HyGLoA: Dual-Granularity Hyperbolic Modeling with Local Geodesic-Ray Alignment for Zero-Shot Learning

Anonymous authors
Paper under double-blind review

Abstract

Zero-shot learning (ZSL) recognizes unseen classes by transferring knowledge from auxiliary semantics. Existing methods have explored various visual-semantic alignment schemes at either attribute or class granularity, yet most perform in Euclidean spaces. The polynomial volume growth of Euclidean space makes it difficult to allocate increasing capacity to finer semantic levels, while its geometry does not naturally encode the asymmetric specificity relations among classes, attributes, and visual evidence. We introduce **HyGLoA** (**Hy**perbolic **G**lobal–**Lo**cal **A**lignment), a dual-granularity framework that represents global class identity and local attribute evidence in two independent hyperbolic spaces under the Lorentz model. The global branch aligns the image CLS token with class-name prototypes, whereas the local branch projects patch tokens, aggregates them via Lorentzian centroid pooling, and aligns the resulting representation with attribute prototypes. Both branches use branch-specific direction–radius decoupled projections to control semantic direction and geodesic radius separately. To capture the asymmetric relation between partial local evidence and complete attribute semantics, we derive local geodesic-ray alignment from the zero-aperture limit of hyperbolic entailment. It removes cone-angle tolerance and penalizes deviation of the pooled local representation from the matched prototype's origin-facing geodesic ray. We further perform prototype bundle expansion on normalized semantic prototypes before hyperbolic lifting, improving angular discrimination without interfering with radial modeling. The two branch distributions undergo probability-level fusion for final prediction. Experiments on three popular benchmarks yield the highest generalized-ZSL harmonic mean among the compared methods, while ablations validate the principal design choices.

1 Introduction

Zero-shot learning (ZSL) recognizes unseen classes by transferring knowledge through semantics shared with seen categories. Early visual ZSL predominantly used human-defined attribute signatures and learned image–attribute compatibility (Lampert et al., 2009; 2014; Akata et al., 2013; 2015; Romera-Paredes & Torr, 2015; Xian et al., 2017), while semantic output codes, class-label word embeddings, and synthesized classifiers provided parallel class-level interfaces (Palatucci et al., 2009; Frome et al., 2013; Norouzi et al., 2014; Changpinyo et al., 2016). Attribute-based methods later evolved toward dense attention, attribute prototypes, and patch-level interaction, retaining attributes as localizable evidence (Huynh & Elhamifar, 2020; Xu et al., 2020; Chen et al., 2022b; Liu et al., 2025; Chen et al., 2025b). More recently, CLIP and prompt learning have made class names effective semantic anchors for zero-shot recognition (Radford et al., 2021; Zhou et al., 2022a;b; Khattak et al., 2023). Although CLIP-style protocols differ from conventional attribute-supervised ZSL, the two lines expose complementary granularities: class names compactly encode global category identity, whereas attributes decompose it into localized visual evidence. Existing methods typically emphasize one interface or couple both within Euclidean representations (Liu et al., 2025), leaving their granularity-specific geometric organization underexplored.

This geometric choice matters because most ZSL and CLIP-based methods perform visual–semantic matching with dot products or cosine similarities in Euclidean or normalized hyperspherical embeddings. Order Embeddings introduced an explicit formulation of images and language as an

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.

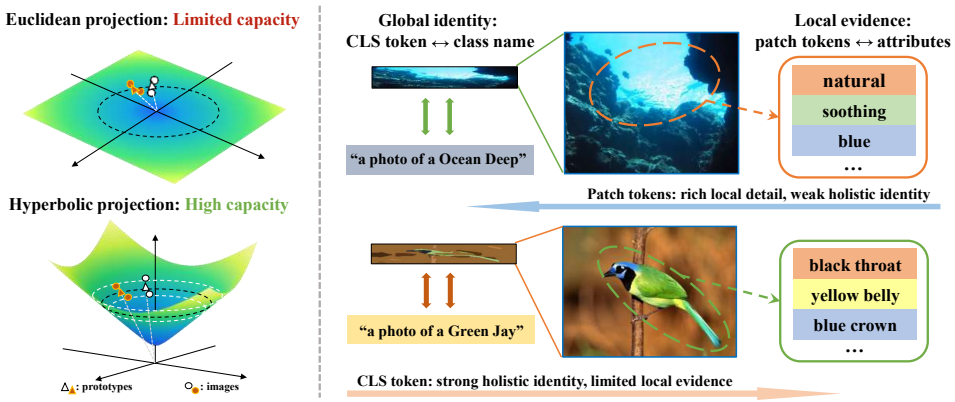


Figure 1: Conceptual motivation for dual-granularity hyperbolic modeling. *Left*: Euclidean volume grows polynomially with radius, which can crowd hierarchical representations, whereas hyperbolic volume grows exponentially and offers increasing capacity toward the boundary. *Right*: CLS tokens and class-name prototypes capture global identity, while patch tokens and attribute prototypes provide localized evidence. HyGLoA models these complementary granularities in separate hyperbolic spaces under the Lorentz model and combines their predictions.

asymmetric visual–semantic partial order (Vendrov et al., 2016), but Euclidean and normalized hyperspherical spaces offer limited radial structure: Euclidean volume grows only polynomially with distance from the origin, and unit normalization removes the radial degree of freedom that can encode semantic specificity. Hyperbolic space instead provides exponentially increasing capacity and a natural radial organization for hierarchical data (Nickel & Kiela, 2017; 2018); it can express asymmetric partial orders through hyperbolic entailment cones (Ganea et al., 2018). MERU brings this radial and entailment geometry to large-scale image–text representation learning, placing generic concepts near the origin and increasingly specific concepts farther away (Desai et al., 2023); HyCoCLIP further studies compositional entailment (Pal et al., 2025). These methods do not, however, address the distinct roles of global class identity and local attribute evidence in ZSL. As illustrated in Fig. 1, jointly exploiting these cues calls for both richer geometry and granularity-specific modeling.

We therefore propose **HyGLoA** (**Hy**perbolic **G**lobal–**Lo**cal **A**lignment), a dual-granularity ZSL framework with two independent hyperbolic spaces under the Lorentz model. Its global branch aligns an image CLS token with class-name prototypes to capture holistic identity. Its local branch embeds patch tokens, summarizes them using Lorentzian centroid pooling, and aligns the result with attribute-derived class prototypes to capture decomposable evidence. Each branch makes an independent prediction, followed by probability-level fusion. This separation allows each semantic granularity to acquire its own curvature-aware organization without forcing heterogeneous cues into a single shared embedding.

The local branch further needs to encode an asymmetric relation: an observed local representation of the image contains partial visual evidence, whereas the matched attribute prototype describes the class more completely. This follows the visual–semantic partial-order view of Order Embeddings (Vendrov et al., 2016), while using the geometry of hyperbolic entailment cones (Ganea et al., 2018) and their image–text adaptation in MERU (Desai et al., 2023). We derive *local geodesic-ray alignment* (LGRA) from a zero-aperture, direction-reversed limit, with the matched attribute prototype as the vertex. LGRA penalizes deviation of the pooled local representation from the prototype's origin-facing geodesic ray, favoring an organization in which partial visual evidence lies closer to the origin than the more complete semantic prototype. It thereby replaces tolerant cone membership with a cone-free axial constraint and removes the need to specify or learn a cone aperture.

HyGLoA also separates the two geometric roles of a hyperbolic representation in both branches. Branch-specific direction–radius decoupled projections learn semantic direction and radial specificity through separate mappings, rather than tying them to a single projection. In addition, we introduce *prototype bundle expansion* (PBE) before hyperbolic lifting. PBE spreads unit-normalized class-name and attribute prototypes in Euclidean angular space, improving their discrimination before geodesic radius is introduced by the Lorentz lift. Consequently, angular prototype separation and radial semantic organization can be optimized without conflating their functions. Experiments

on CUB, SUN, and AWA2 yield the best generalized-ZSL harmonic mean among the methods compared on all three benchmarks, while maintaining competitive conventional-ZSL accuracy; comprehensive ablations assess the contribution of each design.

Our main contributions are summarized as follows. (1) We introduce a dual-granularity framework that models global class-name semantics and local attribute semantics in separate Lorentz spaces and combines their predictions through probability-level fusion. (2) We propose Local Geodesic-Ray Alignment (LGRA), a zero-aperture, cone-free reformulation of hyperbolic entailment that constrains the pooled local representation toward the matched attribute prototype's origin-facing geodesic ray. (3) We design branch-specific direction–radius factorization to control semantic direction and geodesic radius through separate projections. (4) We develop Prototype Bundle Expansion (PBE) to redistribute normalized semantic prototypes in Euclidean angular space before hyperbolic lifting, improving prototype separation without interfering with radial modeling.

2 Related Work

2.1 Zero-Shot Learning

Zero-shot learning (ZSL) recognizes unseen classes by transferring knowledge through class-level semantic descriptors, including attributes, word embeddings, and textual descriptions (Larochelle et al., 2008; Palatucci et al., 2009). Generative methods use GANs (Su et al., 2022; Xian et al., 2018b; Zhu et al., 2018), VAEs (Narayan et al., 2020; Wang et al., 2018; Keshari et al., 2020; Chen et al., 2021b), or hybrid models (Narayan et al., 2020; Chen et al., 2021a) to synthesize visual features for unseen categories and train a classifier. RLVC (Hou et al., 2026) further explores reward-based optimization with class-wise visual cues. Embedding-based methods instead learn visual–semantic compatibility, with early approaches aligning holistic image features and class descriptors (Xian et al., 2016; Zhang et al., 2017; Akata et al., 2015).

Recent attribute-based methods improve local visual–semantic correspondence. TransZero++ (Chen et al., 2022a;b) uses cross-attribute-guided interaction, ZSLViT (Chen et al., 2024) progressively refines semantic-related tokens, and PSVMA+ (Liu et al., 2023; 2025) performs multi-granularity semantic–visual adaptation. SAGE (Zhang et al., 2026) introduces structured attribute-guided enhancement, while SVIP (Chen et al., 2025b) learns semantic patch selection using aggregated cross-layer attention as supervision. These advances motivate a complementary question: how should the semantic direction and radial coordinate of local evidence be parameterized? In an unfactorized embedding, both are derived from the same vector, even when the surrounding architecture contains multiple branches or granularity levels. HyGLoA explicitly predicts a semantic direction and a geodesic radius, making their roles in alignment and aggregation accessible to separate analysis at the coordinate level.

CLIP (Radford et al., 2021) and related approaches (Zhou et al., 2022a; Xian et al., 2018a) provide strong zero-shot transfer through global image–text alignment. Global class-name supervision, however, may not expose all localized attribute cues needed to distinguish visually similar categories. Conversely, attribute-guided local evidence need not fully preserve global object identity. We therefore retain both signals in a patch–attribute branch and a $[\mathrm{CLS}]$–class-name branch, combining their class probabilities through probability-level fusion. This complements local alignment with global class semantics without requiring the two prototype spaces to share an embedding dimension.

2.2 Hyperbolic Geometry in Visual Representation

Hyperbolic space has constant negative curvature and exponential volume growth, making it useful for representing hierarchical structure (Sarkar, 2011; Nickel & Kiela, 2017; Krioukov et al., 2010). Hyperbolic embeddings have also been studied for visual representation (Khrulkov et al., 2020) and zero-shot recognition (Liu et al., 2020). Order Embeddings (Vendrov et al., 2016) introduced a partial-order formulation of words, sentences, and images as a visual–semantic hierarchy. In hyperbolic vision–language learning, MERU (Desai et al., 2023) models this cross-modal hierarchy using Lorentz embeddings and entailment constraints; HyCoCLIP (Pal et al., 2025) additionally studies compositional hierarchies within and across modalities.

