HyCoPro: Hyperbolic Compositional Prototypes for Compositional Zero-Shot Learning

Anonymous authors
Paper under double-blind review

Abstract

Compositional Zero-Shot Learning (CZSL) aims to recognize unseen state–object compositions by learning from seen compositions. Recent CLIP-based methods have improved CZSL by turning each candidate pair into a text prompt, but the resulting composition prototypes are still often optimized as independent labels. This becomes fragile in open-world CZSL, where the model must score all possible compositions, many of which have no corresponding training images. To address this problem, we propose **HyCoPro**, a hyperbolic compositional closure framework for CZSL. HyCoPro organizes CLIP visual and textual features as role-specific state, object, and composition prototypes in a Poincaré ball, where Möbius addition closes primitive prototypes into composition prototypes. This closure is enforced from both visual and textual sides, strengthening unseen prototypes without requiring extra images. We further provide theoretical analysis explaining why HyCoPro is well suited to compositional recognition in CZSL. Experiments on three benchmarks demonstrate consistent gains, particularly under open-world evaluation.

1 Introduction

Humans rarely learn every visual concept as an isolated category. After seeing "sliced bread" and "red apple", one can readily recognize the novel concept "sliced apple", despite never observing this particular composition. Compositional Zero-Shot Learning (CZSL) formalizes this ability by training on a subset of state–object compositions and testing on novel combinations of known primitives. The central challenge is therefore not merely to recognize unseen labels, but to learn representations from which observed primitives can form reliable new compositions. This difficulty is amplified in open-world CZSL, where the model must score the full Cartesian product of states and objects, including many plausible but unobserved pairs.

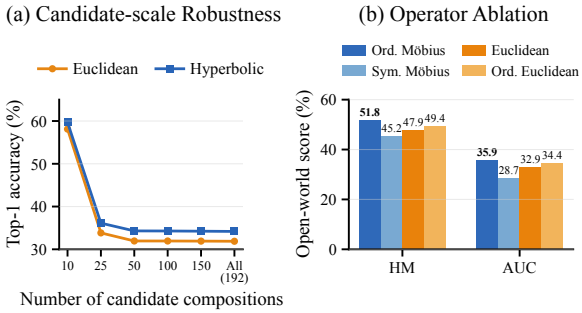

Figure 1: Diagnostic analysis on UT-Zappos. (a) As candidate compositions increase, the hyperbolic model maintains a clearer Top-1 advantage over its matched Euclidean counterpart, suggesting improved candidate-scale robustness from additional representational capacity. (b) Open-world HM and AUC averaged over three seeds. Ordered Möbius achieves the best performance, validating the benefit of Möbius composition. Moreover, both order-aware variants outperform their order-agnostic counterparts, demonstrating that preserving state–object order benefits CZSL.

Large-scale vision-language models such as CLIP (Radford et al., 2021) have become a strong foundation for CZSL. A common strategy is to convert every candidate pair into a prompt, e.g., "a photo of red apple", and classify an image by image-text similarity. Prompt learning methods further adapt textual representations to the CZSL domain (Nayak et al., 2023; Xu et al., 2024). While effective, this paradigm still

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.

tends to optimize each composition prototype as an independent candidate label. The prototype of "red apple" is available through a prompt, but it is not explicitly required to be induced by the primitive prototypes of "red" and "apple". In other words, prompt-derived prototypes name candidate compositions without explicitly enforcing primitive-to-composition closure.

We argue that CZSL benefits from jointly leveraging the representational capacity of hyperbolic space and its intrinsic composition operations. In CZSL, the candidate space spans the Cartesian product of the state set $\mathcal{S}$ and the object set $\mathcal{O}$, yielding $|\mathcal{S}| \times |\mathcal{O}|$ possible compositions, of which only a sparse subset is observed during training. Compared with normalized cosine spaces, where separation is mainly angular and scale-invariant, hyperbolic prototype spaces provide an additional radial degree of freedom under geodesic distance. Their exponentially expanding volume can therefore organize a polynomially growing set of prototypes with only a logarithmic increase in radial distance (Nickel & Kiela, 2017; Ganea et al., 2018), providing a capacity-efficient geometry for dense state–object candidates, especially hard negatives sharing a common state or object. Furthermore, states and objects occupy distinct, non-interchangeable semantic roles within a composition. A composition operator should therefore preserve the ordered state–object assignment rather than treating primitive inputs symmetrically. Möbius addition naturally captures this role sensitivity because it is generally non-commutative while remaining closed within the Poincaré ball. Consequently, it provides a geometrically consistent and manifold-preserving operation for composing primitives.

To this end, we propose **HyCoPro**, a hyperbolic compositional closure framework for CZSL. HyCoPro uses CLIP as a visual-language backbone and maps visual and textual features into role-specific hyperbolic prototypes for states, objects, and compositions. For each candidate pair (s, o), with state s and object o, let $\mathbf{t}_s^h$ and $\mathbf{t}_o^h$ denote its hyperbolic state and object text prototypes. HyCoPro constructs the primitive-composed prototype as $\tilde{\mathbf{t}}_{s,o}^h = \mathbf{t}_s^h \oplus_c \mathbf{t}_o^h$, where $\tilde{\mathbf{t}}_{s,o}^h$ denotes the resulting primitive-composed prototype and $\oplus_c$ denotes Möbius addition in the Poincaré ball with curvature $-c$. HyCoPro enforces this closure from both modalities. On the visual side, image-side state and object features are encouraged to compose toward the target composition prototype. On the text side, HyCoPro aligns the primitive-composed prototype $\tilde{\mathbf{t}}_{s,o}^h$ with the explicit composition prototype $\mathbf{t}_{s,o}^h$ encoded from the full prompt, i.e., $\tilde{\mathbf{t}}_{s,o}^h \approx \mathbf{t}_{s,o}^h$, where $\approx$ indicates alignment under the text-side closure objective. This dual-sided closure strengthens unseen composition prototypes without requiring additional images.

The key contributions of this work are summarized as follows:

- We explore the usage of hyperbolic geometry for CZSL and propose **HyCoPro**, a role-specific prototype framework that represents state, object, and composition semantics in a Poincaré ball. HyCoPro connects primitive prototypes through Möbius addition and incorporates two complementary closure objectives to promote compositional consistency without additional visual supervision.
- We theoretically substantiate the geometric and discriminative advantages of HyCoPro through hyperbolic prototype capacity bounds, geodesic classification margins, and primitive-aware score guarantees.
- We conduct extensive experiments on MIT-States, UT-Zappos, and C-GQA, demonstrating the effectiveness of HyCoPro, particularly in the challenging open-world CZSL setting.

