HyFL-CLIP: Hyperbolic Fine-Tuning of CLIP for Robust Long-Context Understanding

Ji Ha Jang[1,*], Hayeon Kim[1,*], Chulwon Lee[2], Junghun James Kim[2], Se Young Chun[1,2,3,†]

[1]Dept. of Electrical and Computer Engineering, [2]IPAI, [3]INMC & AIIS, Seoul National University, Republic of Korea
{jeeit17, khy5630, chul0e, jonghean12, sychun}@snu.ac.kr

Abstract. CLIP (Contrastive Language-Image Pre-training) has become a de facto paradigm for image-text alignment, but it struggles with long-context descriptions (> 77 tokens) due to absolute positional encoding and pretraining on short captions. In long contexts, sentences are often reordered, summarized, or partially omitted. Although prior works extend CLIP with longer positional encodings, they often suffer from degraded image-text alignment under such text perturbations. We attribute this limitation to the Euclidean contrastive objective, which enforces strict one-to-one matching and lacks explicit mechanisms for modeling hierarchical relationships between global context and its constituent elements. To address this issue, we propose HyFL-CLIP, a hyperbolic fine-tuning framework that distills the well-established text-image alignment learned in Euclidean CLIP into hyperbolic space via cross-manifold similarity distillation, leveraging its geometry to capture hierarchical and entailment relations. Our method models hierarchical semantics by linking summarized token-wise features, long-context descriptions, constituent short textual components, and images, capturing part–whole relationships via hyperbolic entailment with Einstein midpoint aggregation. Experiments on diverse benchmarks, including long-context cross-modal retrieval, cross-modal retrieval with caption perturbations, intra-modality retrieval, and short-text cross-modal retrieval, show that HyFL-CLIP achieves more robust long-context understanding. In particular, it yields up to 19.5% improvement in long-text cross-modal retrieval under textual perturbations over the best prior method. We also show HyFL-CLIP can be seamlessly integrated into other model frameworks by applying it to Stable Diffusion XL (SDXL). The project page is available at `https://janeyeon.github.io/hyflclip`.

Keywords: hyperbolic representation learning · long-context vision-language alignment · cross-manifold distillation

1 Introduction

Vision-language contrastive pre-training has laid the foundation for a wide range of vision-language learning tasks. Models such as CLIP [48], ALIGN [23] have

[*] Authors contributed equally. [†] Corresponding author.

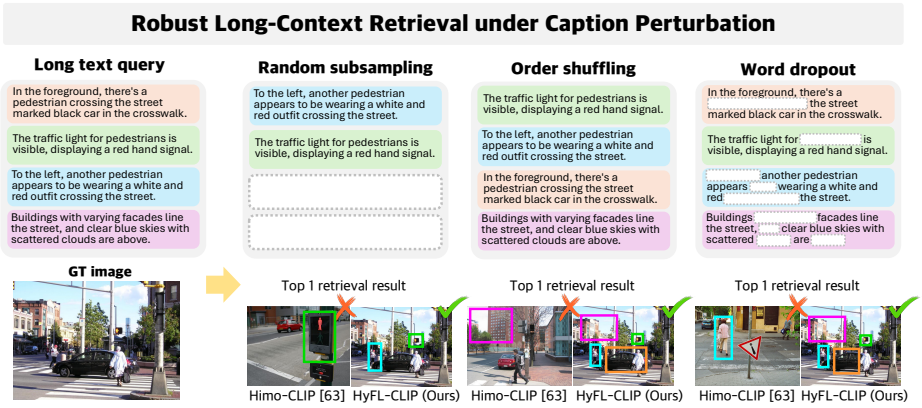

Fig. 1: Robust long-context retrieval under caption perturbation. In long contexts, sentence order or structure may change preserving overall semantics, making robust text-image alignment essential. However, prior works on long-context understanding CLIP often exhibit degraded alignment under such perturbations. For example, even when only the sentence order is shuffled (Order shuffling), they fail to retrieve the correct image, missing key semantic elements such as the red pedestrian traffic signal, the black car, and the pedestrian wearing a white and red outfit.

achieved remarkable performance in text-image alignment by projecting image and text representations into a shared embedding space via contrastive learning. As a result, such models have been widely adopted not only for downstream tasks such as zero-shot classification and retrieval, but also for broader applications including image generation [46, 50] and large multi-modal systems [2, 32, 35, 71].

