.40	80.25	67.33	59.76	65.22	79.89	71.48	76.71
ARGENT	✓	91.87	87.73	**81.10**	**91.33**	84.07	75.60	88.92	80.88	67.71	59.89	65.32	80.00	72.59	77.11
AURORA	✓	**91.93**	**87.80**	80.77	91.30	**84.69**	**76.53**	**89.77**	**81.85**	68.42	60.00	65.47	**80.16**	**74.27**	**77.75**

5.3 Geometric Visualizations

We inspect AURORA's learned geometry from two complementary perspectives: spatial-norm distributions reveal radial allocation across four representation types, while HoroPCA (Chami et al., 2021) provides a low-dimensional view of their organization. Figure 4 compares AURORA ViT-S with its w/o cone variant. Without RAOC, full inputs and their corresponding boxes have largely overlapping norm distributions within each modality. RAOC produces more distinct distributions, particularly for texts and text boxes, indicating stronger radial differentiation. In the HoroPCA projections, the w/o cone variant exhibits a narrow text configuration and a compact visual cluster. With RAOC, both modalities occupy broader, more structured regions, with text and image boxes tending to lie closer to the projected origin than their full-input counterparts. This projected arrangement is consistent with the intended ordering from general concepts to contextualized instances.

We further examine local hierarchical organization through radial gaps and root angles. Figure 5 compares AURORA ViT-S with and without RAOC for the text query "bag." With RAOC, the image-region tree forms distinct branches: root angles decrease along the path from the full image through "mat bag" to "bag," while the umbrella–chair branch remains less aligned with the query. Without RAOC, the branches are less clearly distinguished. These observations suggest that RAOC helps capture local tree-like semantic structure.

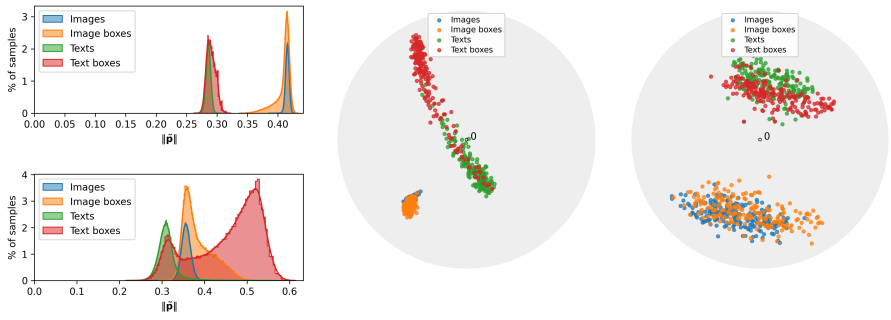

Figure 4: **Representation geometry with and without RAOC.** (a) Spatial-norm distributions and HoroPCA projections (b) without and (c) with RAOC for AURORA ViT-S. RAOC yields more distinct norm distributions and broader projected organization, with boxes tending to lie nearer the projected origin than full inputs.

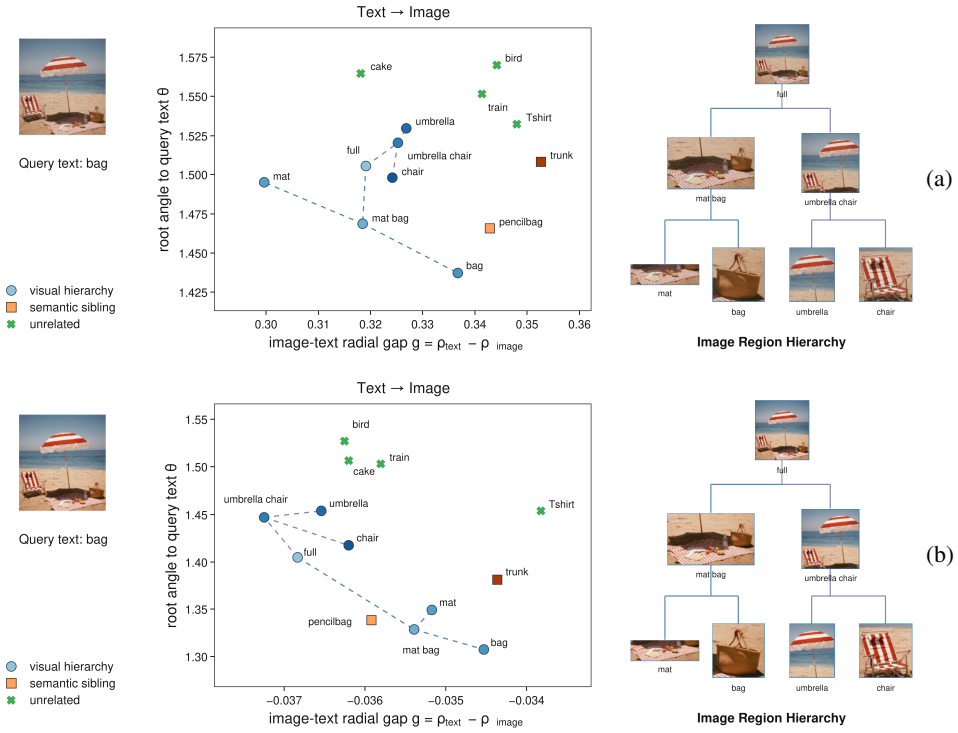

Figure 5: **Local hierarchical organization with and without RAOC.** AURORA ViT-S (a) w/o cone and (b) w/ cone for the text query "bag." Scatter plots show radial gaps and root angles for visual regions, semantic siblings, and unrelated controls; the corresponding visual hierarchy is shown on the right. RAOC produces clearer query-relative organization along the relevant hierarchy branch.

5.4 Ablation Study

Table 4 isolates the main components of AURORA. Replacing RAOC with conventional Entailment Cone (EC) compares alternative order regions; removing the norm regularizer tests radial-scale control; removing RAOC isolates the effect of directed hierarchy supervision; and removing both RAOC and norm regularization leaves the root-angle alignment alone. Replacing the pairwise sigmoid loss with InfoNCE further examines the contrastive loss term under different alignment scores. With $-d_{\mathbb{H}}$, positive pairs are still encouraged toward smaller absolute distance, retaining the radial pressure discussed in Sec. 3. Appendix D.1 provides the corresponding gradient analysis.

Table 4: **Ablation study of AURORA-ViT-S/16 pre-trained on grounded GRIT,** evaluated on image classification, COCO retrieval, and hierarchical classification. Controlled variants isolate the contribution of individual training components and loss terms.

AURORA-ViT-S/16					Classification		COCO R@5		Hierarchy	
ID	**Alignment score**	**Order objective**	**Norm**	**Loss**	ImageNet	Food-101	T→I	I→T	TIE↓	J ↑
(a)	root angle	EC	✓	Sigmoid	38.12	45.92	53.48	66.72	3.79	0.76
(b)	root angle	RAOC	✗	Sigmoid	40.21	48.57	54.98	68.41	<u>3.70</u>	<u>0.77</u>
(c)	root angle	RAOC	✓	InfoNCE	39.42	47.71	54.43	67.93	3.76	0.76
(d)	root angle	none	✓	Sigmoid	<u>40.27</u>	47.06	<u>55.29</u>	<u>68.92</u>	3.74	<u>0.77</u>
(e)	root angle	none	✗	Sigmoid	38.91	45.38	54.56	67.88	3.85	0.76
(f)	$-d_{\mathbb{H}}$	EC	✓	Sigmoid	37.32	44.63	52.66	65.84	3.85	0.75
(g)	$-d_{\mathbb{H}}$	EC	✓	InfoNCE	36.05	43.24	51.79	64.87	3.94	0.74
Full	root angle	RAOC	✓	Sigmoid	**41.54**	**50.13**	**55.77**	**69.42**	**3.62**	**0.78**

6 Related Work

Vision and language representation learning. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn transferable visual and textual representations through contrastive learning on image and text pairs. SigLIP (Zhai et al., 2023) replaces softmax normalization with independent sigmoid losses over pairs. RegionCLIP (Zhong et al., 2022) extends alignment to image regions and textual concepts. AURORA retains separate image and text encoders and examines how semantic alignment interacts with hierarchical supervision.

Hyperbolic vision and language representations. Hyperbolic embeddings capture hierarchical structure in language and vision (Nickel & Kiela, 2017; Khrulkov et al., 2020). MERU combines Lorentz representations with entailment constraints (Desai et al., 2023). HyCoCLIP incorporates object regions and grounded phrases through compositional contrastive and entailment learning (Pal et al., 2025). PHyCLIP uses an ℓ_1 product of hyperbolic factors to model hierarchy and composition (Yoshikawa & Matsubara, 2026). Wu et al. (Wu et al., 2026) align visual and textual trees across hyperbolic manifolds with different curvatures. ATMG addresses the tension between distance alignment and entailment through an exterior angle contrastive objective (Ramasinghe et al., 2024). AURORA separates root angle alignment from radial ordering to preserve differences in specificity.

Hierarchical representations and entailment regions. Hierarchical embeddings encode asymmetric relations through coordinatewise partial orders (Vendrov et al., 2016). Hyperbolic entailment cones define nested descendant regions (Ganea et al., 2018) and support concept hierarchy induction from text (Le et al., 2019). Shadow Cones (Yu et al., 2024) generalizes entailment regions through umbral and penumbral constructions based on geometric occlusion. ARGENT (Huynh et al., 2026) combines adaptive entailment supervision with norm regularization to stabilize hierarchical representations. AURORA defines descendant membership through relative radial separation and root angle difference, with contrastive supervision over matched and unmatched pairs.

7 Conclusion

We introduced AURORA, a hyperbolic vision–language pretraining framework that separates angular semantic alignment from radial hierarchical ordering. Our analysis shows that distance minimization can conflict with hierarchical separation, while semantic discrimination alone does not ensure valid entailment. AURORA combines angular alignment with the Radial–Angular Order Cone, which defines a partial order through radial gaps and angular differences. Hierarchical entailment contrastive training reinforces annotated relations and suppresses spurious entailment, while norm regularization stabilizes training. Experiments demonstrate improvements over competitive Euclidean and hyperbolic baselines across classification, retrieval, hierarchical classification, and compositional understanding. Future work could discover richer hierarchical relations within and across image–text examples and explore how to learn from them with fewer explicit annotations.

AI USE STATEMENT

Generative AI tools were used for literature retrieval, language editing, LaTeX assistance, and discussions of research ideas and experimental design. They were also used as a supporting tool during the verification and organization of mathematical derivations. All formulations, implementations, experiments, proofs, and scientific claims were developed, verified, and validated by the authors, who take full responsibility for the final content of this manuscript.

REPRODUCIBILITY STATEMENT

The main paper presents the objectives, methodology, and key experimental findings. Appendices F, H, and related supplementary sections provide implementation details, training configurations, evaluation protocols, mathematical derivations, and computational settings. We will release code, configurations, data processing details, evaluation resources, and experiment outputs to facilitate reproduction.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29.

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

Ines Chami, Albert Gu, Dat P. Nguyen, and Christopher Ré. HoroPCA: Hyperbolic dimensionality reduction via horospherical projections. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1419–1429. PMLR, 2021.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223. PMLR, 2011. URL https://proceedings.mlr.press/v15/coates11a.html.

Ofer Dekel, Joseph Keshet, and Yoram Singer. Large margin hierarchical classification. In *Proceedings of the 21st International Conference on Machine Learning*, 2004. URL https://www.microsoft.com/en-us/research/publication/large-margin-hierarchical-classification/.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7694–7731. PMLR, 2023.

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

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178, 2004. doi: 10.1109/CVPR.2004.383. URL https://cs.stanford.edu/groups/vision/documents/Fei-FeiFergusPerona_GMBV04.pdf.

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.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1355.

Chuong Huynh, Hossein Souri, Abhinav Kumar, Vitali Petsiuk, Deen Dayal Mohan, and Suren Kumar. ARGENT: Adaptive hierarchical image-text representations. *arXiv preprint arXiv:2603.23311*, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6418–6428, 2020.

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, June 2026a.

Hayeon Kim, Ji Ha Jang, Junghun James Kim, and Se Young Chun. UNCHA: Official code release. GitHub repository, 2026b. URL https://github.com/jeeit17/UNCHA/tree/b9aa27d085ec901285189c08ca7a1296b750874d. Commit b9aa27d085ec901285189c08ca7a1296b750874d; accessed 2026-09-25.