2 Related Work

Compositional Zero-Shot Learning (CZSL). Compositional zero-shot learning (CZSL) recognizes unseen state–object pairs by transferring primitives learned from seen compositions. Early methods modeled attribute–object interactions through compositional operators and structured constraints, while disentanglement and invariant learning sought more transferable primitives (Misra et al., 2017; Nagarajan & Grauman, 2018; Li et al., 2020; Atzmon et al., 2020; Zhang et al., 2022). Open-world CZSL expands the candidate set to the full Cartesian product, motivating graph propagation, prototype sharing, and feasibility modeling (Naeem et al., 2021; Mancini et al., 2021; Karthik et al., 2022). With CLIP (Radford et al., 2021), prompt-based methods represent candidate pairs as text: CSP learns compositional soft prompts (Nayak et al., 2023), whereas GIPCOL and PLID add graph relations and language-informed distributions (Xu et al., 2024; Bao et al., 2024). Other approaches introduce concept-aware adapters or separate state, object, and composi-

tion pathways (Zheng et al., 2024; Huang et al., 2024; Li et al., 2024). Recent studies investigate complex-space embeddings, logic-induced embeddings, causal decomposition, and expert adaptation (Jiang et al., 2025; Wu et al., 2025; Yuan et al., 2026; Zhang et al., 2026a). Complementary CLIP-based work explores decomposed prompts, retrieval-augmented primitives, visual proxies, test-time knowledge accumulation, composable flows, and sparse cross-modal alignment (Lu et al., 2023; Jing et al., 2024; Zhang et al., 2025; Yan & Feng, 2025; He et al., 2026; Zhang et al., 2026b).

HyCoPro follows this line of CZSL research, but focuses on the geometric relation among primitives and compositions. It retains CLIP's prompt semantics while explicitly requiring state and object prototypes to close into their corresponding composition prototypes.

Hyperbolic Representation Learning. Hyperbolic space is commonly used for hierarchical structures and spaces with rapid growth, due to its exponential volume property (Nickel & Kiela, 2017). Hyperbolic neural networks provide differentiable tools such as exponential maps and Möbius operations, while hyperbolic image embeddings extend these tools to visual representation learning (Ganea et al., 2018; Khrulkov et al., 2020). Recent vision-language work also explores hyperbolic relations for compositional or entailment-style alignment (Pal et al., 2025). In HyCoPro, hyperbolic geometry is used to support compositional closure rather than to encode an external hierarchy or entailment graph. Möbius addition serves as the operation that connects state, object, and composition prototypes for CZSL.

3 METHOD

3.1 OVERALL FRAMEWORK

HyCoPro is built around the idea that compositional recognition should be supported by the geometry of the prototype space. Given an image and a candidate state-object pair, we first use CLIP to obtain visual and textual features. Instead of mapping all semantics into a single embedding space, HyCoPro separates them into three roles: a state branch, an object branch, and a composition branch. The state and object branches model primitive semantics, while the composition branch preserves the holistic semantics of the full pair.

All role-specific features are projected into a shared Poincaré ball, where hyperbolic distance is used for classification and Möbius addition is used for primitive composition. For each pair (s, o), HyCoPro maintains two complementary composition prototypes: an explicit prototype encoded from the full prompt, and a primitive-composed prototype obtained by closing the state and object prototypes under Möbius addition. The two views are tied together by dual-sided closure constraints. On the visual side, image-side state and object features are encouraged to compose toward the target composition. On the textual side, composed primitive text prototypes are aligned with explicit full-prompt prototypes over candidate pairs. At inference time, these prototypes are combined to score both closed-world and open-world candidate compositions.

3.2 PROBLEM FORMULATION

Let $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$ denote the state set and $\mathcal{O} = \{o_1, \ldots, o_{|\mathcal{O}|}\}$ denote the object set. The full composition space is the Cartesian product $\mathcal{C} = \mathcal{S} \times \mathcal{O}$. The training set contains images annotated with seen compositions:

$$\mathcal{D}_{\text{train}} = \{(x_i, s_i, o_i)\}_{i=1}^{N}, \quad (s_i, o_i) \in \mathcal{C}_{\text{seen}} \subset \mathcal{C}.$$

Given an image x, a CZSL model learns a compatibility score $F(x, s, o)$ and predicts

$$(\hat{s}, \hat{o}) = \arg \max_{(s,o) \in \mathcal{C}_{\text{test}}} F(x, s, o).$$

In the closed-world setting, $\mathcal{C}_{\text{test}}$ consists of seen pairs and a predefined unseen split. In the open-world setting, $\mathcal{C}_{\text{test}} = \mathcal{C}$, so the model must score every state-object pair. This makes prototype reliability and seen-unseen calibration central to the problem.

3.3 ROLE-SPECIFIC HYPERBOLIC PROTOTYPES

HyCoPro uses CLIP as the visual-language backbone. For an image x, CLIP provides a visual feature $\mathbf{v}_x$. For a candidate pair (s, o), we use the prompts $t_r =$ "a photo of r" for $r \in \{s, o\}$, and

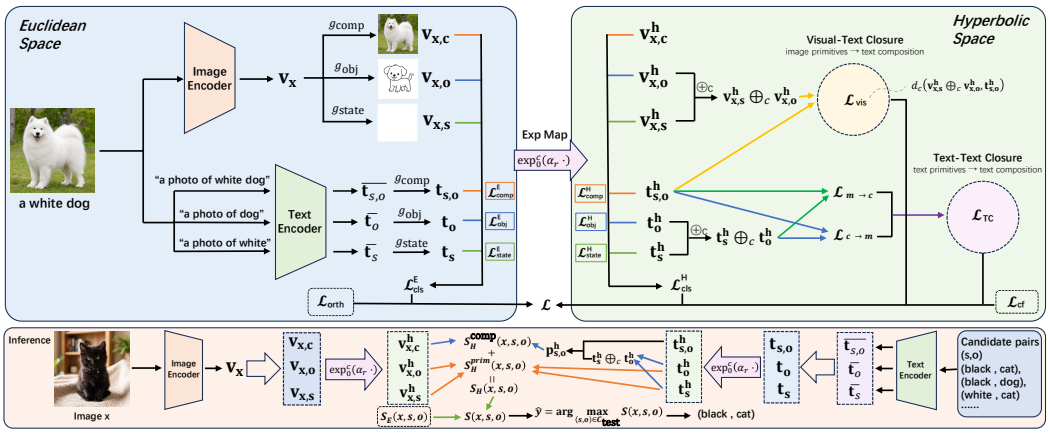

Figure 2: **Overview of HyCoPro. Top:** during training, CLIP visual and textual features are decomposed into role-specific state, object, and composition branches and mapped into the Poincaré ball. Visual-Text Closure aligns the Möbius composition of image-side primitives with the explicit text composition prototype, while Text-Text Closure bidirectionally aligns primitive-composed text prototypes with full-prompt composition prototypes. These closure objectives are optimized jointly with the classification losses. **Bottom:** during inference, explicit and primitive-composed prototypes are interpolated for each candidate pair; composition, primitive, and Euclidean scores are then fused to rank candidates and predict the final composition.

$t_{s,o}$ = "a photo of s o" for the full composition. Their CLIP text features are denoted as $\bar{\mathbf{t}}_s$, $\bar{\mathbf{t}}_o$, and $\bar{\mathbf{t}}_{s,o}$.