Our focus is the relation between pooled local visual evidence and dataset-defined attribute prototypes. A photograph may reveal only part of a class attribute signature, motivating a different radial ordering from the generic-text–specific-image relation used in caption-based alignment. HyGLoA investigates this task-specific organization through Local Geodesic-Ray Alignment (LGRA): the at-

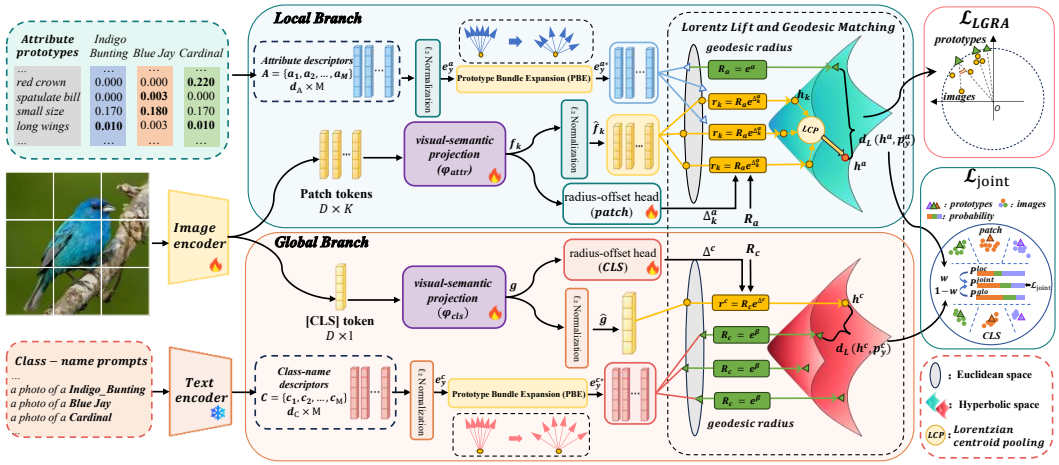


Figure 2: Overview of HyGLoA. PBE redistributes normalized attribute and class-name prototypes before branch-specific Lorentz lifts. LCP pools the patch points into $\mathbf{h}^a$, while the global branch produces $\mathbf{h}^c$. Their prototype-distance predictions are fused, and LGRA constrains only the local patch–attribute relation.

tribute prototype anchors an inward ray, and the pooled local representation is penalized for off-ray deviation. Inspired by MERU-style asymmetric entailment geometry, LGRA reverses the reference direction and takes the zero-aperture limit of an inward cone loss. It therefore supplies a directed ray constraint without prescribing a visual target geodesic radius or claiming the full partial-order properties of entailment cones. Unlike prior hyperbolic ZSL work (Liu et al., 2020), our focus is not a single visual–semantic embedding, but branch-specific geometry for global class-name and local attribute evidence. The following method section develops LGRA together with factorized visual embeddings, prototype bundle expansion, and dual-prototype probability-level fusion.

3 Method

3.1 Problem Formulation

Let $\mathcal{Y}^s$ and $\mathcal{Y}^u$ be disjoint seen and unseen class sets. Training uses $\mathcal{D}^s = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_s}$ with $y_i \in \mathcal{Y}^s$ and no unseen images. Each class y has an attribute vector $\boldsymbol{a}_y \in \mathbb{R}^{d_a}$ and a class-name embedding $\boldsymbol{c}_y = E_{\text{text}}(\mathcal{T}(y)) \in \mathbb{R}^{d_c}$, where $\mathcal{T}(y)$ is the fixed prompt "a photo of a {class}" instantiated with class y, and E_{text} is the frozen CLIP text encoder. Figure 2 denotes their collections by $\boldsymbol{A}$ and $\boldsymbol{C}$. ZSL predicts over $\mathcal{Y}^u$, whereas GZSL predicts over $\mathcal{Y}^s \cup \mathcal{Y}^u$.

3.2 Overview

Figure 2 summarizes the computation. The image encoder produces patch tokens $\boldsymbol{T}$ and a global token $\boldsymbol{t}^{\text{cls}}$. DRF lifts them with separately predicted directions and radii; LCP aggregates the patch points into $\mathbf{h}^a$, and the CLS pathway produces $\mathbf{h}^c$. In parallel, PBE transforms attribute and class-name directions into prototypes $\mathbf{p}_y^a$ and $\mathbf{p}_y^c$. The two branches predict from Lorentz distances and fuse their probabilities, while LGRA imposes the proposed inward local relation. The following subsections define these operations without repeating the branch flow.

3.3 Factorized Hyperbolic Visual Representations

Direction–Radius Factorization (DRF). Each branch uses a semantic direction projection and a scalar *radius-offset head*. The image encoder produces

$$(\boldsymbol{T}, \boldsymbol{t}^{\text{cls}}) = E_{\text{img}}(\boldsymbol{x}), \qquad \boldsymbol{T} = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K] \in \mathbb{R}^{D\times K}, \quad \boldsymbol{t}^{\text{cls}} \in \mathbb{R}^D, \tag{1}$$

where D and K denote token dimension and patch count. The learnable visual–semantic projections in Figure 2, $\phi_{\text{attr}} : \mathbb{R}^D \to \mathbb{R}^{d_a}$ and $\phi_{\text{cls}} : \mathbb{R}^D \to \mathbb{R}^{d_c}$, produce local and global features. With

learnable base log-radii α, β and $R_a = e^{\alpha}, R_c = e^{\beta}$, DRF computes

$$\begin{aligned} \boldsymbol{f}_k &= \phi_{\text{attr}}(\boldsymbol{t}_k), & \boldsymbol{g} &= \phi_{\text{cls}}(\boldsymbol{t}^{\text{cls}}), \\ \Delta_k^a &= \psi_{\text{rad}}(\boldsymbol{f}_k), & \Delta^c &= \psi_{\text{rad}}^{\text{cls}}(\boldsymbol{g}), \\ r_k &= R_a \exp(\Delta_k^a), & r^c &= R_c \exp(\Delta^c). \end{aligned} \tag{2}$$

Here ψ_{rad} and $\psi_{\text{rad}}^{\text{cls}}$ are learnable affine heads and Δ_k^a, Δ^c are input-dependent log-radius offsets. Directions are $\widehat{\boldsymbol{f}}_k = \boldsymbol{f}_k / \|\boldsymbol{f}_k\|_2$ and $\widehat{\boldsymbol{g}} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$; hence feature norms do not determine geodesic radius. Initialization and stability details are given in Appendix A.1.

Dual-branch Lorentz embedding. We use the Lorentz model $\mathbb{L}^d = \{\mathbf{p} = (p_0, \widetilde{\mathbf{p}}) : \langle \mathbf{p}, \mathbf{p} \rangle_{\mathbb{L}} = -1,\ p_0 > 0\}$ at curvature -1, with inner product $\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} = -p_0 q_0 + \widetilde{\mathbf{p}}^{\top} \widetilde{\mathbf{q}}$ and origin $\mathbf{o} = (1, \mathbf{0})$. Its distance and origin-centered radial lift are

$$\begin{aligned} d_{\mathbb{L}}(\mathbf{p}, \mathbf{q}) &= \operatorname{arcosh}(-\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}}), \\ \operatorname{Exp}_{\mathbf{o}}(0, r\widehat{\boldsymbol{v}}) &= (\cosh r, \sinh r\, \widehat{\boldsymbol{v}}), \qquad \|\widehat{\boldsymbol{v}}\|_2 = 1. \end{aligned} \tag{3}$$

The local patch points $\mathbf{h}_k$ and global image point $\mathbf{h}^c$ are therefore

$$\mathbf{h}_k = \operatorname{Exp}_{\mathbf{o}}(0, r_k \widehat{\boldsymbol{f}}_k) \in \mathbb{L}^{d_a}, \tag{4}$$

$$\mathbf{h}^c = \operatorname{Exp}_{\mathbf{o}}(0, r^c \widehat{\boldsymbol{g}}) \in \mathbb{L}^{d_c}. \tag{5}$$

Thus $d_{\mathbb{L}}(\mathbf{h}_k, \mathbf{o}) = r_k$ and $d_{\mathbb{L}}(\mathbf{h}^c, \mathbf{o}) = r^c$: the normalized and radius-offset paths in Figure 2 directly specify direction and geodesic radius.

Lorentzian Centroid Pooling (LCP). To aggregate distributed attribute evidence, we average the patch points in the ambient space and normalize their Minkowski mean back onto the manifold:

$$\overline{\mathbf{h}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{h}_k, \qquad \mathbf{h}^a = \frac{\overline{\mathbf{h}}}{\sqrt{-\langle \overline{\mathbf{h}}, \overline{\mathbf{h}} \rangle_{\mathbb{L}}}} \in \mathbb{L}^{d_a}. \tag{6}$$

Here $\overline{\mathbf{h}}$ is the ambient mean and $\mathbf{h}^a$ is the pooled local representation. Each spatial contribution is $\sinh r_k \widehat{\boldsymbol{f}}_k$, so DRF modulates aggregation without additional pooling parameters; further geometry is in Appendix A.1.

3.4 Semantic Prototypes and Probability-Level Fusion

Prototype bundle expansion. Shared attributes and prompt context can crowd prototype directions. For branch $b \in \{a, c\}$, let $\boldsymbol{e}_y^b$ be the normalized descriptor, $\boldsymbol{\mu}_b$ the seen-class mean axis, $\theta_{b,y}$ its angle to that axis, $\theta_b^{\max}$ the largest seen angle, and $\boldsymbol{\nu}_{b,y}$ the orthogonal unit component. PBE preserves this meridian while redistributing its angle:

$$\begin{aligned} \overline{\theta}_{b,y} &= m_b \theta_{b,y}, \qquad m_b = \frac{\pi/2}{\theta_b^{\max}}, \quad \theta_{b,y}^{\star} = \operatorname{clip}(\overline{\theta}_{b,y}, \epsilon, \pi/2), \\ \boldsymbol{e}_y^{b\star} &= \cos\theta_{b,y}^{\star} \boldsymbol{\mu}_b + \sin\theta_{b,y}^{\star} \boldsymbol{\nu}_{b,y}. \end{aligned} \tag{7}$$

The fixed multiplier m_b is computed from seen classes independently for each branch and is not learned; the expansion strength is fixed to one. The class prototype is

$$\mathbf{p}_y^b = \operatorname{Exp}_{\mathbf{o}}(0, R_b \boldsymbol{e}_y^{b\star}), \qquad b \in \{a, c\}. \tag{8}$$

Thus $\mathbf{p}_y^a$ and $\mathbf{p}_y^c$ reuse the corresponding DRF base radii rather than introduce new radial parameters. This shared branch-wise reference places visual offsets and semantic prototypes on a common radial scale while leaving instance-dependent visual variation to the ROH. Seen classes determine the fixed bundle statistics, which are reused for unseen prototypes. Exact construction and boundary handling are in Appendix A.2.