Jaeyoung Kim, Eunseok Kim, and Dongsuk Jang. Is the geometry doing the work? an operating-point audit of hierarchy in hyperbolic vision-language models. *arXiv preprint arXiv:2607.05268*, 2026c.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015. doi: 10.1007/s10618-014-0382-x.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

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

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.

Matthew Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3231–3241, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1313. URL https://aclanthology.org/P19-1313/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6338–6347, 2017.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3779–3788. PMLR, 2018.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://www.robots.ox.ac.uk/~vgg/publications/2008/Nilsback08/.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27263–27272, June 2024.

John G. Ratcliffe. *Foundations of Hyperbolic Manifolds*, volume 149 of *Graduate Texts in Mathematics*. Springer, 1994. doi: 10.1007/978-1-4757-4013-4.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4460–4469. PMLR, 2018.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *Graph Drawing*, volume 7034 of *Lecture Notes in Computer Science*, pp. 355–366. Springer, 2012. doi: 10.1007/978-3-642-25878-7_34.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations*, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Wei Wu, Xiaomeng Fan, Yuwei Wu, Zhi Gao, Pengxiang Li, Yunde Jia, and Mehrtash Harandi. Modality alignment across trees on heterogeneous hyperbolic manifolds. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1370f2bf31dc0ecaeac17042193a50a1-Abstract-Conference.html.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Daiki Yoshikawa and Takashi Matsubara. PHyCLIP: ℓ_1-product of hyperbolic factors unifies hierarchy and compositionality in vision-language representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Tao Yu, Toni J. B. Liu, Albert Tseng, and Christopher De Sa. Shadow cones: A generalized framework for partial order embeddings. In *The Twelfth International Conference on Learning Representations*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 30–37, Abu Dhabi, UAE, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.4.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

A NOTATION AND EQUIVALENT GEOMETRIC VIEWS

Table 5: Core notation used in the geometric analysis.

Symbol	Meaning	Role
c	positive curvature magnitude; manifold curvature is $-c$	learned global scalar
$\tilde{\mathbf{x}} \in \mathbb{R}^d$	spatial component of a Lorentz point $\mathbf{x}$	Lorentz coordinate
$r_x = \|\tilde{\mathbf{x}}\|$	Lorentz spatial norm	radial regularization variable
$\rho_x = \sqrt{c}\, d_c(O, \mathbf{x})$	scaled distance from the origin	radial order coordinate
$\mathbf{u}_x = \tilde{\mathbf{x}}/\|\tilde{\mathbf{x}}\|$	spatial direction of non-origin $\mathbf{x}$	angular coordinate
$\mathbf{w}_x = \Phi_c(\mathbf{x})$	Poincaré coordinate of $\mathbf{x}$	equivalent coordinate view
$\theta_{xy} = \angle \mathbf{x}O\mathbf{y}$	root angle	directional difference
$g(\mathbf{p}, \mathbf{q}) = \rho_q - \rho_p$	parent–candidate radial gap	radial precedence
$f_\beta(g) = \min\{\pi, \beta g\}$	admissible angle for $g \geq 0$	RAOC aperture
m_r, m_a	bounded radial and angular margins	optimization margins
$D_{\mathrm{RAOC}} = \min\{m_r, m_a\}$	bounded entailment compatibility	training score
β	radial-gap-to-angle slope	order geometry
σ_r	radial margin scale	score scaling
a_e, b_e	entailment-logit scale and bias	score calibration

A.1 RADIAL–ANGULAR COORDINATES AND ROOT-ANGLE IDENTITIES

Throughout the appendix we use the radial–angular representation induced by the Lorentz origin. We write Lorentz vectors as $\mathbf{x} = [x_0; \tilde{\mathbf{x}}] \in \mathbb{R}^{d+1}$, where $x_0 \in \mathbb{R}$ and $\tilde{\mathbf{x}} \in \mathbb{R}^d$; the semicolon denotes vertical concatenation. For any non-origin $\mathbf{x} \in \mathbb{H}^d_c$, the manifold constraint together with $\rho_x = \sqrt{c}\, d_c(O, \mathbf{x})$ gives

$$\mathbf{x} = \frac{1}{\sqrt{c}} \big[\cosh \rho_x;\, \sinh \rho_x\, \mathbf{u}_x\big], \qquad \mathbf{u}_x = \frac{\tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|} \in \mathbb{S}^{d-1}. \tag{13}$$

Equivalently,

$$r_x = \frac{\sinh \rho_x}{\sqrt{c}}, \qquad \rho_x = \operatorname{asinh}(\sqrt{c}\, r_x). \tag{14}$$

For the origin-exponential parameterization in Eq. (4), this simplifies to $\rho_x = \sqrt{c}\, \|\mathbf{z}_x\|$ and $\mathbf{u}_x = \mathbf{z}_x/\|\mathbf{z}_x\|$ whenever $\mathbf{z}_x \neq \mathbf{0}_d$. Consequently, for non-origin $\mathbf{x}$ and $\mathbf{y}$,

$$\cos \theta_{xy} = \mathbf{u}_x^\top \mathbf{u}_y = \frac{\tilde{\mathbf{x}}^\top \tilde{\mathbf{y}}}{\|\tilde{\mathbf{x}}\|\, \|\tilde{\mathbf{y}}\|} = \frac{\cosh \rho_x \cosh \rho_y - \cosh \delta_{xy}}{\sinh \rho_x \sinh \rho_y}, \qquad \delta_{xy} = \sqrt{c}\, d_c(\mathbf{x}, \mathbf{y}). \tag{15}$$

The first equality gives the score used by AURORA, and the last follows directly from the hyperbolic cosine law. The resulting semantic score is invariant to radial position and varies with root direction.

A.2 POINCARÉ REALIZATION OF ROOT-ANGLE ALIGNMENT AND RAOC

A natural question is how the same geometry appears when the embeddings are parameterized directly in the Poincaré ball. For curvature $-c$,

$$\mathbb{B}^d_c = \{\mathbf{w} \in \mathbb{R}^d : c\|\mathbf{w}\|^2 < 1\}, \qquad g^{\mathbb{B}}_w = \left(\frac{2}{1 - c\|\mathbf{w}\|^2}\right)^2 I. \tag{16}$$

The metric is conformal to the Euclidean metric, so Riemannian angles and Euclidean coordinate angles agree at each ball point (Nickel & Kiela, 2017; Ganea et al., 2018). The origin-preserving Lorentz–Poincaré isometry is

$$\Phi_c(\mathbf{x}) = \frac{\tilde{\mathbf{x}}}{\sqrt{c}\, x_0 + 1}. \tag{17}$$

Substituting Eq. (13) gives

$$\mathbf{w}_x = \Phi_c(\mathbf{x}) = \frac{1}{\sqrt{c}} \tanh\Big(\frac{\rho_x}{2}\Big)\, \mathbf{u}_x, \qquad \rho_x = 2\, \operatorname{artanh}(\sqrt{c}\, \|\mathbf{w}_x\|), \qquad \frac{\mathbf{w}_x}{\|\mathbf{w}_x\|} = \mathbf{u}_x. \tag{18}$$

Root-angle alignment. The root angle is the intrinsic angle between the two radial geodesics leaving the origin. The isometry Φ_c preserves this angle and maps the Lorentz origin O to the ball origin 0. In the Poincaré ball, radial geodesics through 0 are Euclidean radial segments; conformality then identifies their intrinsic angle with the ordinary Euclidean polar angle. For corresponding non-origin points $\mathbf{w}_x = \Phi_c(\mathbf{x})$ and $\mathbf{w}_y = \Phi_c(\mathbf{y})$,

$$\begin{aligned} \theta^{\mathbb{L}}_{xy} &= \angle^{\mathbb{L}}_O(\mathbf{x}, \mathbf{y}) = \angle^{\mathbb{B}}_0(\mathbf{w}_x, \mathbf{w}_y) \\ &= \arccos\left(\frac{\mathbf{w}_x^\top \mathbf{w}_y}{\|\mathbf{w}_x\| \, \|\mathbf{w}_y\|} \right), \end{aligned} \tag{19}$$

and therefore

$$s^{\mathbb{L}}_{\angle}(\mathbf{x}, \mathbf{y}) = \frac{\tilde{\mathbf{x}}^\top \tilde{\mathbf{y}}}{\|\tilde{\mathbf{x}}\| \, \|\tilde{\mathbf{y}}\|} = \frac{\mathbf{w}_x^\top \mathbf{w}_y}{\|\mathbf{w}_x\| \, \|\mathbf{w}_y\|} = s^{\mathbb{B}}_{\angle}(\mathbf{w}_x, \mathbf{w}_y). \tag{20}$$

A Poincaré implementation can consequently optimize root-angle alignment with the Euclidean cosine between normalized ball coordinates. The Lorentz spatial-direction cosine and the Poincaré polar-angle cosine are two coordinate expressions of the same intrinsic root angle.

RAOC in Poincaré coordinates. The isometry also preserves distance from the origin. Define the scaled Poincaré radial depth

$$\rho_{\mathbb{B}}(\mathbf{w}) = \sqrt{c}\, d_{\mathbb{B},c}(0, \mathbf{w}) = 2 \operatorname{artanh}(\sqrt{c}\, \|\mathbf{w}\|). \tag{21}$$

Then $\rho_{\mathbb{B}}(\mathbf{w}_x) = \rho_x$, and Eq. (32) has the equivalent ball form

$$\begin{aligned} \mathbf{q} \in \mathcal{C}_\beta(\mathbf{p}) \quad \Longleftrightarrow \quad & \rho_{\mathbb{B}}(\mathbf{w}_q) - \rho_{\mathbb{B}}(\mathbf{w}_p) \geq 0, \\ & \theta_{\mathbb{B}}(\mathbf{w}_p, \mathbf{w}_q) \leq \beta \big[\rho_{\mathbb{B}}(\mathbf{w}_q) - \rho_{\mathbb{B}}(\mathbf{w}_p) \big], \end{aligned} \tag{22}$$

where $\theta_{\mathbb{B}}$ is the Euclidean polar angle from Eq. (19). The order is therefore expressed by the same radial gap and the same direction-sphere angle in either hyperbolic model. The partial-order and descendant-set nesting results below carry directly to the Poincaré representation.

For a fixed parent $\mathbf{w}_p$, Eq. (22) can also be written in the Euclidean ball radius:

$$\|\mathbf{w}_q\| \geq \frac{1}{\sqrt{c}} \tanh\left[\operatorname{artanh}(\sqrt{c}\, \|\mathbf{w}_p\|) + \frac{\theta_{\mathbb{B}}(\mathbf{w}_p, \mathbf{w}_q)}{2\beta} \right]. \tag{23}$$

The linear RAOC boundary in (g, θ) coordinates thus appears as a curved radial–angular contour when expressed with the Euclidean ball radius. The angular coordinate remains the Euclidean polar angle, while radial order is measured by the intrinsic depth $\rho_{\mathbb{B}}$.

When an encoder feature is inserted through the origin exponential map, the tangent-coordinate conventions differ by a constant scale. Since $g^{\mathbb{B}}_0 = 4I$,

$$(\mathrm{d}\Phi_c)_O([0; \mathbf{z}]) = \mathbf{z}/2, \qquad \Phi_c(\exp^c_O([0; \mathbf{z}])) = \frac{\tanh(\sqrt{c}\|\mathbf{z}\|/2)}{\sqrt{c}} \frac{\mathbf{z}}{\|\mathbf{z}\|}. \tag{24}$$

Thus the Lorentz tangent coordinate $\mathbf{z}$ and the Poincaré coordinate tangent vector $\mathbf{z}/2$ parameterize the corresponding manifold point with the same normalized direction. Coordinate-wise optimization trajectories can vary with parameterization and optimizer choice. At the geometric level, corresponding manifold configurations receive identical root-angle scores and RAOC membership decisions under Φ_c.