Instead of using a single representation for all semantic roles, HyCoPro learns three role-specific projection heads g_{state}, g_{obj}, and g_{comp}. For notational simplicity, we henceforth write g_s, g_o, and g_c for g_{state}, g_{obj}, and g_{comp}, respectively. These heads produce Euclidean role features for both images and text:

$$\begin{aligned} \mathbf{v}_{x,r} &= g_r(\mathbf{v}_x), \quad r \in \{s, o, c\}, \\ \mathbf{t}_s = g_s(\bar{\mathbf{t}}_s), \quad \mathbf{t}_o &= g_o(\bar{\mathbf{t}}_o), \quad \mathbf{t}_{s,o} = g_c(\bar{\mathbf{t}}_{s,o}). \end{aligned} \tag{1}$$

The role features are then mapped to the Poincaré ball $\mathbb{B}_c^d = \{\mathbf{u} \in \mathbb{R}^d : c\|\mathbf{u}\|^2 < 1\}$ by the exponential map at the origin:

$$\begin{aligned} \mathbf{v}_{x,r}^h &= \exp_0^c(\alpha_r \mathbf{v}_{x,r}), \quad r \in \{s, o, c\}, \\ \mathbf{t}_s^h = \exp_0^c(\alpha_s \mathbf{t}_s), \mathbf{t}_o^h &= \exp_0^c(\alpha_o \mathbf{t}_o), \mathbf{t}_{s,o}^h = \exp_0^c(\alpha_c \mathbf{t}_{s,o}). \end{aligned}$$

where α_r is a learnable role-specific scale. The definitions above specify the role-specific Euclidean features and their mappings into the Poincaré ball. This gives image-side hyperbolic features $\mathbf{v}_{x,s}^h$, $\mathbf{v}_{x,o}^h$, and $\mathbf{v}_{x,c}^h$, together with text-side prototypes $\mathbf{t}_s^h$, $\mathbf{t}_o^h$, and $\mathbf{t}_{s,o}^h$. The state and object prototypes represent primitives, while $\mathbf{t}_{s,o}^h$ is the explicit composition prototype encoded from the full prompt.

3.4 Hyperbolic Compositional Closure

The central operation in HyCoPro is Möbius addition. For $\mathbf{x}, \mathbf{y} \in \mathbb{B}_c^d$, it is defined as

$$\mathbf{x} \oplus_c \mathbf{y} = \frac{\left(1 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c\|\mathbf{y}\|^2\right)\mathbf{x} + \left(1 - c\|\mathbf{x}\|^2\right)\mathbf{y}}{D(\mathbf{x}, \mathbf{y})} \tag{2}$$

where $D(\mathbf{x}, \mathbf{y}) = 1 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c^2\|\mathbf{x}\|^2\|\mathbf{y}\|^2$. The Möbius operation in Eq. (2) keeps the composed vector inside the Poincaré ball and gives HyCoPro a geometry-aware way to turn two primitive prototypes into one composition prototype. For a candidate pair (s, o), the primitive-composed prototype is $\tilde{\mathbf{t}}_{s,o}^h = \mathbf{t}_s^h \oplus_c \mathbf{t}_o^h$. This prototype is distinct from the explicit prompt prototype $\mathbf{t}_{s,o}^h$. HyCoPro trains the representation so that $\tilde{\mathbf{t}}_{s,o}^h \approx \mathbf{t}_{s,o}^h$, and extends this text-side closure to image-side features, thereby enforcing compositional closure across both modalities.

Visual-Text Closure. For a training image x_i with label (s_i, o_i), the image-side state and object features should also be composable. We therefore impose a visual Möbius consistency loss:

$$\mathcal{L}_{\text{vis}} = \frac{1}{N}\sum_{i=1}^{N} d_c\left(\mathbf{v}^h_{x_i,s} \oplus_c \mathbf{v}^h_{x_i,o}, \mathbf{t}^h_{s_i,o_i}\right). \tag{3}$$

The loss in Eq. (3) prevents the visual primitive branches from acting only as auxiliary classifiers; they must support the same closure relation used by the text prototypes.

Text-Text Closure. Seen images only supervise $\mathcal{C}_{\text{seen}}$, but open-world CZSL requires reliable scores for many unseen candidates. We therefore apply text-side closure on a candidate set $\mathcal{Q} \subseteq \mathcal{C}$. For each $q = (s, o) \in \mathcal{Q}$, let $\tilde{\mathbf{t}}^h_q = \mathbf{t}^h_s \oplus_c \mathbf{t}^h_o$ be the primitive-composed prototype and $\mathbf{t}^h_q$ be the explicit composition prototype. We define bidirectional logits, where $q' = (s', o')$ indexes candidate compositions in $\mathcal{Q}$:

$$a_{q,q'} = -\tau d_c(\tilde{\mathbf{t}}^h_q, \mathbf{t}^h_{q'}), \quad b_{q,q'} = -\tau d_c(\mathbf{t}^h_q, \tilde{\mathbf{t}}^h_{q'}), \tag{4}$$

Using the logits in Eq. (4), we define the bidirectional contrastive objectives:

$$\begin{aligned}\mathcal{L}_{m\to c} &= -\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \log \frac{\exp(a_{q,q})}{\sum_{q'\in\mathcal{Q}}\exp(a_{q,q'})},\\ \mathcal{L}_{c\to m} &= -\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \log \frac{\exp(b_{q,q})}{\sum_{q'\in\mathcal{Q}}\exp(b_{q,q'})}.\end{aligned} \tag{5}$$

The final closure loss aggregates the two directional objectives in Eq. (5):

$$\mathcal{L}_{\text{TC}} = \frac{1}{2}\left(\mathcal{L}_{m\to c} + \mathcal{L}_{c\to m}\right). \tag{6}$$

3.5 Training Objective

HyCoPro is trained with Euclidean and hyperbolic classification objectives, together with closure and regularization losses. The Euclidean objective preserves the original CLIP matching ability before hyperbolic projection. We apply standard cross-entropy losses to composition, state, and object prediction:

$$\mathcal{L}^E_{\text{cls}} = \mathcal{L}^E_{\text{comp}} + \lambda_s \mathcal{L}^E_{\text{state}} + \lambda_o \mathcal{L}^E_{\text{obj}}. \tag{7}$$

The hyperbolic objective trains the role-specific prototype space. For a training image x_i with ground-truth composition $y_i = (s_i, o_i)$, let $k = (s_k, o_k)$ denote a candidate composition. We first compute the geodesic distance between the image-side composition feature and the candidate composition prototype as $d^{\text{comp}}_{i,k} = d_c(\mathbf{v}^h_{x_i,c}, \mathbf{t}^h_{s_k,o_k})$. Following margin-based classification, we make the positive pair harder by adding a distance margin m, giving $\tilde{d}^{\text{comp}}_{i,k} = d^{\text{comp}}_{i,k} + m\mathbf{1}[k = y_i]$. We also up-weight hard negatives that share the state or object with the positive pair, since such candidates are visually and semantically close in CZSL. Let

$$w_{i,k} = \begin{cases} w_{\text{hard}}, & k \neq y_i \text{ and } k \text{ shares the state or object with } y_i,\\ 1, & \text{otherwise.}\end{cases}$$