Probability-level fusion. We retain both semantic spaces and fuse their distributions over candidate set $\mathcal{Y}$. Learnable log-scales s_a, s_c convert distances to logits:

$$\begin{aligned} z_y^{\text{loc}} &= -e^{s_a} d_{\mathbb{L}}(\mathbf{h}^a, \mathbf{p}_y^a), \quad z_y^{\text{glo}} = -e^{s_c} d_{\mathbb{L}}(\mathbf{h}^c, \mathbf{p}_y^c), \\ P_y^q &= \frac{\exp z_y^q}{\sum_{j \in \mathcal{Y}} \exp z_j^q}, \qquad q \in \{\text{loc}, \text{glo}\}. \end{aligned} \tag{9}$$

A learnable scalar ω gives $w = \sigma(\omega) \in (0, 1)$. For $(\boldsymbol{x}, y)$, fusion and classification are

$$P_y^{\text{joint}} = w P_y^{\text{loc}} + (1 - w) P_y^{\text{glo}}, \qquad \mathcal{L}_{\text{joint}} = -\log P_y^{\text{joint}}. \tag{10}$$

Training uses $\mathcal{Y} = \mathcal{Y}^s$ and w is shared across samples. Both branches receive classification gradients through $\mathcal{L}_{\text{joint}}$, while the local branch is additionally regularized by LGRA below.

3.5 Cone-Free Local Geodesic-Ray Alignment

Because an image may reveal only part of a class attribute signature, we model the pooled patch representation $\mathbf{h}^a$ as less specific than the matched attribute prototype. LGRA reverses the reference direction of outward entailment cones (Ganea et al., 2018; Desai et al., 2023) and removes their positive angular tolerance, directly penalizing deviation from the prototype's origin-facing ray.

For an anchor $\mathbf{u} = (u_0, \widetilde{\mathbf{u}}) \neq \mathbf{o}$ and a query $\mathbf{q} = (q_0, \widetilde{\mathbf{q}}) \neq \mathbf{u}$ on the same manifold, let $\vartheta_{\leftarrow}(\mathbf{u}, \mathbf{q}) \in [0, \pi]$ be the angle at $\mathbf{u}$ between the geodesic toward the origin and the geodesic toward $\mathbf{q}$. The angle and local alignment loss are

$$\vartheta_{\leftarrow}(\mathbf{u}, \mathbf{q}) = \arccos\left(\frac{-q_0 - u_0\langle \mathbf{u}, \mathbf{q}\rangle_{\mathbb{L}}}{\|\widetilde{\mathbf{u}}\|_2 \sqrt{\langle \mathbf{u}, \mathbf{q}\rangle_{\mathbb{L}}^2 - 1}}\right), \qquad \mathcal{L}_{\mathrm{LGRA}} = \vartheta_{\leftarrow}(\mathbf{p}_y^a, \mathbf{h}^a). \tag{11}$$

For ground-truth y, $\mathbf{p}_y^a$ is the anchor and $\mathbf{h}^a$ the query, so their order matters. The loss is the zero-aperture inward-cone limit and does not prescribe a target visual geodesic radius. Appendix A.3 derives its zero-loss set; Appendix A.3 explains why the directed constraint is applied only to the patch–attribute branch and relates it to the radial behavior of Lorentz-distance classification.

3.6 Training Objective and Inference

The complete objective averages the classification and alignment losses over seen training pairs:

$$\mathcal{L} = \mathbb{E}_{(\boldsymbol{x}, y)\sim\mathcal{D}^s}\left[\mathcal{L}_{\mathrm{joint}} + \lambda_{\mathrm{LGRA}}\mathcal{L}_{\mathrm{LGRA}}\right], \qquad \lambda_{\mathrm{LGRA}} \geq 0. \tag{12}$$

Here $\lambda_{\mathrm{LGRA}} \geq 0$ is a fixed loss weight. We optimize all learnable quantities introduced above while keeping semantic annotations, the prompt, text encoder, seen-class bundle statistics, and curvature fixed. At inference, equation 8 constructs unseen prototypes and equation 9–equation 10 are evaluated over $\mathcal{Y}^u$ for ZSL or $\mathcal{Y}^s \cup \mathcal{Y}^u$ for GZSL. All remaining GZSL evaluation settings follow SVIP (Chen et al., 2025b) without modification; details are in Appendix A.4.

4 Experiments

4.1 Experimental Settings

We evaluate on CUB (Wah et al., 2011), SUN (Patterson & Hays, 2012), and AWA2 (Xian et al., 2018a), using the proposed splits adopted by f-CLSWGAN (Xian et al., 2018b): 150/50, 645/72, and 40/10 seen/unseen classes, respectively. Training uses only the designated seen-class training images. Following Xian et al. (2018a), we report average per-class top-1 accuracy. Conventional ZSL (CZSL) evaluates unseen-class images with predictions restricted to unseen classes ($T1$). In GZSL, predictions are made over the union of seen and unseen classes. We report average per-class top-1 accuracy separately for seen (S) and unseen (U) classes, and use their harmonic mean, $H = (2 \times S \times U)/(S + U)$, as the primary measure of balanced performance. Architecture initialization, optimization, numerical safeguards, and all dataset-specific settings are provided in Appendix B.1.

4.2 Comparison with State-of-the-Art Methods

We compare HyGLoA with state-of-the-art generative and embedding-based methods. The generative group covers approaches that synthesize unseen-class features, including CANZSL (Chen et al., 2020), SDGZSL (Chen et al., 2021c), CFSP (Cetin et al., 2022), SC-EGG (Hong et al., 2022), DSP (Chen et al., 2023a), TF-VAEGAN (Narayan et al., 2020), VADS (Hou et al., 2024), and ViFR (Chen et al., 2025a). Our embedding comparisons emphasize the visual–semantic relationships central to HyGLoA. RFF (Han et al., 2020) reduces feature redundancy, while TransZero++ (Chen et al., 2022a) models attribute–region interactions. I2DFormer (Naeem et al., 2022), DUET (Chen et al., 2023b), and I2MVFormer (Naeem et al., 2023) further investigate alignment with textual semantics. More closely related to our use of patches, ZSLViT (Chen et al., 2024), PSVMA+ (Liu et al., 2025), and SVIP (Chen et al., 2025b) study semantic-guided token learning, multi-granularity aggregation, and patch contextualization, respectively. VSPCN (Jiang et al., 2025) and SAGE (Zhang et al., 2026) complement these comparisons through prompt collaboration and structured attribute modeling. Baseline scores are taken from published reports except for the frozen CLIP row, evaluated under the same dataset splits and evaluation settings as HyGLoA; details of the CLIP-based comparisons are provided in Appendix B.2.

Table 1: Comparison with state-of-the-art methods on CUB, SUN, and AWA2 (%). $T1$ denotes CZSL accuracy; U, S, and H denote GZSL unseen, seen, and harmonic-mean accuracy. The CLIP row reports our reproduced results under the same evaluation settings as HyGLoA. † marks CLIP-based evaluations reported by SHIP (Wang et al., 2023): ‡ marks the reported results of APAN w/o ResNet configuration, which has a closer parameter-scale comparability to ours.

	Method	Venue	Backbone	CUB				SUN				AWA2			
				T1	U	S	H	T1	U	S	H	T1	U	S	H
Generative	CANZSL	WACV'20	Res101	60.6	47.9	58.1	52.5	60.1	46.8	35.0	40.0	68.9	49.7	70.2	58.2
	TF-VAEGAN†	ECCV'20	CLIP	–	21.1	**84.4**	34.0	–	–	–	–	–	43.7	**96.3**	60.1
	SDGZSL	ICCV'21	Res101	75.5	59.9	66.4	63.0	62.4	48.2	36.1	41.3	72.1	64.6	73.6	68.8
	CFSP	ICLR'22	Res101	–	51.6	60.4	55.7	–	48.2	36.5	41.5	–	51.3	78.0	61.8
	SC-EGG	IJCAI'22	Res101	75.1	64.1	73.6	68.5	69.2	45.1	43.6	44.3	78.2	60.9	89.3	72.4
	DSP	ICML'23	Res101	–	51.4	63.8	56.9	–	48.3	43.0	45.5	–	60.0	86.0	70.7
	VADS	CVPR'24	ViT	**86.8**	74.1	74.6	74.3	76.3	64.6	49.0	55.7	82.5	75.4	83.6	79.3
	ViFR	IJCV'25	Res101	74.5	63.9	72.0	67.6	69.2	51.3	40.0	44.7	77.8	68.2	78.9	73.2
Embedding	RFF	CVPR'20	Res101	–	52.6	56.6	54.6	–	45.7	38.6	41.9	–	59.8	75.1	66.5
	TransZero++	TPAMI'22	Res101	78.3	67.5	73.6	70.4	67.6	48.6	37.8	42.5	72.6	64.6	82.7	72.5
	I2DFormer	NeurIPS'22	ViT	45.4	35.3	57.6	43.8	–	–	–	–	76.4	66.8	76.8	71.5
	DUET	AAAI'23	ViT	72.3	62.9	72.8	67.5	64.4	45.7	45.8	45.8	69.9	63.7	84.7	72.7
	I2MVFormer	CVPR'23	ViT	42.1	32.4	63.1	42.8	–	–	–	–	73.6	66.6	82.9	73.8
	ZSLViT	CVPR'24	ViT	78.9	69.4	78.2	73.6	68.3	45.9	48.4	47.3	70.7	66.1	84.6	74.2
	PSVMA+	TPAMI'25	ViT	78.8	71.8	77.8	74.6	74.5	61.5	49.4	54.8	79.2	74.2	86.4	79.8
	SVIP	ICCV'25	ViT	79.8	72.1	78.1	75.0	71.6	53.7	48.0	50.7	69.8	65.4	87.7	74.9
	VSPCN	CVPR'25	ViT	80.6	72.8	78.9	75.7	75.3	59.4	49.1	53.8	76.6	71.8	84.3	77.6
	SAGE	AAAI'26	ViT	82.1	75.1	77.5	76.3	77.2	63.7	52.2	57.4	83.7	78.1	87.2	82.4
CLIP-based	CLIP	ICML'21	CLIP	71.3	56.6	55.2	55.9	85.6	54.9	51.9	53.4	**97.1**	89.7	92.2	90.9
	CoOp†	IJCV'22	CLIP	–	49.2	63.8	55.6	–	–	–	–	–	72.7	95.3	82.5
	CoOp+SHIP	ICCV'23	CLIP	–	55.3	58.9	57.1	–	–	–	–	–	84.1	94.4	89.0
	APAN‡	TNNLS'25	CLIP	81.3	64.8	74.6	69.4	88.3	71.6	**62.0**	66.5	96.9	89.6	92.2	90.9
	HyGLoA	Ours	CLIP	85.2	**77.6**	79.1	**78.3**	**88.8**	**76.0**	61.0	**67.7**	97.0	**95.3**	95.1	**95.2**

Table 2: Key ablations (%). P/C denote patch/CLS branches, Hyp. denotes hyperbolic embedding, and ←/→ denote inward/outward LGRA. $T1$ and H are CZSL and GZSL metrics; ✓/– mark enabled/inapplicable entries, bold marks maxima, and IDs match the appendix.