B GEOMETRIC ANALYSIS OF ALIGNMENT AND ENTAILMENT

B.1 DISTANCE ALIGNMENT AND CONVENTIONAL CONE GEOMETRY

Radial pressure from pairwise distance. Let $\delta_{pq} = \sqrt{c}\, d_c(\mathbf{p}, \mathbf{q})$. Starting from Eq. (5), fix $\rho_p > 0$ and θ_{pq}, with $\mathbf{p} \neq \mathbf{q}$ and $\rho_q \geq \rho_p$. Differentiating with respect to ρ_q gives

$$\sinh \delta_{pq} \frac{\partial \delta_{pq}}{\partial \rho_q} = \cosh \rho_p \cosh \rho_q \left(\tanh \rho_q - \tanh \rho_p \cos \theta_{pq}\right). \tag{25}$$

The factor in parentheses can be decomposed as

$$(\tanh \rho_q - \tanh \rho_p) + \tanh \rho_p (1 - \cos \theta_{pq}).$$

Both terms are nonnegative, and simultaneous equality would imply $\rho_q = \rho_p$ and $\theta_{pq} = 0$, hence $\mathbf{p} = \mathbf{q}$. Since $\delta_{pq} > 0$, it follows that $\partial \delta_{pq} / \partial \rho_q > 0$. Holding the parent and both directions fixed, decreasing the outward child's radius toward the parent's therefore decreases their distance.

Moreover, as their directions align,

$$\lim_{\theta_{pq} \to 0} d_c(\mathbf{p}, \mathbf{q}) = \frac{\rho_q - \rho_p}{\sqrt{c}}, \qquad \rho_q > \rho_p. \tag{26}$$

Thus, even at zero angular separation, distance minimization continues to penalize the parent–child radial gap.

Pairwise distance does not determine cone membership. Let $\mathcal{C}_{\mathrm{EC}}(\mathbf{p})$ denote an outward conventional Entailment Cone (EC) rooted at a non-origin parent $\mathbf{p}$. Its half-aperture $\omega(\mathbf{p})$ specifies the maximum permitted angle from the outward continuation of the radial geodesic $O\mathbf{p}$. For $\mathbf{q} \neq \mathbf{p}$, let $\varphi(\mathbf{p}, \mathbf{q})$ be the angle at $\mathbf{p}$ between this outward direction and the geodesic from $\mathbf{p}$ to $\mathbf{q}$. Then $\mathbf{q} \in \mathcal{C}_{\mathrm{EC}}(\mathbf{p})$ if and only if $\varphi(\mathbf{p}, \mathbf{q}) \leq \omega(\mathbf{p})$.

Proposition 1 (Distance and cone membership)**.** *Suppose $\mathcal{C}_{\mathrm{EC}}(\mathbf{p})$ has half-aperture $\omega(\mathbf{p}) \in (0, \pi/2]$. For distinct non-origin embeddings $\mathbf{p}$ and $\mathbf{q}$:*

1. *If $\rho_p = \rho_q$ and $\theta_{pq} > 0$, then $\varphi(\mathbf{p}, \mathbf{q}) > \pi/2$ and $\mathbf{q} \notin \mathcal{C}_{\mathrm{EC}}(\mathbf{p})$, even though $d_c(\mathbf{p}, \mathbf{q}) \to 0$ as $\theta_{pq} \to 0$.*
2. *If $\theta_{pq} = 0$ and $\rho_q > \rho_p$, then $\varphi(\mathbf{p}, \mathbf{q}) = 0$ and $\mathbf{q} \in \mathcal{C}_{\mathrm{EC}}(\mathbf{p})$, while $d_c(\mathbf{p}, \mathbf{q}) = (\rho_q - \rho_p)/\sqrt{c}$ grows without bound as ρ_q increases.*

Proof. The hyperbolic cosine law at $\mathbf{p}$ gives

$$\cos \varphi(\mathbf{p}, \mathbf{q}) = \frac{\cosh \rho_q - \cosh \rho_p \cosh \delta_{pq}}{\sinh \rho_p \sinh \delta_{pq}}. \tag{27}$$

If $\rho_p = \rho_q = \rho > 0$ and $\theta_{pq} > 0$, Eq. (5) gives

$$\cosh \delta_{pq} = 1 + \sinh^2 \rho \left(1 - \cos \theta_{pq}\right) > 1. \tag{28}$$

The numerator of Eq. (27) is therefore negative and the denominator is positive, so $\varphi(\mathbf{p}, \mathbf{q}) > \pi/2$. Any outward cone with half-aperture in $(0, \pi/2]$ excludes $\mathbf{q}$, although $d_c(\mathbf{p}, \mathbf{q}) \to 0$ as $\theta_{pq} \downarrow 0$.

Conversely, if $\theta_{pq} = 0$ and $\rho_q > \rho_p$, then $\mathbf{p}$ and $\mathbf{q}$ lie on the same outward radial geodesic. Hence $\delta_{pq} = \rho_q - \rho_p$ and $\varphi(\mathbf{p}, \mathbf{q}) = 0$. The point $\mathbf{q}$ remains inside the cone while $d_c(\mathbf{p}, \mathbf{q}) = (\rho_q - \rho_p)/\sqrt{c}$ can grow arbitrarily large. □

Proposition 1 shows that distance attraction does not guarantee cone membership, and distance repulsion does not guarantee exclusion. In particular, an unrelated candidate can move farther outward along the parent's radial ray without leaving its descendant region. Positive-only entailment supervision provides no direct penalty for this erroneous relation.

Why parent-radius apertures can lose selectivity. For conventional Lorentz entailment cones (Ganea et al., 2018; Desai et al., 2023), the half-aperture is

$$\omega(\mathbf{p}) = \arcsin\left(\frac{2\kappa}{\sqrt{c}\,\|\tilde{\mathbf{p}}\|}\right) = \arcsin\left(\frac{2\kappa}{\sinh\rho_p}\right), \qquad \kappa > 0. \tag{29}$$

Here κ is a fixed constant controlling cone width, and the admissible domain is $\rho_p \geq \operatorname{arsinh}(2\kappa)$. Thus, $\sin\omega(\mathbf{p}) = 2\kappa/\sinh\rho_p$, so directional selectivity is controlled by the *absolute* parent radius. As ρ_p decreases toward the admissible lower bound, $\omega(\mathbf{p})$ approaches $\pi/2$: the permitted directions fill the outward hemisphere, and the corresponding geodesic cone approaches an outward half-space. The same phenomenon is visible in Poincaré coordinates: from Eq. (18),

$$\sinh\rho_p = \frac{2\sqrt{c}\,\|\mathbf{w}_p\|}{1 - c\|\mathbf{w}_p\|^2}, \tag{30}$$

so the identical aperture becomes

$$\omega(\mathbf{w}_p) = \arcsin\left(\kappa\,\frac{1 - c\|\mathbf{w}_p\|^2}{\sqrt{c}\,\|\mathbf{w}_p\|}\right). \tag{31}$$

The absolute-depth coupling between cone width and parent position is therefore identical in the two coordinate models.

This limiting behavior also appears at practical operating points. ARGENT identifies parent-norm contraction and half-space widening as a cone-collapse failure mode in hyperbolic VLM training (Huynh et al., 2026). More recently, an operating-point audit of released MERU, HyCoCLIP, and PHyCLIP models reports regimes in which conventional cone or traversal mechanisms are saturated, inactive, or weakly aligned with the intended hierarchy (Kim et al., 2026c). These observations motivate an order geometry whose angular selectivity is governed by the *relative parent–child radial gap*. RAOC couples this gap to root-angle separation, and contrastive entailment supplies supervision for matched and unmatched relations.

C RAOC AS AN ORDER GEOMETRY

RAOC assigns each non-origin parent $\mathbf{p}$ the descendant set

$$\mathcal{C}_\beta(\mathbf{p}) = \left\{\mathbf{q} \in \mathbb{H}^d_c \setminus \{O\} : \rho_q - \rho_p \geq \frac{\theta_{pq}}{\beta}\right\}. \tag{32}$$

This form gives a direct semantic interpretation. Consider the parent description "*a dog*" and the more specific description "*a dog running on grass*." Their close directions allow a modest outward radial gap to encode the increase in specificity. An unrelated description such as "*a rocket successfully lifting off*" typically forms a much larger root angle, which raises the radial depth required to enter the dog's descendant region. Contrastive entailment supplies the semantic signal that separates these matched and unmatched relations. Once a parent–child relation is formed, the geometric rule composes consistently along deeper chains.

Theorem 1 (RAOC partial order)**.** *For any fixed $\beta > 0$, define $\mathbf{p} \preceq_\beta \mathbf{q}$ if and only if $\mathbf{q} \in \mathcal{C}_\beta(\mathbf{p})$. Then $\preceq_\beta$ is reflexive, antisymmetric, and transitive on $\mathbb{H}^d_c \setminus \{O\}$. Moreover,*

$$\mathbf{p} \preceq_\beta \mathbf{q} \quad \Longleftrightarrow \quad \mathcal{C}_\beta(\mathbf{q}) \subseteq \mathcal{C}_\beta(\mathbf{p}). \tag{33}$$

Proof. By definition,

$$\mathbf{p} \preceq_\beta \mathbf{q} \quad \Longleftrightarrow \quad \rho_q - \rho_p \geq \theta_{pq}/\beta. \tag{34}$$

Reflexivity follows because both sides vanish for $\mathbf{p} = \mathbf{q}$.

For antisymmetry, suppose $\mathbf{p} \preceq_\beta \mathbf{q}$ and $\mathbf{q} \preceq_\beta \mathbf{p}$. Adding the two inequalities gives $0 \geq 2\theta_{pq}/\beta$, hence $\theta_{pq} = 0$. The two radial inequalities then force $\rho_p = \rho_q$. Equation (13) gives $\mathbf{p} = \mathbf{q}$.

For transitivity, suppose $\mathbf{p} \preceq_\beta \mathbf{q}$ and $\mathbf{q} \preceq_\beta \mathbf{r}$. The spherical angle is a metric on the direction sphere, so

$$\theta_{pr} \le \theta_{pq} + \theta_{qr} \le \beta(\rho_q - \rho_p) + \beta(\rho_r - \rho_q) = \beta(\rho_r - \rho_p). \tag{35}$$

Hence $\mathbf{p} \preceq_\beta \mathbf{r}$.

Finally, if $\mathbf{p} \preceq_\beta \mathbf{q}$, transitivity places every descendant of $\mathbf{q}$ inside $\mathcal{C}_\beta(\mathbf{p})$, proving the forward implication in Eq. (33). Conversely, reflexivity gives $\mathbf{q} \in \mathcal{C}_\beta(\mathbf{q})$; therefore $\mathcal{C}_\beta(\mathbf{q}) \subseteq \mathcal{C}_\beta(\mathbf{p})$ implies $\mathbf{q} \in \mathcal{C}_\beta(\mathbf{p})$. This proves the reverse implication. □

If $\mathbf{p} \preceq_\beta \mathbf{q}$ and $\mathbf{p} \neq \mathbf{q}$, then the inclusion is strict. For example, if "*a dog*" precedes "*a dog running on grass*," every representation that is a descendant of the latter is automatically a descendant of the former. This nested-region property expresses increasing specificity through a shared relative-gap rule across radial levels.

C.1 Why the capped linear aperture is a natural transitive choice

More generally, consider the radial-gap aperture family

$$\mathbf{p} \preceq_f \mathbf{q} \quad \Longleftrightarrow \quad g(\mathbf{p}, \mathbf{q}) \ge 0, \qquad \theta_{pq} \le f(g(\mathbf{p}, \mathbf{q})), \tag{36}$$

where $f : [0, \infty) \to [0, \pi]$ and $f(0) = 0$. The next result characterizes exactly when such a construction is transitive over the full direction sphere.

Theorem 2 (Transitive radial-gap apertures)**.** *Let $d \ge 2$. The relation $\preceq_f$ above is transitive on $\mathbb{H}^d_c \setminus \{O\}$ if and only if*

$$f(a + b) \ge \min\{\pi, f(a) + f(b)\}, \qquad a, b \ge 0. \tag{37}$$

Proof. For sufficiency, let $\mathbf{p} \preceq_f \mathbf{q}$ and $\mathbf{q} \preceq_f \mathbf{r}$, and set $a = g(\mathbf{p}, \mathbf{q})$ and $b = g(\mathbf{q}, \mathbf{r})$. Radial gaps add, while the spherical triangle inequality gives

$$\theta_{pr} \le \min\{\pi, \theta_{pq} + \theta_{qr}\} \le \min\{\pi, f(a) + f(b)\} \le f(a + b). \tag{38}$$

Thus $\mathbf{p} \preceq_f \mathbf{r}$.