However, long-text inputs ($>$ 77 tokens) remain challenging for CLIP, as it is primarily trained on datasets composed of short captions and inherently supports a maximum input length of 77 tokens [67]. Existing approaches typically address this by extending positional encoding and aligning coarse and fine grained features in CLIP [3, 41, 63, 67], and designing training strategies to better learn long-text representations [13, 41, 60]. Although prior works on CLIP finetuning for long-context understanding [3, 13, 41, 60, 63, 67] have significantly improved long text-image retrieval performance, we observe that they remain highly sensitive to (i) changes in the number or ordering of sentences that constitute the long context and (ii) the deletion or substitution of only a few words (See Fig. 1). This limitation may partly stem from CLIP's training paradigm, which primarily relies on one-to-one text-image matching and does not explicitly model entailment or part-whole relationships among semantic components. As a result, the global meaning of a long description can become concentrated on a few salient tokens, making the model sensitive to the removal or reordering of tokens or sentences and leading to degraded image–text alignment. Modeling long descriptions therefore requires capturing hierarchical relationships between the global context and its constituent semantic components.

Recently, hyperbolic space has been actively explored in vision–language contrastive learning to better model hierarchical relationships [11,43,49]. Unlike Euclidean space used by CLIP, hyperbolic space has negative curvature and naturally represents tree-like structures. However, most hyperbolic Vision-Language Models (VLMs) [11, 19, 43, 49, 56, 66] or Large Language Models (LLMs) [18] are trained from scratch, limiting their ability to leverage strong pretrained Euclidean models such as CLIP [48]. While recent studies have begun exploring fine-tuning approaches, they typically focus on single-modality adaptation or specific tasks [33, 44, 61, 66, 68, 69].

We propose HyFL-CLIP (Hyperbolic Fine-tuning for Long-context CLIP), which enhances robust long-context understanding by modeling semantic part-whole relationships. HyFL-CLIP transfers the well-established text-image alignment learned by Euclidean CLIP into hyperbolic space. To the best of our knowledge, this is the first work to leverage Euclidean vision–language relationships for generalized tasks in hyperbolic space. We summarize token-wise features using Einstein midpoint aggregation to connect long-context descriptions and their constituent elements with the visual modality. Part-whole relationships are modeled through hierarchical entailment, which relaxes strict one-to-one matching imposed by contrastive learning and aligns semantically related elements to better capture hierarchical structures. We demonstrate that HyFL-CLIP outperforms prior arts [3,13,41,60,63,67] across diverse long-context benchmarks. By effectively leveraging hyperbolic space, our model remains robust even when parts of the input text are missing or perturbed, whereas Euclidean baselines exhibit larger performance drops. Furthermore, HyFL-CLIP achieves up to 19.5% performance improvement on long-context cross-modal retrieval under caption perturbations, highlighting its ability to more robustly capture hierarchical semantic relationships in long contexts. We further adapt HyFL-CLIP to SDXL [46], demonstrating that it can be seamlessly integrated into existing frameworks. Our contributions are as follows:

- We introduce HyFL-CLIP, a hyperbolic fine-tuning framework that transfers the well-established image–text alignment learned by Euclidean CLIP into hyperbolic space, enabling well-pretrained Euclidean Vision–Language Models to operate effectively in hyperbolic representations.
- We bridge Euclidean and hyperbolic representations via cross-manifold similarity distillation, while aligning image and text using a hyperbolic geodesic contrastive loss. A hierarchical entailment loss further relaxes strict one-to-one contrastive matching, enabling the model to better capture relationships between global context and its semantic components.
- HyFL-CLIP achieves strong performance across diverse long-context understanding benchmarks, consistently outperforming Euclidean baselines. In particular, it shows improved robustness under caption perturbations, maintaining stable retrieval performance even when the input text is reordered, partially removed, or corrupted.