ID	Components							CUB		SUN		AWA2	
	P	C	Hyp.	PBE	DRF	LCP	LGRA	$T1$	H	$T1$	H	$T1$	H
a	✓	✓	–	✓	–	–	–	82.8	75.6	87.7	65.5	96.7	94.4
b	✓	✓	✓	✓	✓	✓	–	83.8	75.2	88.3	67.0	96.9	95.1
c	✓	–	✓	✓	✓	✓	←	81.2	74.8	77.5	53.9	72.0	73.4
d	–	✓	✓	✓	✓	–	–	72.1	65.5	87.9	65.6	96.8	94.7
e	✓	✓	✓	–	✓	✓	←	84.7	77.9	87.9	67.0	96.9	94.5
f	✓	✓	✓	✓	✓	✓	→	83.8	75.9	88.3	67.3	96.9	95.1
g	✓	✓	✓	✓	–	✓	←	84.9	78.1	88.6	67.5	96.9	95.0
h	✓	✓	✓	✓	✓	–	←	84.0	76.8	88.3	66.7	96.8	95.1
Full	✓	✓	✓	✓	✓	✓	←	**85.2**	**78.3**	**88.8**	**67.7**	**97.0**	**95.2**

HyGLoA achieves the highest GZSL H among the methods compared in Table 1: **78.3/67.7/95.2** on CUB/SUN/AWA2, exceeding SAGE (Zhang et al., 2026), the strongest non-CLIP embedding baseline listed, by 2.0/10.3/12.8 points. These system-level comparisons involve different pretraining; our ablations assess architectural contributions. Relative to our frozen CLIP baseline, HyGLoA improves H by 22.4/14.3/4.3 points. On AWA2, CLIP is 0.1 points higher in CZSL $T1$, whereas HyGLoA obtains a better seen–unseen balance and a 4.3-point higher H. Against the parameter-comparable APAN configuration (Xie et al., 2025), H improves by 1.2/4.3 points on SUN/AWA2; HyGLoA also leads the listed CZSL results on SUN.

4.3 Ablation Studies of HyGLoA Components

Table 2 evaluates the key components of HyGLoA using CZSL $T1$ and GZSL H, with complete results and variant definitions in Appendix B.3. **Granularity complementarity.** Full improves H

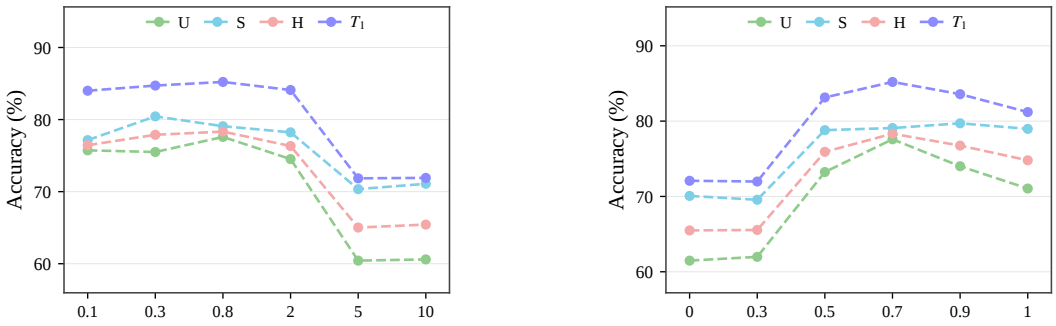


Figure 3: Sensitivity analysis of λ_{LGRA} (left) and w_0 (right) on CUB.

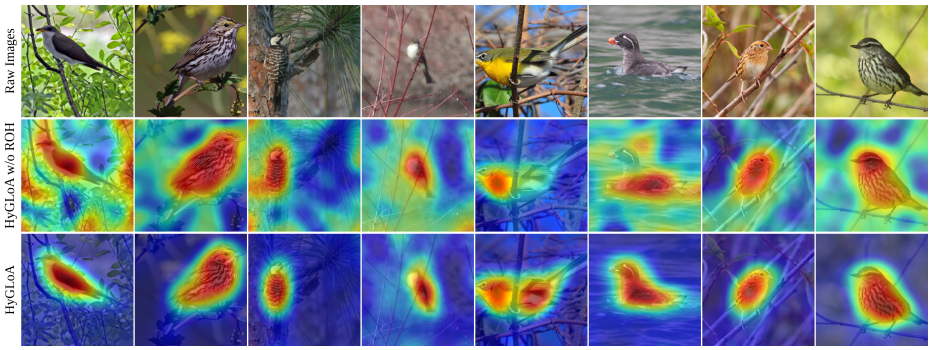


Figure 4: Patch-to-prototype geometric response maps. From top to bottom, the rows show the original images, HyGLoA w/o ROH, and HyGLoA. ROH denotes the radius-offset head, the scalar projection predicting a log-radius offset. Warm colours indicate larger positive responses $[r_k \cos\theta_k]_+$ to the matched class attribute prototype; each map is normalized within its own image.

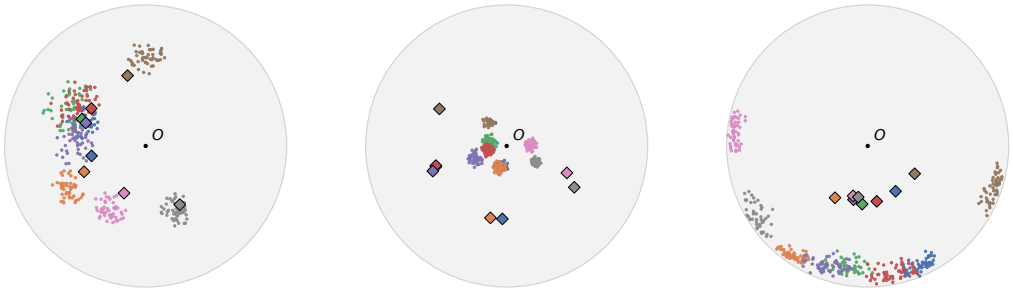
over CLS only (d) by 12.8/2.1/0.5 points on CUB/SUN/AWA2 and also outperforms patch only (c). **Hyperbolic local ordering.** The complete model exceeds the Euclidean dual-branch baseline (a) by 2.7/2.2/0.8 points at the displayed precision. This is a system-level comparison because the Euclidean variant also removes the Lorentz-specific DRF and pooling operations; it should not be interpreted as isolating curvature alone. On CUB, however, removing LGRA (b) lowers H from 78.3 to 75.2, below the Euclidean result (75.6): distance classification can separate classes without establishing the intended patch-inside-prototype ordering. The outward layout in Figure 5 and the weaker outward-LGRA result (f, 75.9) support the proposed inward constraint. **Other components.** Removing PBE reduces H on every dataset; replacing ROH by the projected norm reduces it consistently by 0.2 points; and LCP gives its largest gains on CUB/SUN. Single-branch behavior and the remaining variants are analyzed in Appendix B.3.

4.4 Hyperparameter Analysis

Figure 3 shows that CUB is stable for $\lambda_{\text{LGRA}} \in [0.1, 2]$ and $w_0 \in [0.5, 0.9]$, with the best GZSL H at $\lambda_{\text{LGRA}} = 0.8$ and $w_0 = 0.7$. When λ_{LGRA} increases to 5 or 10, H falls to 65.0/65.4, near CLS only (65.5), because overly strong alignment collapses patch radii before useful local evidence is learned. The learnable fusion weight reaches 0.753 at the best checkpoint: CUB favors local evidence, but the interior optimum shows that neither granularity should dominate. Additional SUN and learning-rate analyses are in Appendix B.4.

4.5 Qualitative Analysis

Although removing the radius-offset head changes H by only 0.2 points, Figure 4 reveals a clearer representational effect. The displayed response $[r_k \cos\theta_k]_+$ combines prototype directional agreement with learned patch radius and is normalized per image. HyGLoA concentrates on informative



(a) HyGLoA w/o LGRA (b) HyGLoA-Patch (c) HyGLoA-CLS

Figure 5: LGRA-induced visual–prototype layouts on CUB visualized by HoroPCA. (a) Pooled patch representations without LGRA; (b) pooled patch representations with LGRA; (c) CLS representations from the same full model. Panels (a)–(b) use attribute prototypes, whereas panel (c) uses class-name prototypes. LGRA directly constrains only the patch branch.

foreground regions even in complex scenes, whereas deriving direction and radius from the same feature vector yields broader background responses. This suggests that independent radial control primarily improves the organization and spatial selectivity of local evidence. Its classification gain can remain small because directional alignment, the global branch, and probability-level fusion can partly compensate when direction and radius are coupled through one projection; moreover, semantically specific patches are not necessarily class-discriminative and may contain shared attributes or irrelevant detail.

Figure 5(a)–(b) further shows that LGRA changes the local visual–prototype ordering: pooled patch representations move from predominantly outside the attribute prototypes to predominantly inside them, and form more compact groups. An original-space radial-residual diagnostic in Appendix B.5.3 independently corroborates this ordering. Together with the recovery from 75.2 to 78.3, these observations support the intended hierarchy in which partial patch evidence lies inside its more complete attribute prototype. The unconstrained patch and CLS layouts also exhibit the outward tendency discussed in the appendix, without implying a universal radial update. The appendix further provides the exact response construction and HoroPCA protocol, pre-Lorentz single-attribute maps, CO-SNE neighborhoods for both branches, and PBE prototype-angle statistics.

5 Conclusion

We presented HyGLoA, a dual-granularity framework that brings hyperbolic modeling to global and local visual–semantic alignment in zero-shot learning. Direction–radius factorization, prototype bundle expansion, Lorentzian centroid pooling, and LGRA jointly organize patch-level attribute evidence while retaining global [CLS]–class-name transfer. Experiments on three benchmarks support the complete design; in our validation, the geometric diagnostics further support the asymmetric local relation in which partial patch evidence lies inside its more complete attribute prototype. This work remains an initial exploration of hyperbolic vision–language modeling for ZSL. Our probability-level fusion is a simple strategy-level combination and is unlikely to be optimal: local patch–attribute and global [CLS]–class-name pairs exhibit different cross-modal hierarchical relations, yet their interactions are not modeled geometrically. We also do not fully investigate how global visual representations and class-name semantics should be adapted specifically to ZSL. More principled cross-granularity fusion and global semantic adaptation are therefore natural directions for future work.

AI use statement

In this work, we used generative AI tools to aid and polish the writing and to assist in drafting selected portions of the manuscript based on content and technical information provided by the authors. The tools were used primarily for grammatical correction, language polishing, improving clarity and readability, and suggesting alternative wording or organization for selected sentences and paragraphs. The scientific ideas, methodology, experimental design, analysis, and conclusions were developed by the authors. All AI-assisted text was carefully reviewed, revised, and verified by the

authors. We did not use generative AI tools for literature discovery, research ideation or execution, synthetic data generation, or proving mathematical claims. The authors take full responsibility for the final content of this work, including all text, claims, results, and conclusions.

REFERENCES

Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 819–826, 2013.

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, 2015.

Samet Cetin, Orhun Bugra Baran, and Ramazan Gokberk Cinbis. Closed-form sample probing for learning generative models in zero-shot learning. In *International Conference on Learning Representations*, 2022.

Ines Chami, Albert Gu, Dat P. Nguyen, and Christopher Ré. HoroPCA: Hyperbolic dimensionality reduction via horospherical projections. In *International Conference on Machine Learning*, pp. 1419–1429, 2021.

Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5327–5336, 2016.

Shiming Chen, Wen Wang, Bing Xia, et al. FREE: Feature refinement for generalized zero-shot learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 122–131, 2021a.

Shiming Chen, Guo-Sen Xie, Yang Liu, et al. HSVA: Hierarchical semantic-visual adaptation for zero-shot learning. In *Advances in Neural Information Processing Systems*, pp. 16622–16634, 2021b.

Shiming Chen, Zhi Hong, Wei Hou, et al. TransZero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:12844–12861, 2022a.

Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. TransZero: Attribute-guided transformer for zero-shot learning. In *AAAI Conference on Artificial Intelligence*, pp. 330–338, 2022b.

Shiming Chen, Wenjin Hou, Ziming Hong, Xiaohan Ding, Yibing Song, Xinge You, Tongliang Liu, and Kun Zhang. Evolving semantic prototype improves generative zero-shot learning. In *International Conference on Machine Learning*, pp. 4611–4622, 2023a.