For necessity, fix $a, b \ge 0$ and a base radius $\rho_0 > 0$. Choose radii $\rho_p = \rho_0$, $\rho_q = \rho_0 + a$, and $\rho_r = \rho_0 + a + b$. Since $d \ge 2$, choose their directions on a common great circle with consecutive increments

$$\eta_1 = f(a), \qquad \eta_2 = \min\{f(b), \pi - f(a)\}. \tag{39}$$

Then $\mathbf{p} \preceq_f \mathbf{q}$ and $\mathbf{q} \preceq_f \mathbf{r}$, while

$$\theta_{pr} = \eta_1 + \eta_2 = \min\{\pi, f(a) + f(b)\}. \tag{40}$$

Transitivity therefore requires Eq. (37). □

Corollary 1 (Pointwise minimality at fixed local slope)**.** *Suppose $\preceq_f$ is transitive and*

$$\lim_{g \downarrow 0} \frac{f(g)}{g} = \beta > 0. \tag{41}$$

Then

$$f(g) \ge \min\{\pi, \beta g\} = f_\beta(g) \qquad \textit{for all } g \ge 0. \tag{42}$$

Proof. For $g > 0$, repeated application of Eq. (37) gives $f(g) \ge \min\{\pi, nf(g/n)\}$ for every positive integer n. Since $nf(g/n) \to \beta g$, taking $n \to \infty$ yields Eq. (42). The case $g = 0$ follows from $f(0) = 0$. □

The capped linear rule $f_\beta(g) = \min\{\pi, \beta g\}$ satisfies Eq. (37) with equality. At fixed local slope β, it is the pointwise tightest globally transitive aperture in this family. Equation (8) therefore selects the smallest descendant region compatible with the same local expansion rate and global transitivity.

C.2 Multiple parents and radial depth

RAOC also gives a simple geometric interpretation of compositional constraints. For non-origin parents $\mathbf{p}_1, \dots, \mathbf{p}_k$ and a fixed candidate direction $\mathbf{u}$, simultaneous membership is equivalent to

$$\rho_q \geq \rho_{\min}(\mathbf{u}) = \max_i \left\{ \rho_{p_i} + \frac{\theta(\mathbf{u}_{p_i}, \mathbf{u})}{\beta} \right\}, \qquad \theta(\mathbf{u}, \mathbf{v}) = \arccos(\mathbf{u}^\top \mathbf{v}). \tag{43}$$

Thus each parent contributes a lower bound on the radial depth required along the chosen direction, and the strictest parent determines the feasible depth.

For two parents with the same radius ρ and angular separation Θ, with $d \geq 2$,

$$\min_{\mathbf{q} \in \mathcal{C}_\beta(\mathbf{p}_1) \cap \mathcal{C}_\beta(\mathbf{p}_2)} \rho_q = \rho + \frac{\Theta}{2\beta}. \tag{44}$$

Indeed, the triangle inequality implies

$$\Theta \leq \theta(\mathbf{u}_{p_1}, \mathbf{u}) + \theta(\mathbf{u}, \mathbf{u}_{p_2}) \leq 2 \max_i \theta(\mathbf{u}_{p_i}, \mathbf{u}), \tag{45}$$

which gives the lower bound, and the midpoint of a shortest spherical path between the two parent directions attains it.

For a chain $\mathbf{p}_0 \preceq_\beta \cdots \preceq_\beta \mathbf{p}_k$, summing the edge constraints yields

$$\sum_{i=0}^{k-1} \theta_{p_i p_{i+1}} \leq \beta(\rho_{p_k} - \rho_{p_0}). \tag{46}$$

The total radial depth therefore acts as a geometric budget for accumulated directional change along a hierarchy.

C.3 Geodesic convexity versus order consistency

RAOC descendant regions can be non-convex along hyperbolic geodesics. A compact counterexample is obtained by fixing a parent direction $\mathbf{u}_p$, choosing $\mathbf{v} \perp \mathbf{u}_p$, and taking

$$\mathbf{u}_\pm = \cos a \, \mathbf{u}_p \pm \sin a \, \mathbf{v}, \qquad 0 < a < \pi/2. \tag{47}$$

Place two descendants in these directions at a common radius $R \geq \rho_p + \pi/\beta$. Their Lorentz geodesic midpoint lies in direction $\mathbf{u}_p$ and has radius ρ_m satisfying

$$\tanh \rho_m = \tanh R \cos a. \tag{48}$$

For a sufficiently close to $\pi/2$, $\rho_m < \rho_p$, and the midpoint falls outside the descendant region.

For hierarchical entailment, transitivity and nested descendant sets govern compositional consistency. Geodesic convexity controls the shape of the region along geodesics, while angular selectivity controls which directions are admitted at a given hierarchical depth. The conventional EC analysis above shows that a geodesically well-structured cone can still approach a broad half-space as the parent contracts, substantially weakening directional discrimination (Huynh et al., 2026; Kim et al., 2026c). RAOC makes this selectivity explicit through parent–child radial difference and root-angle separation.

D Optimization Analysis

D.1 Absolute-distance calibration and radial pressure

Let $\Delta = d_c(\mathbf{x}, \mathbf{y}) \geq 0$ be the unscaled geodesic distance and use the score $s = -\Delta$. With sigmoid logit $\zeta = b - a\Delta$, $a > 0$, the positive- and negative-pair losses are

$$\ell_+(\Delta) = \log(1 + \exp(a\Delta - b)), \qquad \ell_-(\Delta) = \log(1 + \exp(b - a\Delta)). \tag{49}$$

Their derivatives are

$$\frac{\partial \ell_+}{\partial \Delta} = a\,\sigma(a\Delta - b) > 0, \qquad \frac{\partial \ell_-}{\partial \Delta} = -a\,\sigma(b - a\Delta) < 0. \tag{50}$$

Thus every positive distance term favors a smaller absolute distance, whereas every negative term favors a larger one. Combined with Eq. (25), the positive term exerts inward radial pressure on the outward point under the fixed-direction conditions analyzed above.

For comparison, row-wise InfoNCE with score $-a\Delta$ is

$$\ell_i^{\mathrm{NCE}} = a\Delta_{ii} + \log \sum_j \exp(-a\Delta_{ij}). \tag{51}$$

If

$$P_{ij} = \frac{\exp(-a\Delta_{ij})}{\sum_k \exp(-a\Delta_{ik})}, \tag{52}$$

then

$$\frac{\partial \ell_i^{\mathrm{NCE}}}{\partial \Delta_{ii}} = a(1 - P_{ii}), \qquad \frac{\partial \ell_i^{\mathrm{NCE}}}{\partial \Delta_{ij}} = -aP_{ij} \quad (j \neq i). \tag{53}$$

InfoNCE therefore also attracts the positive pair. Pairwise sigmoid calibrates distance on an absolute scale, whereas row-wise InfoNCE normalizes it relative to the other candidates in the query row. Both retain the radial dependence induced by the distance score.

D.2 The bounded RAOC score used for training

The geometric relation is Eq. (8); the bounded score in Eqs. (9)–(10) provides its optimization interface. Its sign matches geometric membership exactly:

$$D_{\mathrm{RAOC}}(\mathbf{p}, \mathbf{q}) \geq 0 \quad \Longleftrightarrow \quad \mathbf{q} \in \mathcal{C}_\beta(\mathbf{p}). \tag{54}$$

Indeed, $m_r \geq 0$ is equivalent to $g \geq 0$, and on that domain $m_a \geq 0$ is equivalent to $\theta_{pq} \leq \min\{\pi, \beta g\}$. Clipping bounds the optimization score and preserves the zero-threshold RAOC membership. The contrastive entailment loss uses this continuous score, and the order analysis above uses the corresponding exact relation.

For negative gaps, the angular allowance is clipped to zero, while $m_r < 0$ ensures a negative entailment score even when $\theta_{pq} = 0$.

E Relation to Recent Hyperbolic Vision–Language Objectives

Exterior-angle alignment. ATMG (Ramasinghe et al., 2024) analyzes the tension between geodesic-proximity alignment and hierarchical organization and introduces an exterior-angle contrastive objective. Its score reduces the dependence of semantic matching on absolute radial position. AURORA uses the common-origin root angle, whose Lorentz and Poincaré realizations are given in Eq. (20); RAOC then carries the radial structure required for hierarchical ordering.

Parent-dependent entailment cones. MERU and HyCoCLIP employ EC descendant regions whose aperture varies with the parent norm, while PHyCLIP applies hierarchy constraints in a factorized hyperbolic representation (Desai et al., 2023; Pal et al., 2025; Yoshikawa & Matsubara, 2026). Section B.1 shows how the EC aperture broadens as the parent approaches its admissible inner radius. ARGENT (Huynh et al., 2026) addresses this training regime with adaptive exterior-angle supervision and radial stabilization, and a later operating-point audit reports saturation or weak activation of conventional cone/traversal mechanisms in several released hyperbolic VLM checkpoints (Kim et al., 2026c). RAOC parameterizes descendant selectivity directly through the relative radial gap and root angle; the resulting relation admits the transitivity and descendant-set nesting guarantees in Theorem 1.

Uncertainty-guided compositional alignment. UNCHA is a concurrent 2026 method that models part-to-whole semantic representativeness through a norm-derived uncertainty signal (Kim et al., 2026a). The uncertainty weights its Lorentz-distance contrastive alignment and calibrates an

entropy-regularized entailment term. Its entailment branch is formulated around observed part–whole relations, whereas AURORA assigns entailment labels to both matched and unmatched parent–candidate pairs. The two methods therefore organize compositional supervision differently: UNCHA modulates distance-based alignment according to part representativeness, and AURORA separates semantic matching into root-angle alignment while learning radial order through contrastive RAOC supervision.

During our reproduction, we found that the public `uncha/data/webdataset_mapper.py` contains the same evaluation-dataset implementation as `uncha/data/evaluation.py`, leaving the training WebDataset mapper implementation absent from the released code path (Kim et al., 2026b). We reconstructed the mapper from the data pipeline described in the paper and the surrounding training interfaces. Since the resulting implementation did not consistently reproduce the reported trends across repeated runs, we exclude reproduced UNCHA numbers from the controlled comparison tables to avoid comparisons based on an implementation whose fidelity could not be fully verified.

F Implementation Details

Model architecture. The text branch uses the CLIP Transformer configuration with 12 layers, hidden width 512, maximum sequence length 77, and vocabulary size 49,408 (Radford et al., 2021). The visual branch uses ViT-S/16 for the principal experiments and ViT-B/16 for the larger-backbone comparison, with 16×16 image patches (Dosovitskiy et al., 2021). Images are resized with border padding and random-resized cropping to 224×224; the crop scale is sampled from $[0.5, 1.0]$. This yields a 14×14 patch grid (196 image tokens), to which fixed two-dimensional sine–cosine positional embeddings are added. Full images and object crops share the visual encoder, while captions and object phrases share the text encoder.

Training data and hyperbolic parameterization. We use the recoverable subset of GRIT (Peng et al., 2024), containing 10.3M valid image–caption pairs and 17.8M grounded box–phrase annotations after local recovery and validation. All four representation types (I, T, I^b, T^b) are projected to the same d-dimensional Lorentz space. The shared curvature magnitude c is learned globally, initialized at 1.0, and constrained to $[0.1, 10]$. The principal ViT-S/16 configuration uses

$$\lambda_e = 0.15, \qquad \lambda_n = 0.05, \qquad \beta = 8, \qquad \sigma_r = 0.1. \tag{55}$$

The parameter β determines the geometric expansion rate of RAOC and is therefore part of the order definition. The parameter σ_r sets the scale of the bounded radial optimization margin while preserving zero-threshold membership. We use the norm regularizer defined in Sec. 4.4 to penalize both very small and excessively large spatial norms, stabilizing training and maintaining a usable radial range for hierarchical organization.

Optimizer, schedule, and hardware. All principal runs use four NVIDIA A40 48GB GPUs, a global batch size of 768, and 500k optimization steps. We use AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and weight decay 0.2; weight decay is disabled for learned scalar parameters. The learning rate is linearly warmed up for the first 4k steps to a maximum of 5×10^{-4} and then follows cosine decay (Loshchilov & Hutter, 2017). We keep these settings fixed across all principal AURORA runs and downstream benchmarks.

Training efficiency. We profile ViT-S/16 training under the common setup of four NVIDIA A40 48GB GPUs, global batch size 768, and 500k optimization steps. HyCoCLIP and PHyCLIP are reproduced using their original source code without implementation changes. Our implementation streamlines the training computation branches and eliminates redundant matrix computations, while keeping all other training settings the same. The efficiency comparison therefore reflects these implementation optimizations as well as differences in the training objectives.