2 Related Works

2.1 Vision-language foundation models

Vision–Language Models (VLMs) learn a shared embedding space between images and text, becoming key components for cross-modal understanding [15, 24, 47, 52, 54]. CLIP [48] aligns image and text using a contrastive objective, achieving strong zero-shot performance in tasks such as image–text retrieval and classification. Building on this paradigm, many pretrained VLMs [20, 21, 25, 29, 37] have been developed and widely used for vision–language tasks.

Although contrastive loss of CLIP is powerful, it also has several limitations. First, contrastive learning treats all non-paired samples as negatives in binary [5], even though some may share semantic overlap and act as false negatives in practice [9, 22]. In addition, CLIP-style training often relies heavily on the CLS token for global alignment, which can limit the model's ability to capture hierarchical [16, 38] or fine-grained structural associations [3, 70]. To address these limitations, various approaches in Euclidean space have been explored [3, 9, 16, 22, 70] to improve the contrastive learning framework. In this work, we instead transfer the well-trained similarity geometry of CLIP into hyperbolic space and further refine it using hyperbolic entailment. This allows the model to more robustly retrieve correct image-text pairs even when long-context descriptions are perturbed.

2.2 Long-context understanding with CLIP

CLIP [48] can face challenges when handling text sequences longer than 77 tokens, as it is trained primarily on short captions and uses absolute positional encoding [67]. A common strategy is to extend CLIP to longer sequences using interpolated positional encoding and coarse to fine cross-modal alignment strategies, as explored in Long-CLIP [67] and HiMo-CLIP [63]. Subsequent works further enhance token-wise fine-grained visual–textual correspondence [3]. Other approaches introduce architectural modifications, such as relative positional encoding [41], dual-branch training for jointly handling short and long contexts [60], and dual-teacher distillation frameworks for long-context learning [13].

However, these approaches still rely on Euclidean embeddings with one-to-one image–text alignment and do not explicitly model hierarchical inclusion relationships, which can lead to performance degradation under semantically preserving text modifications. In contrast, we leverage hyperbolic space to explicitly model hierarchical structures that connect summarized token-wise features, long-context descriptions, and the short textual components that compose them. By linking global semantic summaries with their constituent textual elements, the model can robustly preserve semantic understanding even when parts of the text are removed, reordered, or perturbed.

2.3 Hyperbolic representation learning in Vision-language models

Hyperbolic geometry provides an embedding space well suited for modeling fine-grained and hierarchical relationships [14, 53]. Due to its inherent tree-like struc-

ture, hyperbolic space naturally captures hierarchical data. As a result, hyperbolic embeddings have been widely applied in various domains, including graph, image understanding, and text [4, 7, 12, 27, 31, 36, 55, 57, 65].

Recently, several works have explored the use of hyperbolic geometry in VLMs [11, 19, 43, 49, 56, 66]. These models align image and text embeddings using geodesic contrastive losses, while further model hierarchical structure within the same modality [11, 49] or across modalities via part–to–whole relations [43]. Also, some Large Language Models (LLMs) [18] adopt hyperbolic manifold to better capture the hierarchical structure inherent in language.

Although prior works show promising results, they are generally trained from scratch and remain disconnected from well-pretrained Euclidean models, limiting their ability to leverage strong representations from models such as CLIP. While recent attempts address this gap by fine-tuning pretrained CLIP, they are often restricted to single-modality settings [61, 66, 68] or specific tasks and experimental setups [33,44,69], which may limit their applicability in more general settings. In contrast, we perform hyperbolic fine-tuning starting from a pretrained CLIP model. By guiding the model with an entailment objective, our approach preserves CLIP's strong representations while enabling the text embedding space to capture relationships across different levels of semantic abstraction, leading to improved performance in long-context understanding.