Shiming Chen, Wei Hou, Salman Khan, et al. Progressive semantic-guided vision transformer for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23964–23974, 2024.

Shiming Chen, Ziming Hong, Xinge You, and Ling Shao. Semantics-conditioned generative zero-shot learning via feature refinement. *International Journal of Computer Vision*, 133(7):4504–4521, 2025a.

Zhi Chen, Zecheng Zhao, Jingcai Guo, Jingjing Li, and Zi Huang. SVIP: Semantically contextualized visual patches for zero-shot learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 3346–3356, 2025b.

Zhiyuan Chen, Jingjing Li, Yadan Luo, et al. CANZSL: Cycle-consistent adversarial networks for zero-shot learning from natural language. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 874–883, 2020.

Zhiyuan Chen, Yadan Luo, Rui Qiu, et al. Semantics disentangling for generalized zero-shot learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 8692–8700, 2021c.

Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z. Pan, and Huajun Chen. DUET: Cross-modal semantic grounding for contrastive zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 405–413, 2023b.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731, 2023.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pp. 2121–2129, 2013.

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

Yunhui Guo, Haoran Guo, and Stella X. Yu. CO-SNE: Dimensionality reduction and visualization for hyperbolic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21–30, 2022.

Zongyan Han, Zhiyuan Fu, and Jian Yang. Learning the redundancy-free features for generalized zero-shot object recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12862–12871, 2020.

Ziming Hong, Shiming Chen, Guo-Sen Xie, Wenhan Yang, Jian Zhao, Yuanjie Shao, Qinmu Peng, and Xinge You. Semantic compression embedding for generative zero-shot learning. In *International Joint Conference on Artificial Intelligence*, pp. 956–963, 2022.

Wei Hou, Xia Sun, and Huan Fan. Incentivizing generative zero-shot learning via outcome-reward reinforcement learning with visual cues. *arXiv preprint arXiv:2603.21138*, 2026.

Wenjin Hou, Shiming Chen, Shuhuang Chen, Ziming Hong, Yan Wang, Xuetao Feng, Salman Khan, Fahad Shahbaz Khan, and Xinge You. Visual-augmented dynamic semantic prototype for generative zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23627–23637, 2024.

Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4483–4493, 2020.

Huajie Jiang, Zhengxian Li, Xiaohan Yu, Yongli Hu, Baocai Yin, Jian Yang, and Yuankai Qi. Visual and semantic prompt collaboration for generalized zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20275–20285, 2025.

Rishabh Keshari, Richa Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13297–13305, 2020.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Valentin Khrulkov, Lili Mirvakhabova, Evgeniya Ustinova, et al. Hyperbolic image embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6417–6427, 2020.

Diederik P. Kingma, Jimmy Ba, and J. Adam. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Physical Review E*, 82:036106, 2010.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 951–958, 2009.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence*, 2008.

Man Liu, Huihui Bai, Feng Li, Chunjie Zhang, Yunchao Wei, Meng Wang, Tat-Seng Chua, and Yao Zhao. PSVMA+: Exploring multi-granularity semantic-visual adaption for generalized zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):51–66, 2025.

Mingsheng Liu, Fang Li, Chao Zhang, et al. Progressive semantic-visual mutual adaption for generalized zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15337–15346, 2023.

Shichen Liu, Jing Chen, Liang Pan, et al. Hyperbolic visual embedding learning for zero-shot recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9270–9278, 2020.

Muhammad Ferjad Naeem, Yongqin Xian, Luc Van Gool, and Federico Tombari. I2DFormer: Learning image to document attention for zero-shot image classification. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12283–12294, 2022.

Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2MVFormer: Large language model generated multi-view document supervision for zero-shot image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15169–15179, 2023.

Sanath Narayan, Akshay Gupta, Fahad Shahbaz Khan, et al. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision*, pp. 479–495, 2020.

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

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

Mohammad Norouzi, Tomás Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Gregory S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations*, 2014.

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 *International Conference on Learning Representations*, pp. 87371–87399, 2025.

Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, pp. 1410–1418, 2009.

Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758, 2012.

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 *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pp. 2152–2161, 2015.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366, 2011.

Hao Su, Jingjing Li, Zhen Chen, et al. Distinguishing unseen from seen for generalized zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7875–7884, 2022.

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.

Wen Wang, Yunchen Pu, Vikas Verma, et al. Zero-shot learning via class-conditioned deep generative models. In *AAAI Conference on Artificial Intelligence*, 2018.

Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. Improving zero-shot generalization for CLIP with synthesized prompts. *arXiv preprint arXiv:2307.07397*, 2023.

Yongqin Xian, Zeynep Akata, Gaurav Sharma, et al. Latent embeddings for zero-shot classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 69–77, 2016.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning—the good, the bad and the ugly. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4582–4591, 2017.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2251–2265, 2018a.

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5542–5551, 2018b.

Guo-Sen Xie, Junyi Li, Ting Guo, Xiangbo Shu, Fang Zhao, Zheng Zhang, and Ling Shao. Attribute prompt alignment network for zero-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(12):20457–20463, 2025.

Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *Advances in Neural Information Processing Systems*, pp. 21969–21980, 2020.

Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3010–3019, 2017.

Zhaoyi Zhang, Lin Sun, and Peng Lyu. SAGE: Structured attribute-guided enhancement for GZSL. In *AAAI Conference on Artificial Intelligence*, pp. 28564–28572, 2026.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022b.

Yizhe Zhu, Mohamed Elhoseiny, Bing Liu, et al. A generative adversarial approach for zero-shot learning from noisy texts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1004–1013, 2018.

A Additional Method Details

A.1 Direction–Radius Factorization and Lorentzian Centroid Pooling

For curvature $-\kappa$, $\kappa > 0$, the Lorentz manifold satisfies $\langle \mathbf{p}, \mathbf{p} \rangle_{\mathbb{L}} = -1/\kappa$ and $p_0 > 0$. With origin $\mathbf{o}_\kappa = (1/\sqrt{\kappa}, \mathbf{0})$, unit spatial direction $\boldsymbol{n} \in \mathbb{R}^d$, and geodesic radius $r \geq 0$, its radial parameterization and distance d_κ are

$$\mathrm{Exp}_{\mathbf{o}_\kappa}(0, r\boldsymbol{n}) = \frac{1}{\sqrt{\kappa}}\big(\cosh(\sqrt{\kappa}r), \sinh(\sqrt{\kappa}r)\boldsymbol{n}\big), \quad d_\kappa(\mathbf{p}, \mathbf{q}) = \frac{\operatorname{arcosh}(-\kappa\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}})}{\sqrt{\kappa}}. \tag{13}$$

HyGLoA fixes $\kappa = 1$. For two points with radii r_i, r_j and directions $\boldsymbol{n}_i, \boldsymbol{n}_j$,

$$\cosh d_{\mathbb{L}}(\mathbf{h}_i, \mathbf{h}_j) = \cosh r_i \cosh r_j - \sinh r_i \sinh r_j \, \boldsymbol{n}_i^\top \boldsymbol{n}_j. \tag{14}$$

This identity exposes the radial modulation of angular separation in DRF. The radius-offset heads add to the base log-radii. In the local branch, $\Delta_k^a = \psi_{\mathrm{rad}}(\boldsymbol{f}_k)$ and $r_k = R_a \exp(\Delta_k^a)$ with $R_a = e^\alpha$; the global branch analogously uses $\Delta^c = \psi_{\mathrm{rad}}^{\mathrm{cls}}(\boldsymbol{g})$ and $r^c = R_c \exp(\Delta^c)$ with $R_c = e^\beta$. The head outputs are therefore additive log-radius offsets and multiplicative adjustments of the radii. The base log-radius parameters and both affine offset heads are learned. The heads are initialized with zero weights and biases, so every visual radius initially equals its branch's base radius. Because the base radius is also used by the semantic prototypes, it remains distinct from the visual-only affine bias.

LCP is the normalized Minkowski mean in equation 6. Its spatial component before normalization is $K^{-1} \sum_k \sinh r_k \, \widehat{\boldsymbol{f}}_k$. Its weights grow nonlinearly with learned radius, while misaligned directions can cancel. For a common patch radius r, writing $\overline{\boldsymbol{n}} = K^{-1} \sum_k \widehat{\boldsymbol{f}}_k$ gives $\tanh d_{\mathbb{L}}(\mathbf{h}^a, \mathbf{o}) = \tanh r \, \|\overline{\boldsymbol{n}}\|_2 \leq \tanh r$. Thus pooling can contract radial depth even before any hierarchy is imposed. Following the radial convention used by MERU (Desai et al., 2023), greater geodesic radius represents greater semantic specificity. Specificity is not, however, equivalent to class-discriminative usefulness: a specific patch may encode a shared attribute, background detail, or other non-diagnostic content.

A.2 Construction of Prototype Bundle Expansion

For $b \in \{a, c\}$, let $\boldsymbol{e}_y^b$ be the normalized semantic descriptor of class y from the main text. The branch-specific bundle axis $\boldsymbol{\mu}_b$, angle $\theta_{b,y}$, largest seen-class angle $\theta_b^{\max}$, and unit meridian $\boldsymbol{\nu}_{b,y}$ are

$$\begin{aligned} \boldsymbol{\mu}_b &= \frac{\sum_{y \in \mathcal{Y}^s} \boldsymbol{e}_y^b}{\| \sum_{y \in \mathcal{Y}^s} \boldsymbol{e}_y^b \|_2}, \quad \theta_{b,y} = \arccos(\boldsymbol{e}_y^{b\top} \boldsymbol{\mu}_b), \\ \theta_b^{\max} &= \max_{y \in \mathcal{Y}^s} \theta_{b,y}, \qquad \boldsymbol{\nu}_{b,y} = \frac{\boldsymbol{e}_y^b - (\boldsymbol{e}_y^{b\top} \boldsymbol{\mu}_b)\boldsymbol{\mu}_b}{\| \boldsymbol{e}_y^b - (\boldsymbol{e}_y^{b\top} \boldsymbol{\mu}_b)\boldsymbol{\mu}_b \|_2}. \end{aligned} \tag{15}$$

equation 7 preserves this meridian and changes only the angle to the bundle axis. Orthogonality of $\boldsymbol{\mu}_b$ and $\boldsymbol{\nu}_{b,y}$ ensures $\|\boldsymbol{e}_y^{b\star}\|_2 = 1$. In the full model, the expansion strength is fixed to one and the automatic multiplier is $m_b = (\pi/2)/\theta_b^{\max}$; it is computed separately for the patch and CLS prototype families and is not learned. Thus each angle is expanded as $\overline{\theta}_{b,y} = m_b \theta_{b,y} = (\pi/2)(\theta_{b,y}/\theta_b^{\max})$ and then clipped to $[\epsilon, \pi/2]$. The construction assumes a nonzero mean and $\theta_b^{\max} > 0$; at $\theta_{b,y} = 0$, the mathematical continuous extension is $\boldsymbol{e}_y^{b\star} = \boldsymbol{\mu}_b$, while the implementation uses ϵ-stabilized normalization. The map redistributes angles rather than guaranteeing an increase in every pairwise separation. Only seen classes determine $\boldsymbol{\mu}_b$ and $\theta_b^{\max}$. An unseen prototype uses those same statistics; ratios above one are handled by the same angular clipping and do not affect the seen-only training objective.