The composition logit is $\ell^{\text{comp}}_{i,k} = -\tau\tilde{d}^{\text{comp}}_{i,k} + \log w_{i,k}$, where τ is a logit scale. The hyperbolic composition loss is then the cross-entropy loss over candidate composition prototypes:

$$\mathcal{L}^H_{\text{comp}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\ell^{\text{comp}}_{i,y_i})}{\sum_k \exp(\ell^{\text{comp}}_{i,k})}. \tag{8}$$

State and object logits are defined analogously by replacing composition prototypes with state and object prototypes. These two primitive losses are not merely auxiliary supervision: they train the state and object branches that are later reused at inference time to form the primitive-aware score. Together with the composition term in Eq. (8), they form the hyperbolic classification loss

$$\mathcal{L}^H_{\text{cls}} = \mathcal{L}^H_{\text{comp}} + \lambda_s \mathcal{L}^H_{\text{state}} + \lambda_o \mathcal{L}^H_{\text{obj}}. \tag{9}$$

We further use two auxiliary regularizers. First, an orthogonality loss reduces semantic overlap between the state and object branches before the exponential map. Let $[\cdot]_+ = \max(\cdot, 0)$, let ε be a small tolerance, and let $\mathbf{v}$ and $\mathbf{t}$ denote pre-hyperbolic role features. We define

$$\mathcal{L}_{\text{orth}} = \frac{1}{N}\sum_{i=1}^{N} \left[|\cos(\mathbf{v}_{x_i,s}, \mathbf{v}_{x_i,o})| - \varepsilon\right]_+ + \left[|\cos(\mathbf{t}_{s_i}, \mathbf{t}_{o_i})| - \varepsilon\right]_+ . \tag{10}$$

Second, a counterfactual composition loss forms mismatched state-object pairs within a batch. Let $\pi(i)$ denote a batch permutation. We compose the state from $x_{\pi(i)}$ with the object from x_i, and align this image-side Möbius composition with the corresponding text-side composition:

$$\mathcal{L}_{\text{cf}} = \frac{1}{N}\sum_{i=1}^{N} d_c\left(\mathbf{v}^h_{x_{\pi(i)},s} \oplus_c \mathbf{v}^h_{x_i,o}, \mathbf{t}^h_{s_{\pi(i)}} \oplus_c \mathbf{t}^h_{o_i}\right). \tag{11}$$

This regularizer encourages the closure operation to remain meaningful beyond the observed pair.

The total objective is

$$\mathcal{L} = \eta_E \mathcal{L}^E_{\text{cls}} + \eta_H \mathcal{L}^H_{\text{cls}} + \lambda_{\text{TC}}\mathcal{L}_{\text{TC}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}}. \tag{12}$$

Specifically, in Eq. (12), the hyperbolic classification term in Eq. (9) primarily supervises discriminative prototype learning in the Poincaré ball. Closure objectives $\mathcal{L}_{\text{vis}}$ and $\mathcal{L}_{\text{TC}}$ promote consistency between primitive-composed and explicit composition prototypes. The Euclidean classification term in Eq. (7) preserves CLIP alignment, while the regularizers in Eqs. (10) and (11) promote role separation and cross-composition generalization, respectively.

3.6 Inference

At inference time, each candidate pair has both an explicit composition prototype $\mathbf{t}^h_{s,o}$ and a primitive-composed prototype $\tilde{\mathbf{t}}^h_{s,o}$. We combine them through hyperbolic interpolation:

$$\mathbf{p}^h_{s,o} = \gamma \otimes_c \mathbf{t}^h_{s,o} \oplus_c (1-\gamma) \otimes_c \tilde{\mathbf{t}}^h_{s,o}, \tag{13}$$

where $\otimes_c$ denotes Möbius scalar multiplication. The parameter γ in Eq. (13) controls the balance between the explicit prompt prototype and the primitive-composed prototype. The hyperbolic composition and primitive scores are

$$\begin{aligned} S^{\text{comp}}_H(x,s,o) &= -\tau d_c(\mathbf{v}^h_{x,c}, \mathbf{p}^h_{s,o}), \\ S^{\text{prim}}_H(x,s,o) &= -\tau d_c(\mathbf{v}^h_{x,s}, \mathbf{t}^h_s) - \tau d_c(\mathbf{v}^h_{x,o}, \mathbf{t}^h_o). \end{aligned} \tag{14}$$

We then combine the component scores in Eq. (14):

$$S_H(x,s,o) = (1-\alpha)S^{\text{comp}}_H(x,s,o) + \alpha S^{\text{prim}}_H(x,s,o).$$

Finally, the hyperbolic score is combined with the Euclidean CLIP score:

$$S(x,s,o) = \beta S_H(x,s,o) + (1-\beta)S_E(x,s,o). \tag{15}$$

When feasibility priors are available for a dataset, they are added as a log prior to the score in Eq. (15) during open-world scoring.

4 Theoretical Analysis

We provide a brief theoretical analysis of why HyCoPro is well suited to CZSL, especially in the open-world setting. Detailed proofs are deferred to the supplementary material.

Proposition 1 (Capacity of Hyperbolic Prototype Spaces)**.** *Let $d \geq 2$, $R > 0$, $\rho > 0$, and $c > 0$, where R denotes geodesic radius. Let $M_E(R,\rho)$ be the maximum number of ρ-separated points in a d-dimensional Euclidean ball of radius R, and let $M_H(R,\rho)$ be the corresponding number in a d-dimensional hyperbolic ball with curvature $-c$. Then*

$$M_E(R,\rho) \leq \left(1 + \frac{2R}{\rho}\right)^d. \tag{16}$$

In contrast, for sufficiently large R, there exists a constant $C_{d,c,\rho} > 0$ such that

$$M_H(R,\rho) \geq C_{d,c,\rho} \exp\left((d-1)\sqrt{c}R\right). \tag{17}$$

The capacity bounds in Eqs. (16) and (17) characterize HyCoPro's prototype geometry under geodesic distance. While a Euclidean ball admits only polynomial packing growth, a hyperbolic ball has an exponential packing lower bound in geodesic radius. Thus, the Poincaré space can allocate more separable regions to state–object compositions with slower radial expansion, providing geometric support for HyCoPro's geodesic-distance scoring, especially for dense candidate sets sharing the same state or object.

Proposition 2 (Distance Margins from Geodesic Softmax)**.** *For an image x_i with correct composition label y, let $\mathbf{h}_i$ be its hyperbolic composition feature, $\mathbf{p}_y$ be the positive prototype, and $\mathbf{p}_k$ be a negative prototype for $k \neq y$. Define the geodesic distance margin*

$$\Delta_k = d_c(\mathbf{h}_i, \mathbf{p}_k) - d_c(\mathbf{h}_i, \mathbf{p}_y).$$

Consider the clean weighted distance-softmax loss with logit scale $\kappa > 0$:

$$\mathcal{L}_i = \log\left(1 + \sum_{k \neq y} w_k \exp(-\kappa \Delta_k)\right),$$

where $w_k \geq 1$ and $w_k > 1$ for hard negatives that share a state or object with the positive pair. If $\mathcal{L}_i \leq \epsilon$ for $\epsilon > 0$, then every negative prototype satisfies

$$\Delta_k \geq \frac{\log w_k - \log(\exp(\epsilon) - 1)}{\kappa}. \tag{18}$$

The margin bound in Eq. (18) connects HyCoPro's distance-based classifier to discriminative separation. Reducing the geodesic classification loss enlarges the distance gap between the positive composition and its negatives. Hard negatives receive a larger required margin through the $\log w_k/\kappa$ term, which matches the CZSL intuition that pairs sharing a state or object are more confusable and should be separated more carefully.