3 Preliminaries

3.1 Hyperbolic geometry in the Lorentz model

Hyperbolic space is a Riemannian manifold with constant negative curvature $-\kappa$, where $\kappa \in \mathbb{R}^+$. In this work, we employ the Lorentz (Minkowski) model as the underlying geometric space for hyperbolic fine-tuning, by distilling representations from a pre-trained Open-CLIP [10] model into the Lorentz manifold.

We consider the $(n+1)$-dimensional Minkowski space $\mathbb{R}^{n+1}$ equipped with the Lorentzian inner product defined as below:

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} = -p_{\text{time}} q_{\text{time}} + \langle \mathbf{p}_{\text{space}}, \mathbf{q}_{\text{space}} \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product.

Under this metric, the n-dimensional Lorentz manifold $\mathbb{L}^n$ is defined as the upper sheet of the two-sheeted hyperboloid given by:

$$\mathbb{L}^n = \left\{ \mathbf{p} \in \mathbb{R}^{n+1} \;\middle|\; \langle \mathbf{p}, \mathbf{p} \rangle_{\mathbb{L}} = -\frac{1}{\kappa},\ p_{\text{time}} > 0 \right\}. \tag{2}$$

Each point $\mathbf{p} \in \mathbb{L}^n$ is represented as below:

$$\mathbf{p} = [p_{\text{time}}, \mathbf{p}_{\text{space}}], \qquad p_{\text{time}} = \sqrt{\frac{1}{\kappa} + \|\mathbf{p}_{\text{space}}\|^2}. \tag{3}$$

where $\mathbf{p}_{\text{space}} \in \mathbb{R}^n$ denotes the spatial component, $||\cdot||$ denotes the Euclidean norm, and the expression for p_{time} follows from the Lorentz manifold constraint.

The geodesic distance between two points $\mathbf{p}, \mathbf{q} \in \mathbb{L}^n$ is given by:

$$d_{\mathbb{L}}(\mathbf{p}, \mathbf{q}) = \sqrt{1/\kappa}\, \cosh^{-1}\left(-\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}}\right). \tag{4}$$

The hyperbolic radius of an embedding $\mathbf{p}$ corresponds to its hyperbolic distance from the hyperboloid origin $\mathbf{o}$, measured by $d_{\mathbb{L}}(\mathbf{p}, \mathbf{o})$.

3.2 Tangent spaces

For each point $\mathbf{z} \in \mathbb{L}^n$, the tangent space at $\mathbf{z}$ is defined as:

$$T_{\mathbf{z}}\mathbb{L}^n = \left\{ \mathbf{v} \in \mathbb{R}^{n+1} : \langle \mathbf{z}, \mathbf{v} \rangle_{\mathbb{L}} = 0 \right\}, \tag{5}$$

which forms an n-dimensional Euclidean vector space consisting of all vectors orthogonal to $\mathbf{z}$ under the Lorentzian inner product. A tangent vector $\mathbf{v} \in T_{\mathbf{z}}\mathbb{L}^n$ can be mapped back to the hyperboloid via the exponential map,

$$\exp_{\mathbf{z}}^{\kappa}(\mathbf{v}) = \cosh(\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}})\mathbf{z} + \frac{\sinh(\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}})}{\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}}} \mathbf{v}. \tag{6}$$