A.3 Derivation and Geometry of LGRA

Angle at the prototype. Take an anchor $\mathbf{u} \neq \mathbf{o}$, a query $\mathbf{q} \neq \mathbf{u}$, and $C = -\langle \mathbf{u}, \mathbf{q} \rangle_{\mathbb{L}} = \cosh d_{\mathbb{L}}(\mathbf{u}, \mathbf{q}) > 1$. Unit tangent vectors at $\mathbf{u}$ pointing toward the origin and the query are

$$\boldsymbol{v}_{\mathrm{in}} = \frac{\mathbf{o} - u_0 \mathbf{u}}{\|\widetilde{\mathbf{u}}\|_2}, \qquad \boldsymbol{v}_{\mathbf{q}} = \frac{\mathbf{q} - C\mathbf{u}}{\sqrt{C^2 - 1}}. \tag{16}$$

Their Lorentz inner product equals the cosine in equation 11. The outward tangent is $-\boldsymbol{v}_{\mathrm{in}}$. Writing $\vartheta_{\rightarrow}$ for the angle to this outward reference gives $\vartheta_{\rightarrow} = \pi - \vartheta_{\leftarrow}$. Conventional entailment cones use an angular aperture $A(\mathbf{u}) > 0$ and the hinge $[\vartheta_{\rightarrow} - A(\mathbf{u})]_+$, where $[z]_+ = \max(0, z)$ (Ganea et al., 2018; Desai et al., 2023). Reversing the reference direction and scaling the aperture by $\epsilon > 0$ yields the limit

$$\lim_{\epsilon \downarrow 0} [\vartheta_{\leftarrow}(\mathbf{u}, \mathbf{q}) - \epsilon A(\mathbf{u})]_+ = \vartheta_{\leftarrow}(\mathbf{u}, \mathbf{q}). \tag{17}$$

Consequently, LGRA removes the positive angular tolerance of cone inclusion. It is a directed ray penalty; we do not claim that it retains the full partial-order properties of an entailment-cone construction.

Zero-loss set and radial scope. Writing $\mathbf{u} = (\cosh R, \sinh R\, \boldsymbol{n})$ with anchor radius $R = d_{\mathbb{L}}(\mathbf{u}, \mathbf{o})$ and unit direction $\boldsymbol{n}$, the inward unit-speed geodesic $\boldsymbol{\gamma}(t)$ is parameterized by distance t traveled from $\mathbf{u}$:

$$\boldsymbol{\gamma}(t) = \cosh t\, \mathbf{u} + \sinh t\, \boldsymbol{v}_{\mathrm{in}} = \big(\cosh(R - t), \sinh(R - t)\boldsymbol{n}\big), \qquad t > 0. \tag{18}$$

LGRA is zero precisely for $\mathbf{q} = \boldsymbol{\gamma}(t)$ on this ray. For $0 < t < R$, the query shares the prototype's origin-based direction and has radius $R - t$. The ray reaches the origin at $t = R$ and continues beyond it for $t > R$. Thus zero angular loss alone does not restrict the query to the segment between the origin and the prototype, nor impose a global radial upper bound. At the coincident point $\mathbf{q} = \mathbf{u}$, the query direction is undefined; the zero-loss statement excludes this singular case.

The positive geodesic distance on the same origin ray is $d_{\mathbb{L}}(\mathbf{h}^a, \mathbf{p}_y^a) = |d_{\mathbb{L}}(\mathbf{h}^a, \mathbf{o}) - R_a|$. Classification compares this distance with distances to competing prototypes, whereas LGRA penalizes deviation from the inward reference ray. Their joint optimization supplies a radial reference without a manually assigned visual geodesic radius. Because the angle is measured at the prototype, it depends on relative position and may produce radial as well as angular gradients. The geometry therefore supplies a conditional radial mechanism rather than a universal monotonicity theorem for cross-entropy or negative curvature. The geometric diagnostics in Section 4.5 verify that the learned model realizes the intended inside-prototype organization.

Why LGRA is restricted to the patch branch. The inward relation modeled by LGRA is motivated by a partial local observation relative to a class-level attribute signature. The global [CLS] representation instead summarizes instance-level visual content and is matched to a category-level class-name prototype; the same partial-attribute ordering therefore does not directly transfer to this branch.

The classification behavior provides a second, complementary reason. For a visual point $\mathbf{h}(r)$ moving radially along a fixed direction, define the distance margin against a competing class j as $m_j(r) = d_{\mathbb{L}}(\mathbf{h}(r), \mathbf{p}_j) - d_{\mathbb{L}}(\mathbf{h}(r), \mathbf{p}_y)$. With negative Lorentz-distance logits of scale $\tau > 0$, the corresponding branch-wise cross-entropy can be written as

$$\mathcal{L}_{\mathrm{CE}}(r) = \log\left(1 + \sum_{j \neq y} \exp[-\tau m_j(r)]\right). \tag{19}$$

Consequently, outward motion is favored whenever it increases the relevant distance margins. This is a conditional geometric effect rather than a monotone property of hyperbolic cross-entropy: the sign of the radial gradient depends on the image direction, prototype configuration, and current radius. Our model uses the fused objective $\mathcal{L}_{\mathrm{joint}}$ rather than separate branch-wise cross-entropies, but the same distance-margin mechanism acts through each branch probability. The radial analysis in Section 4 empirically shows this outward tendency for branches without direct LGRA supervision, including the global visual representations relative to their class-name prototypes. This ordering is also compatible with the general-to-specific text–image hierarchy considered by Desai et al. (2023), but that work does not establish the radial gradient of our classification objective. We therefore leave the global radial configuration to $\mathcal{L}_{\mathrm{joint}}$ and reserve explicit inward geodesic-ray alignment for local attribute evidence.

A.4 GZSL Evaluation Protocol

We follow the GZSL evaluation protocol of SVIP (Chen et al., 2025b) without modification. After computing P_y^{joint} over the joint seen–unseen candidate set, prediction is

$$\widehat{y}_{\text{GZSL}} = \arg\max_{y \in \mathcal{Y}^s \cup \mathcal{Y}^u} \left[P_y^{\text{joint}} + \delta \, \mathbb{I}[y \in \mathcal{Y}^u] \right], \tag{20}$$

where δ is the inference-time offset and $\mathbb{I}[\cdot]$ is the indicator function. The offset is not part of the training objective.

B Additional Experimental Details and Results

B.1 Datasets and Implementation

CUB (Wah et al., 2011) contains 11,788 images from 200 bird species with 312-dimensional class attributes; SUN (Patterson & Hays, 2012) has 14,340 images from 717 scene categories with 102-dimensional attributes; and AWA2 (Xian et al., 2018a) contains 37,322 images from 50 animal categories with 85-dimensional attributes. We use the designated train/test partitions of the proposed split protocol (Xian et al., 2018a;b), without training on unseen-class images. Average per-class accuracy is $A(\mathcal{C}) = |\mathcal{C}|^{-1} \sum_{c \in \mathcal{C}} n_c^{\text{correct}} / n_c$, where $\mathcal{C}$ is the evaluated class set, n_c is its number of test images for class c, and n_c^{correct} counts correct predictions. GZSL S and U use the joint seen/unseen candidate set.

We initialize from CLIP ViT-B/16. Its image encoder consists of a ViT backbone followed by a projection into the joint image–text embedding space. Only the CLS branch's visual–semantic projection is initialized from the pretrained projection; the patch counterpart is randomly initialized. Separate learning rates accommodate the newly learned patch head, pretrained CLS head, and backbone (Table 3). The image backbone is fine-tuned, whereas the text encoder remains frozen throughout training. Class-name prototypes use the fixed prompt "a photo of a {class}".

Each dataset is trained for 40 epochs on one of three NVIDIA GeForce RTX 3090 GPUs (24 GB), with batch size 64 per dataset run. We use Adam (Kingma et al., 2014) with $(\beta_1, \beta_2) = (0.5, 0.999)$ and zero weight decay; the learning-rate scheduler and remaining training protocol follow SVIP (Chen et al., 2025b). In Table 3, λ_{LGRA} weights the patch alignment loss, and w_0 initializes the learnable patch weight w in probability-level fusion. Specifically, we initialize $\omega_0 = \log\big(w_0/(1-w_0)\big)$ so that $\sigma(\omega_0) = w_0$; ω is subsequently optimized with the rest of the model. At the best-GZSL-H checkpoints, the learned patch weights are 0.753, 0.481, and 0.107 on CUB, SUN, and AWA2, respectively. We additionally analyze the three learning rates on CUB in Figure 7, assessing the sensitivity of these settings.

Table 3: Dataset-specific training settings. LR denotes learning rate.

Dataset	Patch LR	CLS LR	Backbone LR	λ_{LGRA}	w_0
CUB	6.0×10^{-6}	2.5×10^{-6}	2.0×10^{-6}	0.8	0.7
SUN	8.5×10^{-6}	1.5×10^{-6}	1.0×10^{-6}	0.2	0.6
AWA2	1.0×10^{-5}	3.5×10^{-6}	5.0×10^{-7}	0.1	0.1

For numerical stability, inverse-hyperbolic distance arguments are clamped above one, inverse-cosine arguments are clamped to their valid interval, and near-zero tangent and spatial norms use $\epsilon = 10^{-6}$. Tangent norms in the exponential map are capped at 20 before evaluating hyperbolic functions; this safeguard was not activated in our training runs. In the no-ROH ablation, the Euclidean feature norm parameterizes the geodesic radius as $\min(\|\boldsymbol{f}\|_2, 20)$, while semantic prototypes retain their learned branch base radius.

Table 4: Complete ablations (%). Rows a–h match Table 2; i–n are additional variants. Each row changes the indicated components of Full. Euclidean variants a, i, and k retain PBE and use dot-product scores as in SVIP (Chen et al., 2025b), omitting Lorentz embedding, radius-offset heads, and LGRA; a retains probability fusion. Variants b and j retain Lorentz embedding without LGRA. Variant f reverses only the LGRA direction, keeping zero aperture. Variant g removes the radius-offset head and sets each visual radius to $r = \min(\|\boldsymbol{f}\|_2, 20)$, while retaining the learned branch radius for semantic prototypes. Variant h pools projected patch features in Euclidean space before the Lorentz lift. Rows l–n list patch/CLS prototype targets; the default is attributes/names. LGRA always acts on patches, and the CLS projection is reinitialized for attribute targets in l and n. $T1$ denotes CZSL accuracy; U/S are GZSL unseen/seen accuracies and H their harmonic mean. Bold marks column maxima at the displayed precision, including ties.