Proposition 3 (Primitive-Aware Score Margins)**.** *For an image x with correct composition $y = (s, o)$, let $k = (s', o')$ be a candidate pair. HyCoPro combines a composition score with a primitive score:*

$$S_{\text{prim}}(x, s, o) = S_{\text{state}}(x, s) + S_{\text{obj}}(x, o),$$
$$S_\alpha(x, s, o) = (1 - \alpha) S_{\text{comp}}(x, s, o) + \alpha S_{\text{prim}}(x, s, o),$$

where $\alpha \in [0, 1]$. Define $\Delta_{\text{comp}}(k) = S_{\text{comp}}(x, s, o) - S_{\text{comp}}(x, s', o')$, and assume primitive margins $m_s, m_o > 0$ satisfying

$$S_{\text{state}}(x, s) - S_{\text{state}}(x, \bar{s}) \geq m_s, \qquad \forall \bar{s} \neq s,$$
$$S_{\text{obj}}(x, o) - S_{\text{obj}}(x, \bar{o}) \geq m_o, \qquad \forall \bar{o} \neq o.$$

Then

$$S_\alpha(x, s, o) - S_\alpha(x, s', o') \geq (1 - \alpha)\Delta_{\text{comp}}(k) + \alpha\big(\mathbf{1}[s' \neq s] m_s + \mathbf{1}[o' \neq o] m_o\big).$$

where $\mathbf{1}[\cdot]$ is the indicator function.

This bound explains primitive-aware inference: changing the state, object, or both adds the corresponding primitive margin to the final score. Thus, primitive evidence can improve candidate ranking even when an unseen composition has an uncertain composition score, supporting the fusion of role-specific primitive and composition branches at inference.

5 Experiments

5.1 Datasets and Metrics

We evaluate HyCoPro on three standard CZSL benchmarks: MIT-States (Isola et al., 2015), which contains diverse state–object compositions; UT-Zappos (Yu & Grauman, 2014), which focuses on fine-grained shoe concepts; and C-GQA (Naeem et al., 2021), which provides a larger and more diverse compositional space. Following standard protocols, we report seen and unseen accuracies, harmonic mean (HM), and area under the seen–unseen curve (AUC) in both closed- and open-world settings. HM and AUC are obtained by sweeping a calibration bias between seen and unseen candidate scores.

Table 1: Quantitative comparison on three benchmarks under closed-world and open-world settings. Best results are in bold and second-best results are underlined.

Method	MIT-States				UT-Zappos				C-GQA			
	S↑	U↑	HM↑	AUC↑	S↑	U↑	HM↑	AUC↑	S↑	U↑	HM↑	AUC↑
Closed-World Evaluation												
CLIP (ICML'21)	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
CoOp (IJCV'22)	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8	20.5	26.8	17.1	4.4
CSP (ICLR'23)	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
PLID (ECCV'24)	49.7	52.4	39.0	22.1	67.3	68.8	52.4	38.7	38.8	33.0	27.9	11.0
Troika (CVPR'24)	49.0	53.0	39.3	22.1	66.8	73.8	54.6	41.7	41.0	35.7	29.4	12.4
CDS-CZSL (CVPR'24)	50.3	52.9	39.2	22.4	63.9	74.8	52.7	39.5	38.3	34.2	28.1	11.1
IMAX (TPAMI'25)	48.7	53.8	39.1	21.9	69.3	70.7	54.2	40.6	39.7	35.8	29.8	12.8
LOGICZSL (CVPR'25)	50.8	53.9	40.5	23.4	**69.6**	74.9	57.8	45.8	44.4	39.4	33.3	15.3
MSCI (IJCAI'25)	50.2	53.4	39.9	22.8	67.4	75.5	59.2	45.8	42.4	38.2	31.7	14.2
I2CD (AAAI'26)	50.6	**54.1**	40.0	23.3	69.1	**77.1**	57.9	46.4	44.1	**41.2**	**35.0**	**16.3**
Ours	**57.4**	51.3	**41.9**	**25.3**	69.1	76.3	**60.2**	**47.8**	**45.6**	37.9	33.7	15.4
Open-World Evaluation												
CLIP (ICML'21)	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.3
CoOp (IJCV'22)	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.7
CSP (ICLR'23)	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.2
PLID (ECCV'24)	49.1	18.7	20.4	7.3	67.6	55.5	46.6	30.8	39.1	7.5	10.6	2.5
Troika (CVPR'24)	48.8	18.7	20.1	7.2	66.4	61.2	47.8	33.0	40.8	7.9	10.9	2.7
CDS-CZSL (CVPR'24)	49.4	21.8	22.1	8.5	64.7	61.3	48.2	32.3	37.6	8.2	11.6	2.7
IMAX (TPAMI'25)	50.2	18.6	21.4	7.6	68.4	57.3	47.5	32.3	38.7	7.9	11.2	2.6
LOGICZSL (CVPR'25)	50.7	21.4	22.4	8.7	**69.6**	**63.7**	50.8	36.2	43.7	9.3	12.6	3.4
MSCI (IJCAI'25)	49.2	20.6	21.2	7.9	67.4	63.0	53.2	37.3	42.0	10.6	13.7	3.8
I2CD (AAAI'26)	50.6	20.6	21.4	8.2	69.1	62.1	50.2	35.6	44.0	10.9	14.9	4.3
Ours	**57.4**	**26.4**	**27.3**	**12.5**	69.1	62.3	**53.4**	**37.8**	**45.6**	**12.7**	**16.2**	**4.9**

5.2 Implementation Details

We use CLIP ViT-L/14 as the backbone and AdamW for optimization, training only lightweight visual adaptation parameters and role-specific projection heads. LoRA (Hu et al., 2022) is applied to the CLIP vision tower with dataset-specific configurations: rank-64 attention LoRA for 15 epochs on MIT-States, a combination of rank-64 attention and rank-16 MLP LoRA for 10 epochs on UT-Zappos, and rank-128 attention LoRA for 15 epochs on C-GQA. The curvature is fixed to $c = 1$. Learning rates are set to 5×10^{-5}, 2×10^{-4}, and 5×10^{-4} for visual adaptation, hyperbolic projection heads, and learnable loss scales, respectively, with weight decay 10^{-4}. We set $\lambda_{\mathrm{TC}} = 0.05$, sampling 256 candidates for Text Closure in large open-world spaces while using all candidate pairs on UT-Zappos. During open-world inference, feasibility priors are applied to MIT-States and C-GQA. Remaining dataset-specific settings are detailed in the supplementary material.