Table 6 reports the mean peak memory and end-to-end training time over three runs with randomly selected seeds. AURORA uses 13.8GB per GPU and completes training in 77.51 hours, compared with 39.4GB and 91.8 hours for HyCoCLIP and 40.9GB and 93.4 hours for PHyCLIP. This corresponds to about 65–66% lower peak VRAM and 16–17% shorter training time under the matched environment. The AURORA-InfoNCE variant remains below both hyperbolic baselines at 21.2GB per GPU and 80.83 hours. Together with the stronger downstream results in Sec. 5.2, these mea-

surements give AURORA a favorable accuracy–efficiency profile under the shared hardware and schedule.

Table 6: ViT-S/16 training efficiency in the common four-A40 environment. Values are means over three runs with randomly selected seeds.

Model	GPUs	Peak VRAM/GPU (GB)	Train time (h)
HyCoCLIP	4×A40	39.4	91.80
PHyCLIP	4×A40	40.9	93.40
AURORA-InfoNCE	4×A40	21.2	80.83
AURORA	4×A40	**13.8**	**77.51**

G Complete Experimental Tables

We report complete results for ViT-S/16 and ViT-B/16, extending the comparisons in the main text to both model sizes.

Zero-shot image classification. Table 7 shows that AURORA achieves the highest accuracy on 13 of the 16 datasets with both backbones. Its advantages span general, fine-grained, and specialized recognition, with particularly clear gains on ImageNet, Caltech-101, Pets, and Flowers. Moving from ViT-S/16 to ViT-B/16 improves AURORA on every dataset while preserving its overall ranking relative to the baselines. The results therefore extend the classification gains observed in the main text to the larger backbone.

Table 7: **Complete zero-shot image classification comparison for ViT-S/16 and ViT-B/16.** Top-1 accuracy (%). ViT-S/16 and ViT-B/16 best results are shown in **bold** and second-best results are underlined.

			General datasets						Fine-grained datasets						Specialized datasets			
Backbone	**Model**	w/ boxes	ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Food-101	CUB	Cars	Aircraft	Pets	Flowers	DTD	EuroSAT	RESISC45	Country211
ViT S/16	CLIP	✗	32.66	74.19	42.22	42.91	69.80	90.09	42.72	12.55	3.93	2.06	43.22	18.03	17.57	25.47	33.64	3.88
	CLIP	✓	32.49	73.69	42.15	42.66	70.07	89.83	42.50	12.48	4.01	2.10	43.28	17.79	17.61	26.04	33.37	3.84
	SigLIP	✗	34.30	66.57	38.38	44.73	71.42	91.75	38.98	13.33	4.89	2.13	46.65	23.17	18.40	**32.63**	35.42	4.15
	SigLIP	✓	34.34	67.07	38.50	44.82	71.55	91.82	39.27	13.33	4.92	2.24	46.56	23.03	18.36	31.42	35.37	4.23
	MERU	✗	31.74	76.08	41.24	41.28	67.76	87.24	41.09	11.97	3.80	2.88	42.53	16.75	17.46	27.75	33.88	3.82
	MERU	✓	32.01	76.55	41.35	41.21	67.64	87.30	41.24	12.08	3.78	**2.95**	42.54	16.57	17.52	27.53	33.82	3.83
	ATMG	✗	33.74	77.21	43.28	43.15	68.08	90.84	44.01	13.27	4.38	1.75	42.53	15.70	17.89	28.69	35.75	4.21
	HyCoCLIP	✓	36.38	83.28	47.91	47.88	71.53	93.16	46.73	15.33	5.98	2.45	45.59	20.89	19.47	26.00	38.73	4.09
	PHyCLIP	✓	37.06	83.44	48.41	48.88	71.82	93.06	47.26	16.29	6.58	2.57	46.44	21.46	19.13	25.81	39.79	3.95
	ARGENT	✓	37.55	83.55	48.62	49.29	72.15	**94.01**	47.70	16.90	6.81	2.48	47.10	21.65	19.67	26.32	40.46	4.19
	AURORA	✓	**41.54**	**84.25**	**51.20**	**50.27**	**75.81**	93.19	**50.13**	**17.71**	**8.91**	1.80	**51.41**	**28.99**	**22.87**	30.18	**41.09**	**4.70**
ViT B/16	CLIP	✗	36.28	78.42	48.72	46.88	72.54	92.51	47.18	13.92	5.72	4.12	46.82	22.31	18.62	28.73	36.18	5.02
	SigLIP	✗	38.62	74.01	47.34	49.82	75.11	94.18	43.82	15.28	7.12	4.30	51.62	29.18	19.96	**36.91**	38.41	5.43
	MERU	✗	35.41	79.62	47.91	45.54	70.38	90.21	45.46	13.61	5.38	**5.51**	46.34	20.74	18.41	31.02	36.37	4.86
	ATMG	✗	37.83	80.92	50.22	47.56	71.32	93.36	48.44	15.14	6.21	3.58	46.77	19.42	19.02	32.48	38.32	5.34
	HyCoCLIP	✓	40.51	87.61	55.71	52.63	74.82	95.47	51.82	17.43	8.42	4.78	50.11	25.04	20.62	29.66	41.61	5.28
	PHyCLIP	✓	41.63	87.86	56.44	53.91	75.19	95.81	52.74	18.21	9.31	5.18	50.94	26.41	20.39	29.21	42.72	5.19
	ARGENT	✓	42.31	88.02	57.03	54.37	75.62	**96.46**	53.42	18.96	9.88	4.96	51.83	27.46	21.11	29.83	43.67	5.57
	AURORA	✓	**46.43**	**88.57**	**59.83**	**55.51**	**79.12**	95.96	**55.71**	**19.94**	**12.70**	3.54	**56.42**	**34.46**	**24.47**	33.86	**44.22**	**6.09**

Image–text retrieval and hierarchical classification. As shown in Table 8, AURORA leads all reported retrieval and hierarchical classification metrics with both backbones. The retrieval advantage holds in both directions on COCO and Flickr30K, at Recall@5 and Recall@10. With ViT-B/16, AURORA further improves retrieval performance and reduces hierarchical errors, accompanied by higher Jaccard similarity, hierarchical precision, and hierarchical recall. Thus, stronger semantic matching and closer agreement with the WordNet hierarchy are observed at both model sizes.

Table 8: **Complete zero-shot retrieval and hierarchical classification comparison for ViT-S/16 and ViT-B/16.** Retrieval results are Recall@5/10 (%). Hierarchical metrics are rounded to two decimal places for display; best and second-best markings follow the underlying values.

Backbone	Model	w/ boxes	Text → Image				Image → Text				Hierarchical Classification				
			COCO		Flickr		COCO		Flickr		WordNet				
			R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	TIE↓	LCA↓	J ↑	P_H ↑	R_H ↑
ViT S/16	CLIP	✗	50.31	61.52	76.98	84.48	65.75	76.15	87.10	93.00	4.43	2.51	0.72	0.79	0.80
	CLIP	✓	50.27	61.47	76.75	84.42	65.49	76.08	87.20	93.17	4.42	2.50	0.72	0.78	0.80
	SigLIP	✗	52.32	63.90	77.54	85.24	65.98	76.78	88.60	93.23	4.22	2.47	0.75	0.83	0.82
	SigLIP	✓	52.27	63.81	77.21	84.09	66.03	76.84	88.60	93.40	4.23	2.47	0.74	0.83	0.83
	MERU	✗	49.79	61.14	76.03	84.34	65.01	75.67	86.80	92.77	4.44	2.52	0.71	0.78	0.79
	MERU	✓	49.63	61.08	76.13	84.21	65.16	75.74	86.93	92.77	4.44	2.51	0.71	0.78	0.78
	ATMG	✗	50.95	62.36	77.21	84.74	64.96	75.81	87.10	93.17	4.37	2.48	0.73	0.82	0.81
	HyCoCLIP	✓	51.35	63.13	77.92	85.52	64.84	75.78	87.30	93.23	3.97	2.33	0.76	0.84	0.84
	PHyCLIP	✓	52.24	64.08	78.26	85.84	65.65	76.21	87.20	93.17	3.92	2.31	0.76	0.84	0.84
	ARGENT	✓	52.68	64.46	78.59	86.18	65.75	76.46	87.50	93.23	3.91	2.30	0.77	0.84	0.84
	AURORA	✓	**55.77**	**66.81**	**81.02**	**88.18**	**69.42**	**79.18**	**89.90**	**95.13**	**3.62**	**2.25**	**0.78**	**0.85**	**0.85**
ViT B/16	CLIP	✗	53.08	64.36	79.73	86.63	68.32	78.56	89.63	94.16	3.88	2.34	0.77	0.83	0.84
	SigLIP	✗	54.62	65.92	80.42	87.31	69.15	79.36	92.03	95.63	3.69	2.29	0.78	0.84	0.85
	MERU	✗	52.71	63.98	79.10	86.18	67.69	77.83	88.97	93.91	3.97	2.36	0.76	0.83	0.83
	ATMG	✗	53.94	65.14	80.09	86.82	68.47	78.54	89.88	94.37	3.80	2.31	0.77	0.84	0.84
	HyCoCLIP	✓	55.12	66.41	81.30	87.78	69.84	79.72	90.63	94.93	3.48	2.16	0.79	0.85	0.85
	PHyCLIP	✓	55.83	67.07	82.03	88.31	70.72	80.61	90.72	95.12	3.40	2.14	0.80	0.86	0.86
	ARGENT	✓	56.52	67.86	82.74	88.87	71.36	81.23	91.48	95.40	3.34	2.12	0.80	0.86	0.86
	AURORA	✓	**59.41**	**70.10**	**85.60**	**90.70**	**73.82**	**83.18**	**92.50**	**96.30**	**3.18**	**2.07**	**0.81**	**0.87**	**0.87**

Compositional understanding. Table 9 shows a consistent pattern across both backbones: AURORA leads on nine of the fourteen measures, including the overall SugarCrepe score. On VL-CheckList-Object, it performs best for objects in central and intermediate locations and for medium and small objects. On SugarCrepe, its strengths lie in replacing and adding objects or attributes, while relation replacement and swapping remain more challenging. These trends persist with ViT-B/16, indicating that the compositional benefits extend to the larger model.

Table 9: **Complete hard-negative compositional evaluation for ViT-S/16 and ViT-B/16.**

Backbone	Model	w/ boxes	VL-CheckList-Object						SugarCrepe							
			Location			Size			Replace			Swap		Add		
			Center	Mid	Margin	Large	Medium	Small	Obj	Att	Rel	Obj	Att	Obj	Att	Overall
ViT S/16	CLIP	✗	85.10	79.35	71.22	84.48	74.61	65.72	86.74	78.00	69.04	59.49	65.27	77.53	70.76	75.71
	CLIP	✓	85.24	78.41	70.55	85.32	75.02	64.67	85.94	77.45	**69.94**	59.89	64.41	77.90	70.76	75.69
	SigLIP	✗	89.33	83.04	74.91	88.30	78.50	68.82	87.35	78.55	69.49	60.82	65.77	78.08	71.10	76.26
	SigLIP	✓	89.83	83.03	74.50	88.13	78.13	69.47	87.01	78.43	69.18	**61.38**	**66.67**	78.03	70.28	76.12
	MERU	✗	82.07	76.10	67.93	81.47	71.50	62.40	86.40	77.58	68.73	58.94	64.96	77.17	70.33	75.35
	MERU	✓	82.90	75.40	68.40	81.33	71.73	62.40	87.09	77.03	69.11	58.81	65.17	77.19	70.47	75.55
	ATMG	✗	84.07	78.40	69.83	83.53	73.60	64.53	87.11	78.34	69.35	60.16	65.62	77.98	71.15	76.10
	HyCoCLIP	✓	90.80	86.75	79.95	90.52	82.73	74.23	87.23	78.81	66.64	60.41	64.86	79.44	69.08	75.82
	PHyCLIP	✓	91.40	87.14	80.42	90.97	83.06	74.79	88.40	80.25	67.33	59.76	65.22	79.89	71.48	76.71
	ARGENT	✓	91.87	87.73	**81.10**	**91.33**	84.07	75.60	88.92	80.88	67.71	59.89	65.32	80.00	72.59	77.11
	AURORA	✓	**91.93**	**87.80**	80.77	91.30	**84.69**	**76.53**	**89.77**	**81.85**	68.42	60.00	65.47	**80.16**	**74.27**	**77.75**
ViT B/16	CLIP	✗	86.18	80.72	72.64	85.52	76.15	67.43	88.00	78.72	70.11	59.81	65.54	80.63	72.52	77.31
	SigLIP	✗	90.30	84.35	76.24	89.13	79.79	70.32	88.75	79.35	**70.68**	**61.11**	**66.03**	81.44	73.12	78.02
	MERU	✗	84.28	78.29	70.22	83.27	74.02	65.19	87.55	78.20	69.78	59.28	65.24	80.16	72.10	76.89
	ATMG	✗	86.01	80.70	72.45	85.37	76.10	67.10	88.41	79.03	70.50	60.47	65.91	81.16	73.07	77.76
	HyCoCLIP	✓	91.88	88.12	81.47	91.54	84.29	75.95	88.52	79.50	67.81	60.72	65.15	82.54	71.00	77.46
	PHyCLIP	✓	92.45	88.53	82.02	91.93	84.62	76.56	89.67	80.93	68.50	60.07	65.50	83.04	73.43	78.36
	ARGENT	✓	92.78	89.17	**82.63**	**92.45**	85.55	77.22	90.26	81.57	68.90	60.21	65.62	83.38	74.54	78.85
	AURORA	✓	**93.01**	**89.30**	82.42	92.31	**86.25**	**78.38**	**91.10**	**82.49**	69.84	60.41	65.77	**83.95**	**76.59**	**79.67**