ID	Variant	CUB				SUN				AWA2			
		T1	U	S	H	T1	U	S	H	T1	U	S	H
a	Euclidean dual	82.8	75.6	75.6	75.6	87.7	69.7	61.8	65.5	96.7	93.1	95.7	94.4
b	w/o LGRA	83.8	72.6	78.1	75.2	88.3	75.9	60.0	67.0	96.9	95.2	94.9	95.1
c	Patch only	81.2	71.1	79.0	74.8	77.5	59.9	49.0	53.9	72.0	64.3	85.4	73.4
d	CLS only	72.1	61.5	70.1	65.5	87.9	70.4	61.4	65.6	96.8	**96.1**	93.3	94.7
e	w/o PBE	84.7	76.9	78.9	77.9	87.9	72.9	**61.9**	67.0	96.9	94.2	94.8	94.5
f	Outward LGRA	83.8	74.2	77.7	75.9	88.3	**76.2**	60.3	67.3	96.9	94.9	95.2	95.1
g	Norm-based visual radius	84.9	77.3	79.0	78.1	88.6	75.1	61.3	67.5	96.9	95.1	95.0	95.0
h	Euclidean pooling	84.0	75.3	78.5	76.8	88.3	73.3	61.1	66.7	96.8	95.1	95.2	95.1
i	Euclidean patch only	78.5	69.6	76.6	73.0	77.1	59.4	46.4	52.1	70.8	64.3	84.6	73.1
j	Hyperbolic patch, no LGRA	79.6	70.0	74.7	72.3	76.9	62.0	47.3	53.6	71.1	65.1	82.2	72.6
k	Euclidean CLS only	71.8	58.4	70.6	64.0	87.4	71.3	59.2	64.7	96.8	93.4	**95.8**	94.6
l	Attr. / Attr.	81.4	71.6	**81.5**	76.2	79.3	61.9	54.8	58.1	76.7	73.6	88.2	80.2
m	Name / Name	70.3	59.8	74.2	66.2	88.0	69.8	61.3	65.3	96.9	95.4	94.6	95.0
n	Name / Attr.	81.4	67.5	79.4	73.0	82.2	65.8	53.1	58.8	77.1	75.9	89.3	82.1
Full		**85.2**	**77.6**	79.1	**78.3**	**88.8**	76.0	61.0	**67.7**	**97.0**	95.3	95.1	**95.2**

B.2 Baseline Selection and Reporting

Alongside the generative and embedding-based methods in Section 4.2, we separately compare CLIP (Radford et al., 2021), CoOp (Zhou et al., 2022a), CoOp+SHIP (Wang et al., 2023), and APAN (Xie et al., 2025) to contextualize vision–language pretraining and adaptation. The CLIP row is our inference-only baseline under the same dataset splits and evaluation settings as HyGLoA. We keep both encoders frozen, use the same class-name prompt template, and form the prediction solely from the CLS-token image representation and class-name prototypes; no model parameter is trained or fine-tuned. The APAN row uses the reported w/o ResNet configuration: removing its additional ResNet branch gives closer parameter-scale comparability, without implying identical model sizes. TF-VAEGAN is an ECCV 2020 method (Narayan et al., 2020); the daggered table row instead uses SHIP's 2023 re-evaluation (Wang et al., 2023), in which visual features come from the CLIP image encoder and each class semantic vector is produced by feeding "a photo of a {class}" to the CLIP text encoder. The Venue column therefore records the original method publication, whereas Backbone records the configuration used to obtain the displayed result. SHIP evaluates this setting and CoOp on CUB and AWA2 (as well as FLO), but not SUN; we use dashes rather than mix in an untraced SUN source. All remaining baseline scores are quoted from published reports, including their reported harmonic means; consequently, an H value need not exactly equal the value recomputed from displayed one-decimal U and S. Scores and gains are rounded to one decimal, with gains computed from unrounded results. VADS has higher CZSL $T1$ on CUB than HyGLoA (86.8 vs. 85.2). Frozen CLIP is marginally higher on AWA2 CZSL $T1$ (97.1 vs. 97.0), while HyGLoA yields the higher GZSL H (95.2 vs. 90.9).

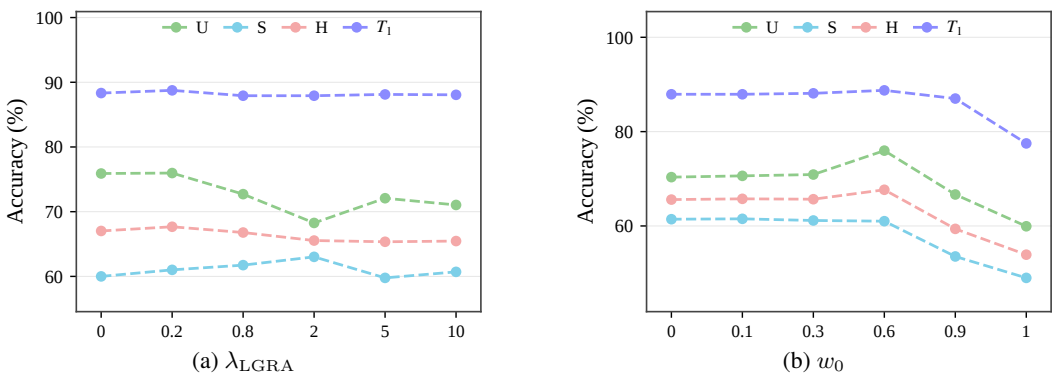


Figure 6: Supplementary hyperparameter sensitivity on SUN. Curves report CZSL $T1$ and GZSL U, S, and H for (a) the LGRA loss weight and (b) the initial patch-fusion weight.

B.3 Complete Ablations and Variant Definitions

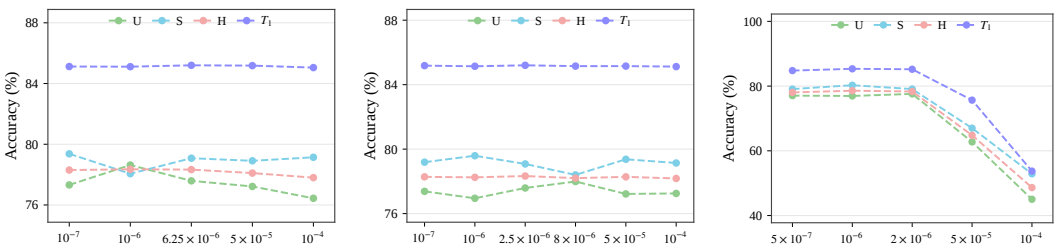
Branch-specific behavior. The complete hyperbolic CLS configuration yields higher CLS-only H than the Euclidean configuration (rows d and k) on all datasets, although the AWA2 difference is small (94.7 vs. 94.6). This comparison includes DRF together with the change of geometry and therefore does not isolate curvature alone. For the patch branch, the corresponding hyperbolic configuration without LGRA declines relative to the Euclidean variant on CUB (73.0 to 72.3) and AWA2 (73.1 to 72.6), while improving on SUN (52.1 to 53.6). Adding LGRA improves patch-only H on all three datasets (rows j and c), supporting its role in local alignment.

On AWA2, Full improves H over CLS only from 94.7 to 95.2. This gain reflects a better seen/unseen balance: S increases from 93.3 to 95.1, while U decreases from 96.1 to 95.3. Local–global fusion thus improves the harmonic mean rather than both component accuracies.

The CUB decline occurs in both dual-branch (a vs. b) and patch-only (i vs. j) settings. The outward patch shift observed without LGRA (Figure 5(a)–(b)) is consistent with the geodesic-distance-based joint classification loss $\mathcal{L}_{\text{joint}}$ learning class separation without the intended local visual–attribute ordering. Inward LGRA regularizes this relation and recovers performance. The stronger global contribution indicated by the learned best-checkpoint patch weights (0.481 on SUN and 0.107 on AWA2) and the single-branch results helps explain why these datasets retain gains in the dual-branch setting, even though the unconstrained patch-only model also declines on AWA2. This accounts for the observed pattern without treating radial expansion as a universal consequence of the classification loss.

Remaining component ablations. Removing PBE (row e) lowers H by 0.4/0.7/0.7 points on CUB/SUN/AWA2. Replacing ROH by the projected feature norm (row g) lowers H by 0.2 points on every dataset, while pooling before the Lorentz lift (row h) trails LCP by 1.5/1.0/0.1 points. Thus PBE, independent radial control, and hyperbolic pooling each contribute within the full system, with the strongest LCP gains on CUB and SUN. These comparisons do not isolate curvature alone.

Semantic roles and initialization. Rows l–n use attributes/attributes, names/names, and names/attributes as patch/CLS targets. Both branches remain hyperbolic, with LGRA always applied to patches. The CLS projection is randomly reinitialized for attribute targets (rows l and n); these variants therefore change initialization along with semantic assignment. The proposed attributes/names assignment yields the highest H across these variants on all datasets. Names/names particularly hurts CUB (66.2 vs. 78.3); attributes/attributes hurts SUN (58.1 vs. 67.7) and AWA2 (80.2 vs. 95.2). Swapping targets also lowers H, favoring the intended local/global semantic roles, although the initialization changes prevent isolating semantic assignment as the sole cause.



(a) Patch LR (b) CLS LR (c) Backbone LR

Figure 7: Learning-rate sensitivity on CUB: (a) patch-head learning rate, (b) CLS-head learning rate, and (c) backbone learning rate. The head learning rates admit broad stable regions, whereas excessively large backbone learning rates lead to a sharp performance decline.

B.4 Hyperparameter Analysis

Figure 6 complements the CUB analysis in Figure 3. SUN attains its best GZSL H at $\lambda_{\mathrm{LGRA}} = 0.2$ and does not exhibit the high-weight collapse observed on CUB, consistent with its stronger reliance on the CLS branch. The learnable fusion weight is initialized by w_0 but continues to update during training. SUN peaks at $w_0 = 0.6$ and deteriorates markedly when $w_0 \geq 0.9$. The learned best-checkpoint patch weights are 0.753, 0.481, and 0.107 on CUB, SUN, and AWA2, respectively, suggesting that CUB benefits from stronger local evidence whereas SUN and especially AWA2 rely more on global semantics. The interior optima on CUB and SUN nevertheless show that neither granularity should dominate universally; this conclusion is specific to our optimization and datasets rather than an inherent granularity ordering for all ZSL tasks.

Figure 7 complements the loss-weight and fusion-weight analysis by examining the three learning rates on CUB. Performance remains relatively stable over broad ranges, suggesting that the reported gains do not depend on a narrowly tuned learning-rate value. Panel (c) shows a sharp decline when the backbone learning rate becomes too large, consistent with overly aggressive updates disrupting useful pretrained CLIP representations. The observation supports conservative backbone fine-tuning in our setting; these experiments assess learning-rate values, rather than alternative learning-rate schedules.

B.5 Qualitative Analysis

B.5.1 Semantic Role of the Radius-Offset Head

The response in Figure 4 measures positive patch evidence along a matched class attribute prototype. For patch token $\boldsymbol{t}_k$, let $\boldsymbol{f}_k = \phi_{\mathrm{attr}}(\boldsymbol{t}_k)$ and let $\mathbf{u}_k = \mathrm{Log}_{\mathbf{o}}(\mathbf{h}_k)[1{:}]$ be the spatial part of its origin-tangent representation. For the matched attribute prototype $\mathbf{p}_y^a$, define $\mathbf{v}_y = \mathrm{Log}_{\mathbf{o}}(\mathbf{p}_y^a)[1{:}]$. The Lorentz logarithmic map is the inverse of the exponential map at the origin: for $\mathbf{p} = (\cosh r, \sinh r\, \boldsymbol{n})$, $\mathrm{Log}_{\mathbf{o}}(\mathbf{p}) = (0, r\boldsymbol{n})$. Consequently, $\mathbf{u}_k = r_k \widehat{\boldsymbol{f}}_k$ and $\mathbf{v}_y = R_a \boldsymbol{e}_y^{a\star}$. The displayed response is

$$a_k = \left[\mathbf{u}_k^\top \frac{\mathbf{v}_y}{\|\mathbf{v}_y\|_2} \right]_+ = [r_k \cos\theta_k]_+ , \tag{21}$$

where $r_k = \|\mathbf{u}_k\|_2$ is the patch's geodesic radius and θ_k measures patch–prototype directional agreement at the origin, distinct from the prototype-anchored LGRA angle. For the full model, $r_k = R_a \exp(\Delta_k^a)$, where $\Delta_k^a = \psi_{\mathrm{rad}}(\boldsymbol{f}_k)$; without the radius-offset head, $r_k = \min(\|\boldsymbol{f}_k\|_2, 20)$. We keep the score polarity fixed and apply no correlation-based sign reversal. Scores are min–max normalized within each image only for rendering, so the maps compare spatial allocation rather than absolute response magnitude across models. A high response requires both positive directional agreement and sufficient geodesic radius: direction identifies the related prototype, while radius modulates the strength of that evidence. This is a diagnostic geometric readout, rather than a ViT self-attention weight or an exact decomposition of the classification score.