5.3 Main Results

Table 1 compares HyCoPro with CLIP and prompt-learning baselines (Radford et al., 2021; Zhou et al., 2022; Nayak et al., 2023), as well as recent structured or adaptive CZSL methods (Bao et al., 2024; Huang et al., 2024; Li et al., 2024; Jiang et al., 2025; Wu et al., 2025; Wang et al., 2025; Yuan et al., 2026). Unlike these approaches, HyCoPro focuses on prototype geometry and explicitly connects role-specific primitives and compositions through hyperbolic closure.

Comparisons under Closed-World Setting. HyCoPro achieves the best HM and AUC on MIT-States and UT-Zappos, surpassing the previous best results by 1.4% and 1.9% on MIT-States, and by 1.0% and 1.4% on UT-Zappos. On C-GQA, it yields the highest seen accuracy and second-best HM and AUC. This success stems from complementary representations: full-prompt prototypes preserve holistic semantics, while Möbius-composed prototypes transfer reusable primitive semantics. Enforcing closure between them discourages isolated-label learning and promotes balanced seen–unseen recognition.

Comparisons under Open-World Setting. The open-world setting is more challenging because models must distinguish among the entire Cartesian product of states and objects, including many compositions without training images. HyCoPro achieves the best HM and AUC on all three

Table 2: Ablation study on inference score components under closed-world and open-world settings.

Setting	Variant	MIT-States				UT-Zappos				C-GQA			
		S↑	U↑	HM↑	AUC↑	S↑	U↑	HM↑	AUC↑	S↑	U↑	HM↑	AUC↑
Closed-World	Full	**57.4**	51.3	**41.9**	**25.3**	69.1	**76.3**	**60.2**	**47.8**	**45.6**	37.9	**33.7**	**15.4**
	w/o Primitive Score	53.4	**52.2**	40.9	23.9	65.6	74.6	58.3	44.0	43.6	37.4	32.9	14.5
	w/o Composition Score	55.5	48.5	39.6	22.9	68.1	73.2	57.8	44.6	45.4	34.8	31.4	13.9
	w/o Hyperbolic Score	56.8	51.1	41.5	24.9	**69.9**	73.4	58.4	44.9	43.7	**38.0**	32.4	14.6
Open-World	Full	**57.4**	**26.4**	**27.3**	**12.5**	69.1	**62.3**	**53.4**	**37.8**	**45.6**	**12.7**	**16.2**	**4.9**
	w/o Primitive Score	53.4	22.7	23.4	9.6	65.6	57.5	49.1	32.7	43.6	11.3	14.8	4.3
	w/o Composition Score	55.5	25.6	25.9	11.5	68.1	59.2	51.7	34.5	45.4	11.0	14.8	4.3
	w/o Hyperbolic Score	56.8	24.5	25.6	11.3	**69.9**	53.4	48.0	32.4	43.7	11.0	14.6	4.1

Table 3: Ablation study on closure constraints on UT-Zappos under closed-world and open-world settings.

(a) Closed-World

Variant	S↑	U↑	HM↑	AUC↑
Full	**69.1**	**76.3**	**60.2**	**47.8**
w/o T-T Closure	68.9	74.6	57.2	44.7
w/o V-T Closure	69.1	73.9	56.4	43.8
w/o Both Closures	68.5	73.1	56.1	43.2

(b) Open-World

Variant	S↑	U↑	HM↑	AUC↑
Full	**69.1**	**62.3**	**53.4**	**37.8**
w/o T-T Closure	68.9	60.4	53.0	37.1
w/o V-T Closure	69.1	59.7	52.3	36.2
w/o Both Closures	68.5	59.2	51.9	36.0

datasets, with particularly clear gains on MIT-States, where it improves the previous best HM and AUC by 4.9% and 3.8%, respectively. This stronger open-world advantage is consistent with the design of HyCoPro: the role-specific hyperbolic space provides additional capacity for organizing the expanded candidate set, while Möbius closure constructs informative composition prototypes by transferring reusable state and object semantics. Consequently, candidates without direct visual supervision can still be represented compositionally rather than treated as unrelated labels.

5.4 Ablation Study

To evaluate HyCoPro's key components, we conduct ablation studies on score terms and closure constraints. As shown in Table 2, removing the primitive, composition, or hyperbolic score degrades performance, confirming the value of decomposable evidence, holistic composition prototypes, and the additional contribution of hyperbolic scoring beyond Euclidean CLIP matching. Additionally, Table 3 evaluates the closure design on UT-Zappos. Removing either the text–text or visual–text closure loss impairs performance, with the visual–text constraint having a more pronounced impact. Omitting both yields the worst overall results, demonstrating their complementary roles in maintaining a coherent composition prototype space.

6 Conclusion

We introduced HyCoPro, a hyperbolic compositional prototype framework for CZSL. It organizes CLIP features into role-specific Poincaré prototypes and composes state and object prototypes through Möbius addition. Dual-sided closure transfers reusable primitive semantics to unseen compositions without extra visual supervision, while primitive-aware inference combines decomposable and holistic evidence. Theoretical analysis connects this design to hyperbolic capacity and score margins, and experiments validate its effectiveness, particularly in open-world CZSL.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, specifically to check the rigor of proofs initially developed by the authors and help refine the mathematical arguments. The research proposal and initial proofs were developed by the authors. We did not use generative AI tools for other tasks requiring disclosure under the ICLR policy. Additionally, we used generative AI tools to edit the manuscript to improve readability, identify relevant literature, and source and search for information. We reviewed all AI-assisted work, checking suggested references against their original sources and examining proof revisions for mathematical correctness and consistency with the stated assumptions. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1462–1473, 2020.

Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. In *Proceedings of the European Conference on Computer Vision*, pp. 107–123, 2024.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhenqi He, Lin Li, and Long Chen. FlowComposer: Composable flows for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24005–24014, 2024.

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1383–1391, 2015.

Chenyi Jiang, Shidong Wang, Yang Long, Zechao Li, Haofeng Zhang, and Ling Shao. Imaginary-connected embedding in complex space for unseen attribute-object discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1395–1413, 2025. doi: 10.1109/TPAMI.2024.3487631.

Chenchen Jing, Yukun Li, Hao Chen, and Chunhua Shen. Retrieval-augmented primitive representations for compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2652–2660, 2024.

Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. KG-SP: Knowledge guided simple primitives for open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9336–9345, 2022.

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.

Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11316–11325, 2020.

Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17037–17046, 2024.

Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23560–23569, 2023.

Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5222–5230, 2021.

Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1792–1801, 2017.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 953–962, 2021.

Tushar Nagarajan and Kristen Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision*, pp. 169–185, 2018.

Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *International Conference on Learning Representations*, 2023.

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

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*, 2025.

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*, pp. 8748–8763, 2021.

Yue Wang, Shuai Xu, Xuelin Zhu, and Yicong Li. MSCI: Addressing CLIP's inherent limitations for compositional zero-shot learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2009–2017, 2025.

Peng Wu, Xiankai Lu, Hao Hu, Yongqin Xian, Jianbing Shen, and Wenguan Wang. LOGICZSL: Exploring logic-induced representation for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30301–30311, 2025.