H Evaluation Protocols

H.1 Zero-shot image classification

We evaluate the same 16 classification datasets for every checkpoint: ImageNet (Deng et al., 2009), CIFAR-10 (Krizhevsky, 2009), CIFAR-100 (Krizhevsky, 2009), SUN397 (Xiao et al., 2010), Caltech-101 (Fei-Fei et al., 2004), STL-10 (Coates et al., 2011), Food-101 (Bossard et al., 2014), CUB-200-2011 (Wah et al., 2011), Stanford Cars (Krause et al., 2013), FGVC Aircraft (Maji et al., 2013), Oxford-IIIT Pets (Parkhi et al., 2012), Oxford Flowers-102 (Nilsback & Zisserman, 2008), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), RESISC45 (Cheng et al., 2017), and Country211 (Radford et al., 2021). For a given dataset, all models use the same class vocabulary and prompt templates. AURORA scores image embeddings against text prototypes using $s_{\angle}$; each baseline is evaluated with the similarity function native to that model. We report top-1 accuracy.

H.2 Image–text retrieval

Bidirectional retrieval is evaluated on the COCO validation set (Lin et al., 2014) and the Flickr30K test set (Young et al., 2014). In text-to-image retrieval, each caption queries the image collection; in image-to-text retrieval, each image queries the caption collection. We report Recall@5 and Recall@10 in both directions, matching the columns in the main and appendix tables. AURORA ranks candidates by root-angle similarity, while each baseline retains its native released retrieval score.

H.3 Hierarchical classification on WordNet

We evaluate whether zero-shot ImageNet predictions remain close to the ground-truth category in the WordNet hierarchy (Deng et al., 2009; Miller, 1995). Let $\hat{y}$ and y denote the predicted and true synsets. Tree-Induced Error (TIE) measures the shortest hierarchical path cost between them (Dekel et al., 2004):

$$\mathrm{TIE}(\hat{y}, y) = \sum_{e \in P(\hat{y}, y)} w_e, \tag{56}$$

where all WordNet edge weights are set to $w_e = 1$. Let $a^\star$ be the deepest common ancestor of $\hat{y}$ and y. Following the hierarchical-classification protocol used by the compared hyperbolic VLMs (Pal et al., 2025; Yoshikawa & Matsubara, 2026), the LCA error is the larger of the two distances from the labels to this common ancestor,

$$\mathrm{LCA}(\hat{y}, y) = \max\{d_{\mathcal{G}}(\hat{y}, a^\star), d_{\mathcal{G}}(y, a^\star)\}\,. \tag{57}$$

Lower TIE and LCA indicate hierarchically closer errors.

For set-based measures, let $\widehat{Y}_{\mathrm{anc}}$ and Y_{anc} be the ancestor sets of $\hat{y}$ and y. We report Jaccard similarity, hierarchical precision, and hierarchical recall (Kosmopoulos et al., 2015):

$$J = \frac{|\widehat{Y}_{\mathrm{anc}} \cap Y_{\mathrm{anc}}|}{|\widehat{Y}_{\mathrm{anc}} \cup Y_{\mathrm{anc}}|}, \qquad P_H = \frac{|\widehat{Y}_{\mathrm{anc}} \cap Y_{\mathrm{anc}}|}{|\widehat{Y}_{\mathrm{anc}}|}, \qquad R_H = \frac{|\widehat{Y}_{\mathrm{anc}} \cap Y_{\mathrm{anc}}|}{|Y_{\mathrm{anc}}|}. \tag{58}$$

Higher J, P_H, and R_H indicate better agreement with the reference taxonomy.

H.4 VL-CheckList–Object

VL-CheckList (Zhao et al., 2022) constructs controlled image–text tests by modifying a caption while keeping the image fixed. We use the object subset (VL-CO), in which an object noun phrase is replaced to create an incorrect alternative. Results are grouped by the scale and image location of the referenced object.

For object scale, let A denote the bounding-box area. The three groups are

$$\text{O-Small} : A < 32^2, \qquad \text{O-Medium} : 32^2 \le A \le 96^2, \qquad \text{O-Large} : A > 96^2. \tag{59}$$

For location, let x be the image half-diagonal and y the Euclidean distance between the object-box center and the image center. The groups are

$$\text{O-Center} : \frac{y}{x} \leq \frac{1}{3}, \qquad \text{O-Mid} : \frac{1}{3} < \frac{y}{x} \leq \frac{2}{3}, \qquad \text{O-Margin} : \frac{y}{x} > \frac{2}{3}. \tag{60}$$

For each subset, accuracy is the fraction of cases in which the model assigns the correct image–text pair a higher compatibility score than the corrupted alternative.

H.5 SugarCrepe

SugarCrepe (Hsieh et al., 2023) evaluates image-to-text compositional discrimination using fluent hard-negative captions. We report the seven released perturbation categories: Replace-Object, Replace-Attribute, Replace-Relation, Swap-Object, Swap-Attribute, Add-Object, and Add-Attribute. Each example asks the model to prefer the original caption over its hard negative; the overall score is the mean over the reported categories.

I Additional Hierarchy Visualizations

We provide additional visualizations of AURORA ViT-S with and without RAOC to examine semantic alignment and hierarchical organization together. Each scatter plot shows the root angle θ to a fixed query and the radial gap g between modalities. Smaller root angles indicate stronger semantic alignment. Dashed lines connect concepts or image regions according to the displayed hierarchy, allowing their radial order to be inspected alongside their similarity to the query. Orange squares denote related alternatives, and green crosses denote unrelated controls. The original axis ranges are retained for each model.

I.1 Image to Text

Figures 6 and 7 compare ten image queries. Each row shows the same example with RAOC on the left and without RAOC on the right. Blue circles trace the semantic path from a broad category to the query class. With RAOC, these paths consistently move toward smaller radial gaps and smaller root angles as the concepts become more specific. Without RAOC, the query class can still have a small root angle, while the radial sequence reverses in examples such as electric guitar, leopard, monarch, and zebra. The piano, laptop, and desk examples retain an ordered path in both variants. Together, these comparisons support the role of RAOC in making hierarchical order more consistent while preserving semantic alignment.

I.2 Text to Image

Figures 8–12 compare five text queries: shoes, eggplant, earphone, chair, and bread. Each figure places the model with RAOC above the model without RAOC. The image region hierarchy on the right provides the same reference structure for both models, and the colored dashed links identify its two branches in the scatter plot. With RAOC, the branch containing the queried object follows an increasing radial gap and a decreasing root angle from the full image through the intermediate region to the object crop. Without RAOC, angular alignment often remains, but the radial progression can reverse or the intermediate region can be less aligned than the full image. These examples provide qualitative evidence that RAOC couples semantic alignment with the radial organization associated with hierarchical partial order.

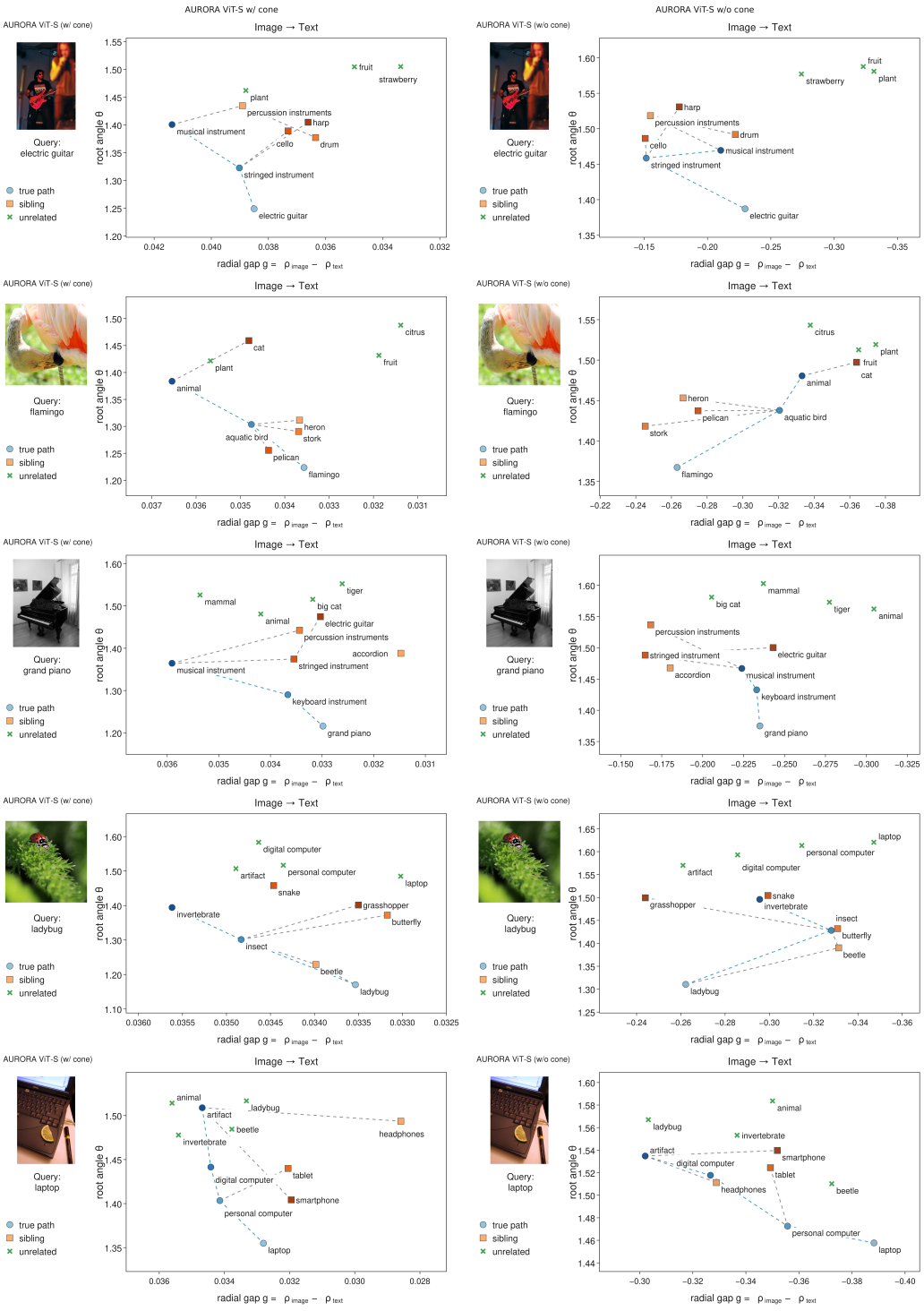

Figure 6: **Image to text hierarchy comparisons for five object queries.** Rows show electric guitar, flamingo, grand piano, ladybug, and laptop. The left column uses RAOC, and the right column omits it. Dashed blue lines trace the query's semantic path, while gray lines show related branches. RAOC preserves the angular preference for the query class and gives a consistent radial progression along each path. The radial reversals for electric guitar and ladybug without RAOC illustrate why angular alignment alone does not ensure hierarchical order.