Although row g of Table 4 differs from Full by only 0.2 H points on each dataset, Figure 4 shows a clearer representational difference. The full model places more compact responses on prototype-

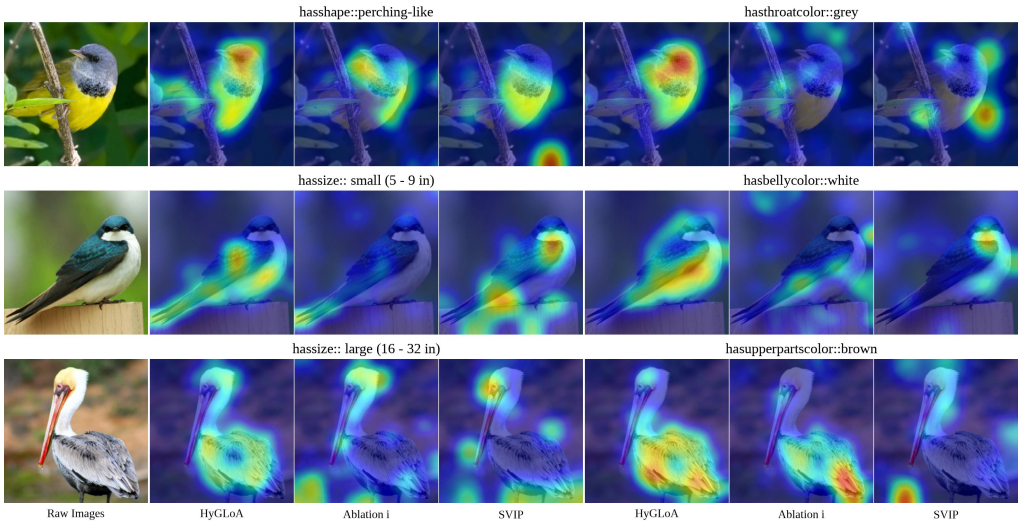


Figure 8: Single-attribute spatial responses on CUB before the Lorentz lift. Original images appear on the left; for each attribute, the response maps on the right show HyGLoA, Ablation i (Euclidean patch only), and SVIP (Chen et al., 2025b), from left to right. HyGLoA and Ablation i use a 14×14 grid, whereas SVIP uses a 7×7 grid; the comparison concerns semantic localization rather than map sharpness.

related foreground regions, including images with complex backgrounds, whereas deriving direction and radius through one projection produces broader background responses. This supports the usefulness of separately parameterized radial control for spatial selectivity. It does not imply that every high-radius patch is discriminative: specific local details may also be shared or irrelevant, which helps explain why the qualitative difference is larger than the observed accuracy gain.

B.5.2 Single-Attribute Localization before the Lorentz Lift

The attribute head produces $\boldsymbol{f}_k = \phi_{\mathrm{attr}}(\boldsymbol{t}_k) \in \mathbb{R}^{d_a}$. For named attribute a, the raw spatial response is $S_k^{(a)} = (\boldsymbol{f}_k)_a$, extracted before normalization, pooling, or the Lorentz lift. For display, we use the fixed-polarity map

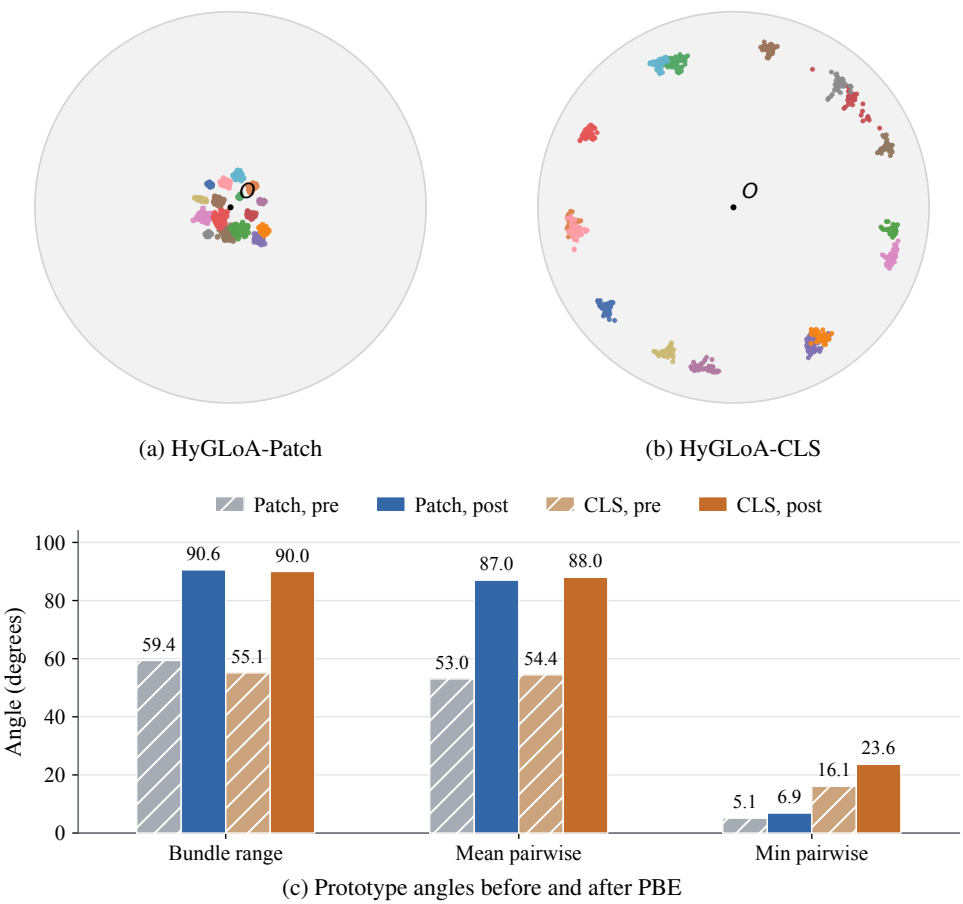
$$A_k^{(a)} = \left[S_k^{(a)} - Q_\rho(\{S_j^{(a)}\}_{j=1}^K) \right]_+, \tag{22}$$

where Q_ρ is a fixed spatial quantile shared by all displayed methods. No map is sign-flipped according to another method. This coordinate-wise readout is a diagnostic of the attribute head, not a ViT attention weight or direct attribute-localization supervision. Figure 8 shows that HyGLoA better localizes spatially extended attributes, such as shape and size, over the corresponding silhouette and object extent. This is consistent with global class supervision helping the shared encoder retain large-scale semantic context for local attribute responses. However, Ablation i removes both the CLS branch and hyperbolic components, so this comparison does not isolate the global branch's contribution. The maps support the proposed local–global interaction while showing that its influence is already visible before the Lorentz lift.

B.5.3 LGRA-Induced Prototype–Image Layout

HoroPCA (Chami et al., 2021) is used to inspect the coarse placement of image representations relative to semantic prototypes. Within each panel, we jointly project the image points and prototypes without recentering the model origin. The resulting view complements CO-SNE (Guo et al., 2022), which is used below to inspect local class neighborhoods. Projected distances are interpreted qualitatively rather than as exact original-space radii. As a complementary original-space diagnostic, we use the pooled radial residual

$$\Delta r_b = \overline{d_{\mathbb{L}}(\mathbf{h}^b, \mathbf{o})} - R_b, \tag{23}$$



(a) HyGLoA-Patch (b) HyGLoA-CLS

(c) Prototype angles before and after PBE

Figure 9: Complementary hyperbolic-structure analysis on CUB. (a)–(b) CO-SNE maps of pooled patch and CLS image embeddings; colours denote classes and only local neighborhoods are interpreted. (c) Prototype angular statistics before and after PBE. Bundle range is the maximum angle to the bundle axis; mean and minimum pairwise angles use all class pairs.

where the overline denotes the logged batch mean and R_b is the shared prototype radius. Negative and positive values indicate visual points lying inside and outside the prototype shell on average, respectively. At the selected CUB best-H checkpoints, the diagnostics are -0.900 for the Full patch branch, $+0.454$ for the patch branch without LGRA, and $+2.417$ for the Full CLS branch. Because these are checkpoint training diagnostics rather than dataset-wide estimates, we use their signs to corroborate, not replace, the qualitative projection.

Figure 5(a)–(b) qualitatively shows a clear change in the local visual–prototype ordering. Without LGRA, pooled patch representations predominantly lie outside their attribute prototypes. With LGRA, they shift inward and predominantly lie closer to the origin than the prototypes, while the projected class groups appear more compact. This supports the intended role of LGRA in encouraging an inward local visual–attribute relation, consistent with the sign change in Eq. equation 23 and the performance recovery in the ablations.

The CLS representations in Figure 5(c) remain outside their class-name prototypes, although they come from the same full model as panel (b). Both this branch and the patch branch in panel (a) lack direct LGRA supervision and are learned through the geodesic-distance-based joint classification loss $\mathcal{L}_{\text{joint}}$, which combines the branch probabilities derived from negative Lorentz-distance logits. Their layouts, together with the observed radial training behavior, support an outward visual radial tendency under this objective in our experiments. For a fixed nonzero angular separation, larger radii can increase pairwise hyperbolic separation between image points, providing geometric intu-

ition for this tendency. Classification nevertheless depends on relative image–prototype distances, so this observation does not establish a universally positive radial update under $\mathcal{L}_{\text{joint}}$. LGRA counterbalances the observed trend in the patch branch while leaving the CLS branch without an explicit radial-ordering constraint.

B.5.4 Class Neighborhoods and PBE-Induced Angular Separation

CO-SNE (Guo et al., 2022) emphasizes local hyperbolic neighborhoods and is used only to inspect class-wise clustering; its apparent radii are not interpreted geometrically. We group the patch/CLS CO-SNE maps with the PBE angular statistics because together they describe local image neighborhoods and the global redistribution of semantic prototype directions, respectively.

Figure 9(a)–(b) shows qualitatively class-consistent local neighborhoods in both branches, indicating discriminative structure at complementary granularities.

Let $\widehat{\boldsymbol{e}}_y$ be a unit prototype direction and $\widehat{\boldsymbol{a}}$ the bundle axis. We compute $\Theta_{\text{range}} = \max_y \arccos(\widehat{\boldsymbol{e}}_y^\top \widehat{\boldsymbol{a}})$, together with the mean and minimum of $\arccos(\widehat{\boldsymbol{e}}_i^\top \widehat{\boldsymbol{e}}_j)$ over all seen and unseen class pairs. For patch prototypes, PBE increases the range and mean pairwise angle from $59.4°/53.0°$ to $90.6°/87.0°$; for CLS prototypes, they increase from $55.1°/54.4°$ to $90.0°/88.0°$. The minimum angles increase more moderately, from $5.1°$ to $6.9°$ and from $16.1°$ to $23.6°$, respectively, as shown in Figure 9(c). The range uses the raw all-class mean axis before PBE and the seen-derived fitted bundle axis afterward, so the range is a descriptive statistic rather than a controlled before/after comparison. The pairwise statistics do not depend on this axis choice and provide the direct comparison. Together with the performance drop in Ablation e, these results support PBE's role in alleviating angular crowding within the full system, without implying a uniformly large pairwise margin.