Guangyue Xu, Joyce Chai, and Parisa Kordjamshidi. Gipcol: Graph-injected soft prompting for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5774–5783, 2024.

Xudong Yan and Songhe Feng. TOMCAT: Test-time comprehensive knowledge accumulation for compositional zero-shot learning. In *Advances in Neural Information Processing Systems*, 2025.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 192–199, 2014.

Zhaoquan Yuan, Zining Wang, Yuankang Pan, Ao Luo, Wei Li, Xiao Wu, and Changsheng Xu. I2CD: An invertible causal framework for compositional zero-shot learning via disentangle-compose-disentangle. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12295–12303, 2026.

Shiyu Zhang, Cheng Yan, Yang Liu, Chenchen Jing, Lei Zhou, and Wenjun Wang. Learning visual proxy for compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2793–2802, 2025.

Tian Zhang, Kongming Liang, Ruoyi Du, Xian Sun, Zhanyu Ma, and Jun Guo. Learning invariant visual representations for compositional zero-shot learning. In *Proceedings of the European Conference on Computer Vision*, pp. 339–355, 2022.

Xiao Zhang, Haodong Jing, Yongqiang Ma, and Nanning Zheng. Decoupling primitive with experts: Dynamic feature alignment for compositional zero-shot learning. In *International Conference on Learning Representations*, 2026a.

Yang Zhang, Zhixiang Chi, Xudong Yan, Yang Wang, and Songhe Feng. Bridging the modality gap in compositional zero-shot learning via sparse alignment and unimodal memory bank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Zhaoheng Zheng, Haidong Zhu, and Ram Nevatia. Caila: Concept-aware intra-layer adapters for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

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, 2022.

SUPPLEMENTARY MATERIAL

A THEORETICAL PROOFS

A.1 PROOF OF PROPOSITION 1: CAPACITY OF HYPERBOLIC PROTOTYPE SPACES

Proof. We first prove the Euclidean upper bound. Let $\{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ be any ρ-separated set in the d-dimensional Euclidean ball $B_E(0, R)$. Then the Euclidean balls $B_E(\mathbf{z}_j, \rho/2)$ have pairwise disjoint interiors; any boundary intersections have zero volume. Moreover, all of them are contained in $B_E(0, R + \rho/2)$. If v_d denotes the volume of the Euclidean unit ball in $\mathbb{R}^d$, then

$$N v_d \left(\frac{\rho}{2}\right)^d \leq v_d \left(R + \frac{\rho}{2}\right)^d.$$

Therefore,

$$N \leq \left(\frac{R + \rho/2}{\rho/2}\right)^d = \left(1 + \frac{2R}{\rho}\right)^d.$$

Taking the maximum over all ρ-separated sets gives

$$M_E(R, \rho) \leq \left(1 + \frac{2R}{\rho}\right)^d.$$

We next prove the hyperbolic lower bound. Here R is the geodesic radius in the hyperbolic metric. All balls in this proof are closed. A finite maximal ρ-separated set exists because $B_H(0, R)$ is compact. Let $\{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ be such a maximal ρ-separated set in the hyperbolic ball $B_H(0, R)$ of curvature $-c$. Maximality implies that the balls $B_H(\mathbf{z}_j, \rho)$ cover $B_H(0, R)$; otherwise, a point not covered by these balls could be added to the set while preserving ρ-separation. Hence,

$$\mathrm{Vol}_H(B_H(0, R)) \leq N \,\mathrm{Vol}_H(B_H(0, \rho)),$$

and therefore

$$M_H(R, \rho) \geq \frac{\mathrm{Vol}_H(B_H(0, R))}{\mathrm{Vol}_H(B_H(0, \rho))}.$$

In a d-dimensional hyperbolic space with curvature $-c$, the volume of a geodesic ball of radius R is

$$\mathrm{Vol}_H(B_H(0, R)) = \omega_{d-1} \int_0^R \left(\frac{\sinh(\sqrt{c}t)}{\sqrt{c}}\right)^{d-1} dt,$$

where ω_{d-1} is the volume of the unit $(d-1)$-sphere. Since $d \geq 2$, the integrand satisfies

$$\left(\frac{\sinh(\sqrt{c}t)}{\sqrt{c}}\right)^{d-1} \asymp \exp((d-1)\sqrt{c}t)$$

as $t \to \infty$. Therefore, there exist constants $R_0(d, c) > 0$ and $A_{d,c} > 0$ such that

$$\mathrm{Vol}_H(B_H(0, R)) \geq A_{d,c} \exp((d-1)\sqrt{c}R)$$

for all $R \geq R_0(d, c)$. Since ρ is fixed, $\mathrm{Vol}_H(B_H(0, \rho))$ is a positive constant. Setting

$$C_{d,c,\rho} = \frac{A_{d,c}}{\mathrm{Vol}_H(B_H(0, \rho))}$$

gives

$$M_H(R, \rho) \geq C_{d,c,\rho} \exp((d-1)\sqrt{c}R),$$

which proves the claim. □

Scope. The bounds concern the maximum number of separated points that the geometry can accommodate. They establish representational capacity, not a guarantee that training learns a maximally separated prototype configuration.

A.2 Proof of Proposition 2: Distance Margins from Geodesic Softmax

Proof. Let the positive logit be

$$z_y = -\kappa d_c(\mathbf{h}_i, \mathbf{p}_y),$$

and let each negative logit be weighted by a hard-negative weight w_k:

$$z_k = -\kappa d_c(\mathbf{h}_i, \mathbf{p}_k) + \log w_k, \qquad k \neq y.$$

The per-sample cross-entropy loss is

$$\mathcal{L}_i = -\log \frac{\exp(z_y)}{\exp(z_y) + \sum_{k \neq y} \exp(z_k)}.$$

Rewriting this expression relative to the positive logit gives

$$\begin{aligned}
\mathcal{L}_i &= \log \left(1 + \sum_{k \neq y} \exp(z_k - z_y) \right) \\
&= \log \left(1 + \sum_{k \neq y} w_k \exp\left(-\kappa [d_c(\mathbf{h}_i, \mathbf{p}_k) - d_c(\mathbf{h}_i, \mathbf{p}_y)]\right) \right) \\
&= \log \left(1 + \sum_{k \neq y} w_k \exp(-\kappa \Delta_k) \right).
\end{aligned}$$

Assume $\mathcal{L}_i \leq \epsilon$. Then

$$\sum_{k \neq y} w_k \exp(-\kappa \Delta_k) \leq \exp(\epsilon) - 1.$$

Since every term in the summation is nonnegative, each individual negative k satisfies

$$w_k \exp(-\kappa \Delta_k) \leq \exp(\epsilon) - 1.$$

Taking logarithms yields

$$\log w_k - \kappa \Delta_k \leq \log(\exp(\epsilon) - 1),$$

and therefore

$$\Delta_k \geq \frac{\log w_k - \log(\exp(\epsilon) - 1)}{\kappa}.$$

This proves the margin lower bound. □

Relation to the training objective. The proposition analyzes a per-sample, unsmoothed weighted geodesic softmax loss. Its assumption $\mathcal{L}_i \leq \epsilon$ is sample-specific; a bound on the average training loss alone does not imply the same bound for every sample. For fixed κ and w_k, decreasing ϵ increases the guaranteed lower bound. If κ is learned, a decrease in loss alone does not establish that the unscaled distance margins increase, since changes in κ also affect the loss.