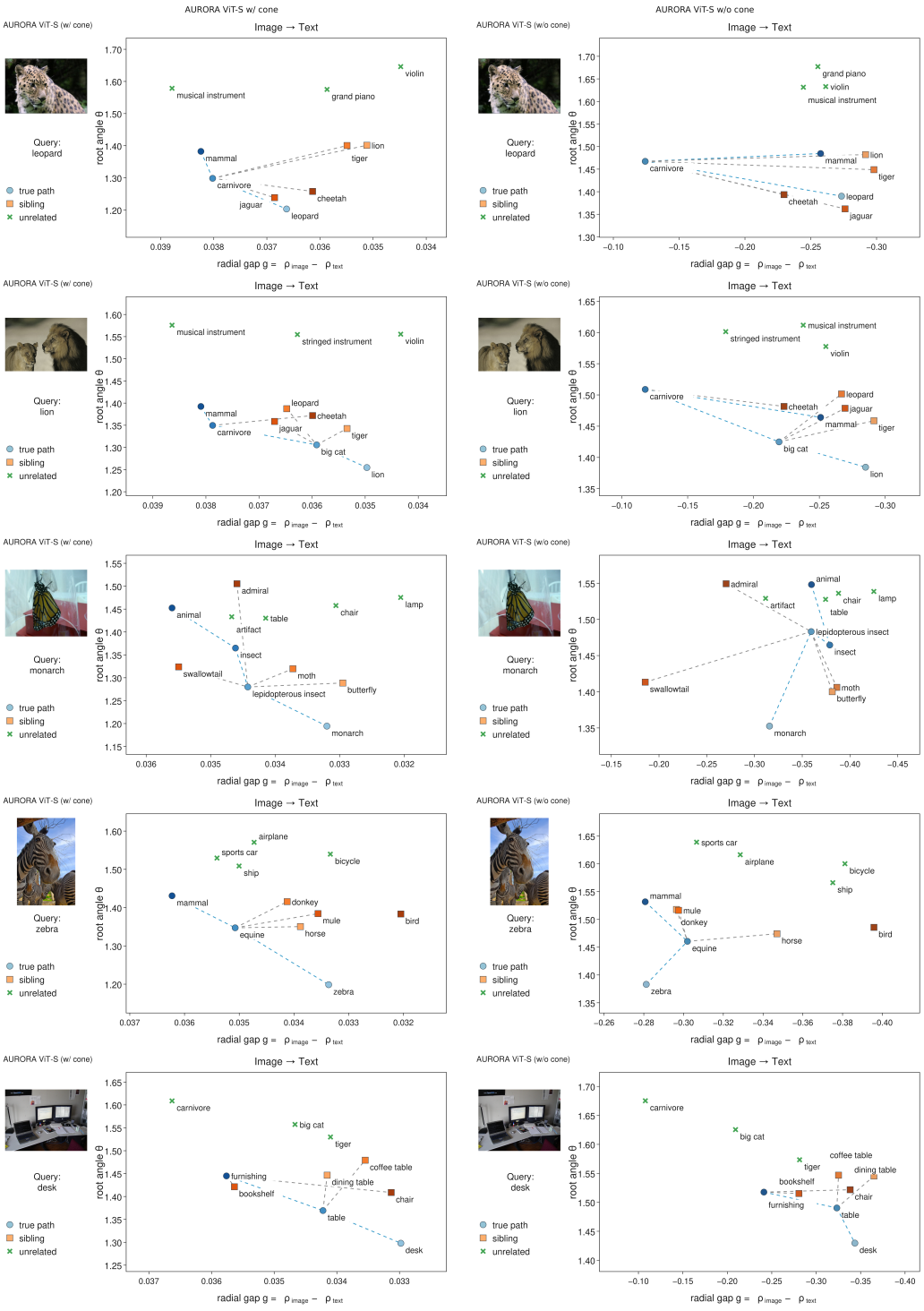

Figure 7: **Image to text hierarchy comparisons for five further queries.** Rows show leopard, lion, monarch, zebra, and desk, with RAOC on the left and without RAOC on the right. RAOC places the displayed semantic paths in a consistent radial order. Without RAOC, the animal examples exhibit reversals between successive concepts even when the query class remains well aligned. The desk path is ordered in both variants, showing that the effect concerns consistency across examples rather than the absence of all hierarchy without RAOC.

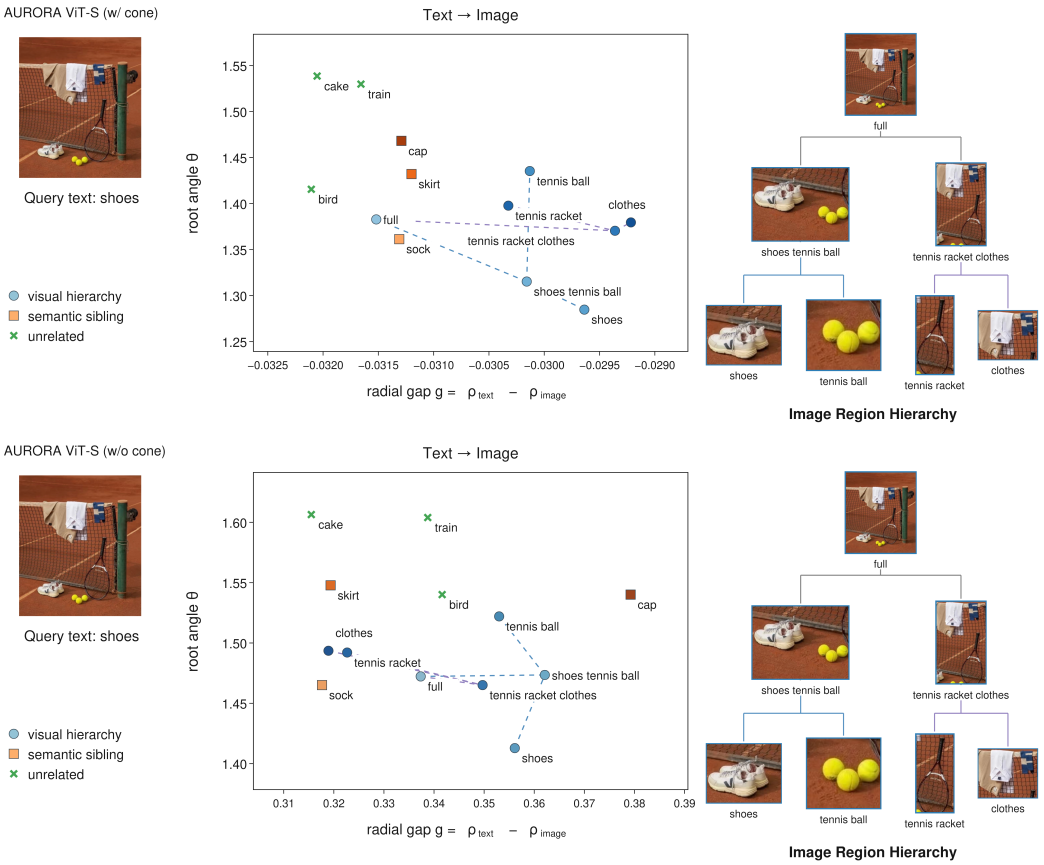

Figure 8: **Text to image hierarchy for the query "shoes."** AURORA ViT-S with RAOC is shown above the variant without RAOC. With RAOC, the full image, shoes and tennis ball region, and shoes crop follow a consistent radial progression as their root angles decrease. Without RAOC, the shoes crop remains the most aligned region, but the final step reverses the radial direction. The region tree provides the common reference hierarchy.

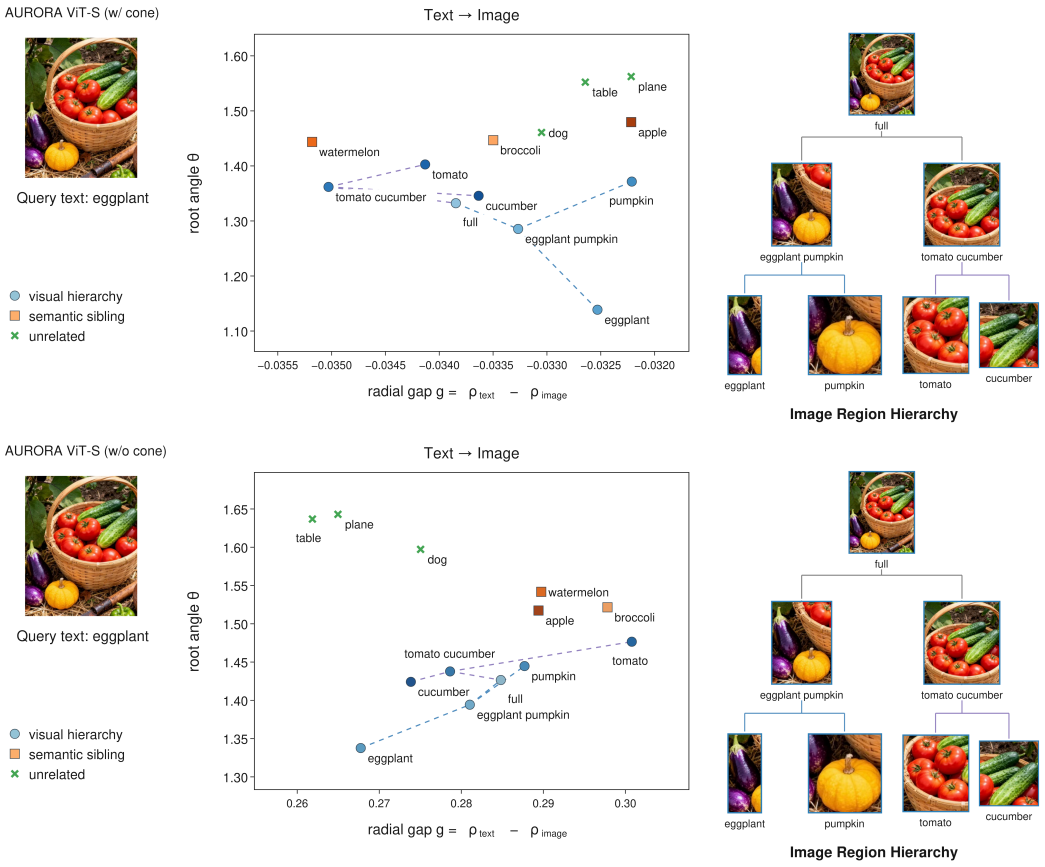

Figure 9: **Text to image hierarchy for the query "eggplant."** The top panel uses RAOC, and the bottom panel omits it. Both models assign the eggplant crop a smaller root angle than the full image. RAOC also places the full image, eggplant and pumpkin region, and eggplant crop in the expected radial order. Without RAOC, this branch progresses in the opposite radial direction despite retaining its angular preference for eggplant.

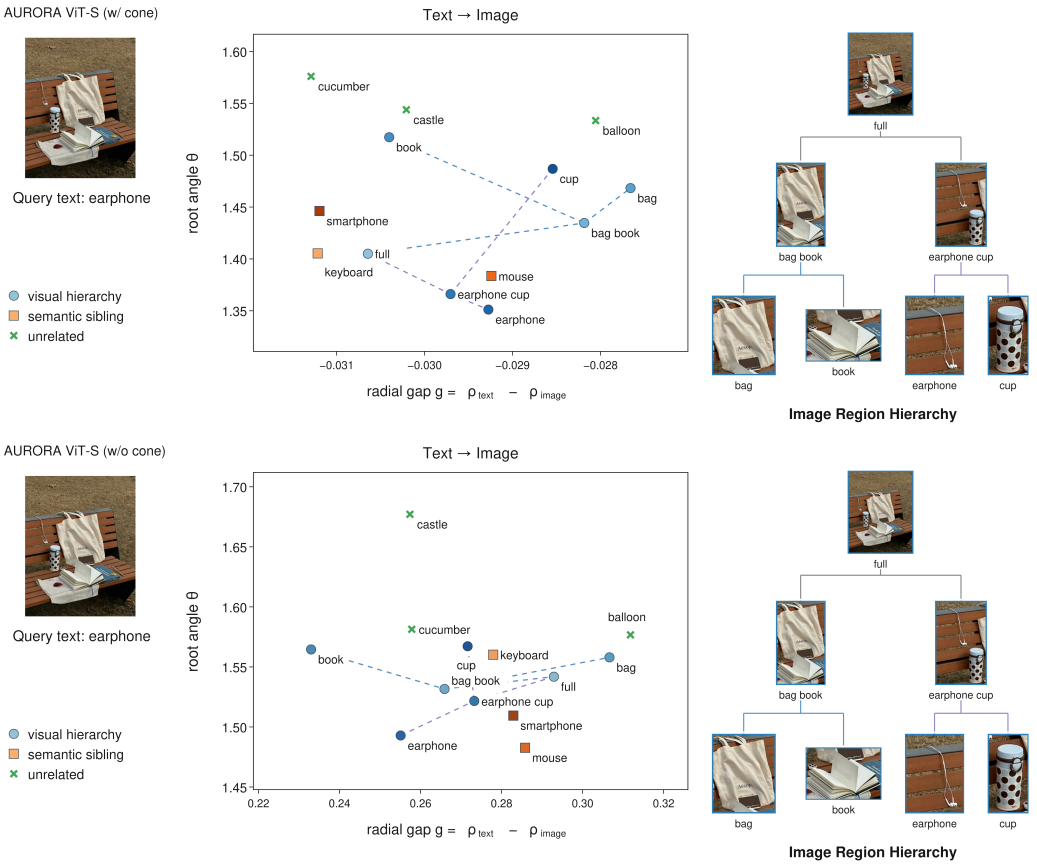

Figure 10: **Text to image hierarchy for the query "earphone."** The top and bottom panels show the models with and without RAOC, respectively. With RAOC, the full image, earphone and cup region, and earphone crop combine progressively smaller root angles with increasing radial gaps. Without RAOC, the same angular trend is present, but the radial sequence is reversed. The comparison separates semantic alignment from the hierarchical ordering encouraged by RAOC.

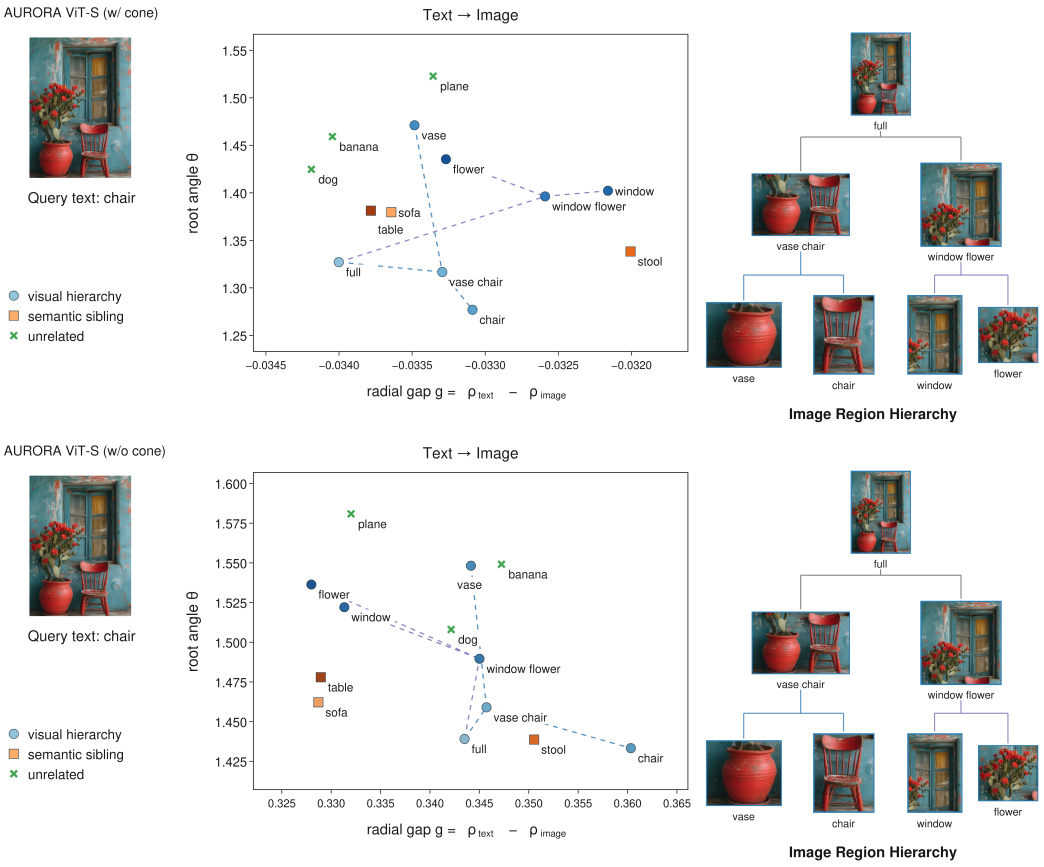

Figure 11: **Text to image hierarchy for the query "chair."** RAOC is used in the top panel and omitted in the bottom panel. Both variants retain radial progression along the full image, vase and chair region, and chair crop. With RAOC, their root angles also decrease consistently. Without RAOC, the intermediate region has a larger root angle than the full image, revealing less consistent angular organization along this branch.

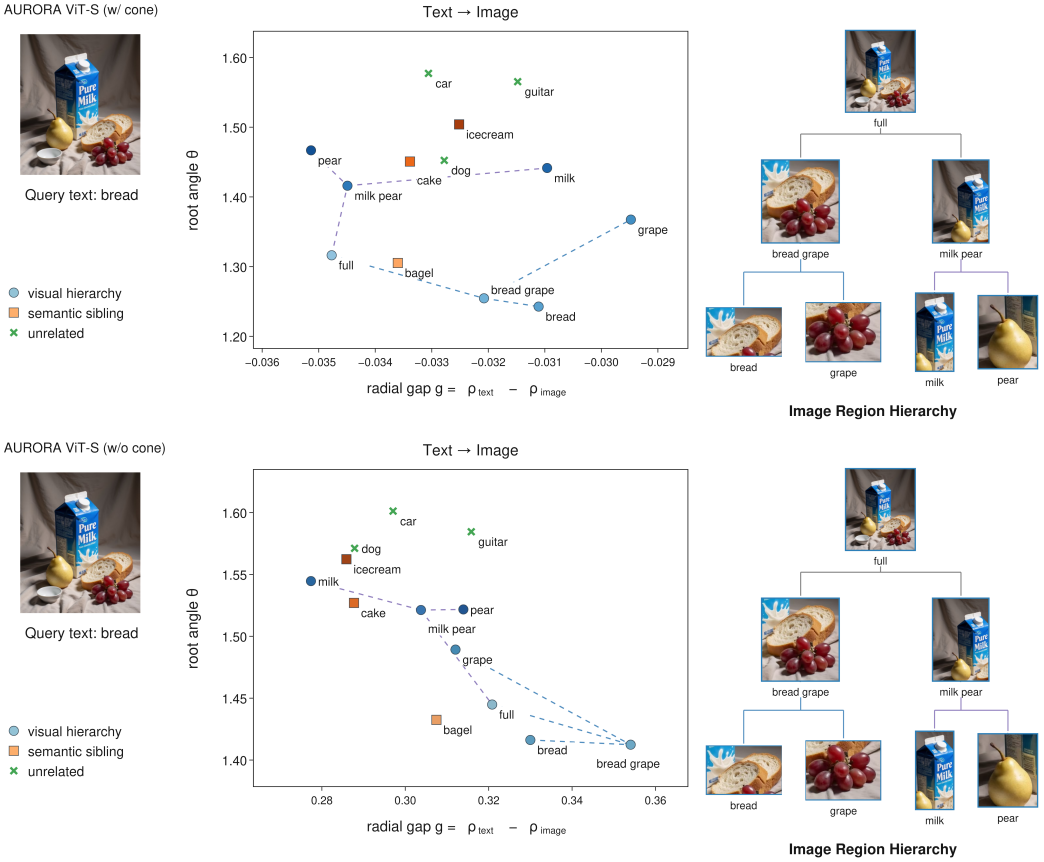

Figure 12: **Text to image hierarchy for the query "bread."** The model with RAOC appears above the model without RAOC. With RAOC, the full image, bread and grape region, and bread crop form a consistent sequence in both radial gap and root angle. Without RAOC, the bread and grape region has a slightly smaller root angle than the bread crop, and the final transition reverses the radial direction. RAOC yields clearer organization along the branch containing the queried object.

J HYPERPARAMETER SENSITIVITY

We examine the sensitivity of AURORA to four geometric and optimization parameters: the RAOC angular allowance β, the entailment weight λ_e, the norm regularization weight λ_n, and the radial margin scale σ_r. Each experiment varies one parameter while keeping the others fixed at the default configuration: $\beta = 8$, $\lambda_e = 0.15$, $\lambda_n = 0.05$, and $\sigma_r = 0.1$. We use ViT-S/16 and the training setup described in Appendix F, evaluating image classification, COCO retrieval, and hierarchical classification. The default configurations are highlighted.

RAOC angular allowance. The parameter β controls the angular allowance associated with a given radial gap. Smaller values impose stricter angular constraints, whereas larger values provide more flexibility. Table 10 examines the effect of this geometric parameter.

Table 10: **Sensitivity to the RAOC parameter β.** All other hyperparameters remain fixed at their default values.

Parameter	Classification		COCO R@5		Hierarchy	
β	ImageNet↑	Food-101↑	T→I↑	I→T↑	TIE↓	J ↑
2	41.31	49.61	55.02	68.71	3.69	0.77
4	41.18	49.82	55.31	68.97	3.67	0.77
8	41.54	50.13	55.77	69.42	3.62	0.78
16	41.12	49.74	55.18	68.84	3.71	0.77
32	41.36	49.96	55.41	69.06	3.66	0.77

Entailment weight. The weight λ_e balances hierarchical entailment supervision and semantic alignment. Table 11 evaluates the effect of different supervision strengths.

Table 11: **Sensitivity to the entailment weight λ_e.** All other hyperparameters remain fixed at their default values.

Parameter	Classification		COCO R@5		Hierarchy	
λ_e	ImageNet↑	Food-101↑	T→I↑	I→T↑	TIE↓	J ↑
0.0375	40.82	48.93	54.72	68.31	3.81	0.76
0.075	41.37	49.88	55.39	69.03	3.68	0.77
0.15	41.54	50.13	55.77	69.42	3.62	0.78
0.30	40.21	48.76	54.65	68.20	3.86	0.76
0.60	39.12	46.94	53.47	66.91	4.05	0.74

Norm regularization weight. The weight λ_n controls the radial scale regularization. Table 12 studies how different regularization strengths affect the learned hierarchy representation.

Table 12: **Sensitivity to the norm regularization weight λ_n.** All other hyperparameters remain fixed at their default values.

Parameter	Classification		COCO R@5		Hierarchy	
λ_n	ImageNet↑	Food-101↑	T→I↑	I→T↑	TIE↓	J ↑
0.0125	40.71	49.12	54.89	68.54	3.76	0.76
0.025	41.08	49.77	55.34	69.01	3.69	0.77
0.05	41.54	50.13	55.77	69.42	3.62	0.78
0.10	41.29	49.96	55.46	69.18	3.70	0.77
0.20	40.62	48.85	54.73	68.37	3.88	0.76

Radial margin scale. The parameter σ_r scales the radial margin during optimization. Table 13 evaluates whether performance depends strongly on this optimization scale.

Table 13: **Sensitivity to the radial margin scale σ_r.** All other hyperparameters remain fixed at their default values.

Parameter	Classification		COCO R@5		Hierarchy	
σ_r	ImageNet↑	Food-101↑	T→I↑	I→T↑	TIE↓	J ↑
0.025	41.08	49.74	55.21	68.92	3.70	0.77
0.05	41.32	50.01	55.54	69.18	3.65	0.77
0.10	41.54	50.13	55.77	69.42	3.62	0.78
0.20	41.46	50.04	55.68	69.31	3.64	0.78
0.40	41.38	49.91	55.59	69.20	3.67	0.77

K REPRODUCIBILITY

The main paper presents the objectives, methodology, and key experimental findings. Appendices F and H provide implementation details, training configurations, evaluation protocols, and computational settings. We will release code, checkpoints, configurations, data processing details, and evaluation resources, including scripts for reproducing RAOC-based geometric analyses and reported benchmark results.