When the positive distance is replaced by $d_c(\mathbf{h}_i, \mathbf{p}_y) + m$ with $m > 0$, define the corresponding unsmoothed per-sample loss as

$$\mathcal{L}_i^{(m)} = \log \left(1 + \sum_{k \neq y} w_k \exp[-\kappa(\Delta_k - m)] \right).$$

If this margin-augmented loss satisfies $\mathcal{L}_i^{(m)} \leq \epsilon$, then the same nonnegative-term argument gives

$$w_k \exp[-\kappa(\Delta_k - m)] \leq \exp(\epsilon) - 1,$$

and hence

$$\Delta_k \geq m + \frac{\log w_k - \log(\exp(\epsilon) - 1)}{\kappa}.$$

Thus, at the same loss threshold and fixed scale and weights, the unsmoothed margin-augmented objective imposes an additional separation requirement of m.

If label smoothing is used, its loss is different from the unsmoothed losses above. The displayed bounds must not be applied directly using the smoothed loss as $\mathcal{L}_i$ or $\mathcal{L}_i^{(m)}$. In that setting, the proposition explains the margin-inducing mechanism of the unsmoothed geodesic softmax component, rather than providing an exact bound for the full training objective.

A.3 Proof of Proposition 3: Primitive-Aware Score Margins

Proof. For an image x with correct composition (s, o) and a candidate pair (s', o'), the primitive-aware score difference is

$$\begin{aligned} &S_\alpha(x, s, o) - S_\alpha(x, s', o') \\ &= (1-\alpha)\left[S_{\text{comp}}(x, s, o) - S_{\text{comp}}(x, s', o')\right] \\ &\quad + \alpha\left[S_{\text{prim}}(x, s, o) - S_{\text{prim}}(x, s', o')\right]. \end{aligned}$$

By definition,

$$S_{\text{prim}}(x, s, o) = S_{\text{state}}(x, s) + S_{\text{obj}}(x, o).$$

Therefore,

$$\begin{aligned} &S_{\text{prim}}(x, s, o) - S_{\text{prim}}(x, s', o') \\ &= S_{\text{state}}(x, s) - S_{\text{state}}(x, s') + S_{\text{obj}}(x, o) - S_{\text{obj}}(x, o'). \end{aligned}$$

The margin assumption on the state branch gives

$$S_{\text{state}}(x, s) - S_{\text{state}}(x, s') \geq \mathbf{1}[s' \neq s] m_s,$$

and the margin assumption on the object branch gives

$$S_{\text{obj}}(x, o) - S_{\text{obj}}(x, o') \geq \mathbf{1}[o' \neq o] m_o.$$

Combining the above inequalities, we obtain

$$\begin{aligned} &S_\alpha(x, s, o) - S_\alpha(x, s', o') \\ &\geq (1-\alpha)\left[S_{\text{comp}}(x, s, o) - S_{\text{comp}}(x, s', o')\right] \\ &\quad + \alpha\left(\mathbf{1}[s' \neq s] m_s + \mathbf{1}[o' \neq o] m_o\right). \end{aligned}$$

Using

$$\Delta_{\text{comp}}(k) = S_{\text{comp}}(x, s, o) - S_{\text{comp}}(x, s', o')$$

gives

$$\begin{aligned} &S_\alpha(x, s, o) - S_\alpha(x, s', o') \\ &\geq (1-\alpha)\Delta_{\text{comp}}(k) + \alpha\left(\mathbf{1}[s' \neq s] m_s + \mathbf{1}[o' \neq o] m_o\right), \end{aligned}$$

which proves the proposition. □

Corollary 1. *Under the assumptions of Proposition 3, the correct pair (s, o) outranks a candidate pair $k = (s', o')$ under the primitive-aware score S_α whenever*

$$(1-\alpha)\Delta_{\text{comp}}(k) + \alpha\left(\mathbf{1}[s' \neq s] m_s + \mathbf{1}[o' \neq o] m_o\right) > 0.$$

Proof. The condition above makes the lower bound in Proposition 3 strictly positive. Therefore,

$$S_\alpha(x, s, o) - S_\alpha(x, s', o') > 0,$$

which means that the correct pair receives a higher primitive-aware score S_α than the candidate pair k. □

Scope of the ranking guarantee. In the main paper, S_α corresponds to the hyperbolic score S_H before fusion with the Euclidean score and addition of feasibility priors. The guarantee is conditional on the stated positive primitive margins; it does not establish that these assumptions hold for every test image. Moreover, the composition contribution is scaled by $1 - \alpha$, so the proposition does not imply that fusion always improves the margin over composition-only scoring.

After fusion, the score is

$$S(x, s, o) = \beta S_\alpha(x, s, o) + (1-\beta) S_E(x, s, o).$$

Writing $\Delta_E(k) = S_E(x, s, o) - S_E(x, s', o')$, its difference is

$$S(x, s, o) - S(x, s', o') = \beta\big[S_\alpha(x, s, o) - S_\alpha(x, s', o')\big] + (1-\beta)\Delta_E(k).$$

An added log feasibility prior contributes its own difference as well. Consequently, positivity of the S_α difference alone does not guarantee the final ranking without further conditions on these additional terms.

B EXPERIMENTAL DETAILS

We use the natural compositional splits for MIT-States, UT-Zappos, and C-GQA, and set the random seed to 42. The physical training batch size is 32, with four gradient-accumulation steps on MIT-States and one on the other two datasets. We use a cosine learning-rate schedule with a minimum learning rate of 10^{-6}. Attention LoRA uses a scaling parameter twice its rank and a dropout rate of 0.05; UT-Zappos additionally uses rank-16 MLP LoRA with a scaling parameter of 32 and a dropout rate of 0.05. The curvature is fixed at $c = 1$. Across datasets, we set $\lambda_{\mathrm{orth}} = 0.2$ and $\lambda_{\mathrm{TC}} = 0.05$.

Table 4: Dataset-specific training hyperparameters.

Dataset	m	w_{hard}	λ_{vis}	λ_{cf}	$\lambda_s = \lambda_o$	Label smoothing
MIT-States	0.15	3.0	0.10	0.10	0.5	0
UT-Zappos	0.05	1.0	0.10	0.10	0.5	0
C-GQA	0.15	1.5	0.15	0.05	0.2	0.05

For text–text closure, we use all candidate pairs on UT-Zappos and sample 256 candidates per training step on MIT-States and C-GQA. The inference weights (γ, α, β), corresponding respectively to the explicit-prototype, primitive-score, and hyperbolic-score weights, are $(0.1, 0.55, 0.5)$ for MIT-States, $(0.1, 0.5, 1.0)$ for UT-Zappos, and $(0.2, 0.3, 0.6)$ for C-GQA. For open-world inference on MIT-States and C-GQA, feasibility scores are min–max normalized, seen compositions are assigned feasibility one, and the log-feasibility prior is added to the fused score.