In contrast, the logarithmic map transports $\mathbf{p} \in \mathbb{L}^n$ to the tangent space at $\mathbf{z}$ as:

$$\log_{\mathbf{z}}^{\kappa}(\mathbf{p}) = \frac{\cosh^{-1}(-\kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}})}{\sqrt{(\kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}})^2 - 1}}\, \mathrm{proj}_{\mathbf{z}}(\mathbf{p}), \tag{7}$$

where $\mathrm{proj}_{\mathbf{z}}(\mathbf{p}) = \mathbf{p} + \kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}} \mathbf{z}$ denotes the projection of $\mathbf{p}$ onto the tangent space $T_{\mathbf{z}}\mathbb{L}^n$. In our formulation, the reference point $\mathbf{z}$ is set to the hyperboloid origin $\mathbf{o} = [\sqrt{1/\kappa}, \mathbf{0}]$. At this point, the tangent space $T_{\mathbf{o}}\mathbb{L}^n$ reduces to an n-dimensional Euclidean space, as tangent vectors have zero time components and are fully parameterized by their spatial coordinates, consistent with standard practice in hyperbolic representation learning [11, 43, 49].

3.3 Einstein midpoint

In hyperbolic space, the centroid of a set of points is computed via the Einstein midpoint. We first project hyperbolic points from the hyperboloid model $\mathbb{L}^n$ to the Klein model $\mathbb{K}^n$, where a weighted average admits a closed-form expression. The Lorentz-Klein projection and its inverse are given by:

$$\Pi_{\mathbb{L}\to\mathbb{K}}(\mathbf{p}) = \frac{\mathbf{p}_{\text{space}}}{p_{\text{time}}} = \frac{\mathbf{p}_{\text{space}}}{\sqrt{\frac{1}{\kappa} + \|\mathbf{p}_{\text{space}}\|^2}}, \qquad \Pi_{\mathbb{K}\to\mathbb{L}}(\mathbf{k}) = \frac{(1, \mathbf{k})}{\sqrt{\kappa(1 - \|\mathbf{k}\|^2)}}, \tag{8}$$

where $\Pi_{\mathbb{L}\to\mathbb{K}}$ maps a point $\mathbf{p} \in \mathbb{L}^n$ to its Klein coordinate representation, and $\Pi_{\mathbb{K}\to\mathbb{L}}$ denotes the inverse projection that lifts a Klein point $\mathbf{k}$ back onto the Lorentz hyperboloid.

Given a set of points $\{\mathbf{x}_j\}_{j=1}^N \subset \mathbb{L}^n$, their Einstein midpoint $\bar{\mathbf{x}}$ is obtained by computing a Lorentz-factor-weighted mean in Klein coordinates as:

$$\bar{\mathbf{x}} = \Pi_{\mathbb{K}\to\mathbb{L}} \left(\frac{\sum_{j=1}^N \gamma_j\, \Pi_{\mathbb{L}\to\mathbb{K}}(\mathbf{x}_j)}{\sum_{j=1}^N \gamma_j} \right), \qquad \gamma_j = \frac{1}{\sqrt{1 - \kappa\, \|\Pi_{\mathbb{L}\to\mathbb{K}}(\mathbf{x}_j)\|^2}}. \tag{9}$$

4 Method

In this section, we introduce our method HyFL-CLIP. We first describe the problem setting, and then present the key training objectives of our approach: short-text guided cross-manifold similarity distillation, hyperbolic geodesic contrastive learning, hierarchical entailment with Einstein midpoint aggregation, and an entropy regularizer for stabilizing hyperbolic embeddings.

4.1 Problem formulation

HyFL-CLIP fine-tunes a pre-trained CLIP model, which uses separate encoders for images and texts to produce aligned representations in a shared embedding space. We denote the model as f, consisting of image encoder f_v and a text encoder f_t. For a text-image pair (I, T), the visual encoder produces a set of embeddings $f_v(I) \in \mathbb{R}^{(K+1)\times n}$, consisting of a class token $\tilde{\mathbf{v}}$ that captures global image representation and a set of token-wise embeddings $\{\tilde{\mathbf{v}}_k\}_{k=1}^{K}$ that encode local visual information. Similarly, the text encoder produces a set of embeddings $f_t(T) \in \mathbb{R}^{(L+1)\times n}$, consisting of a sentence-level token $\tilde{\mathbf{t}}$ that represents the overall semantic meaning of the text and token-wise embeddings $\{\tilde{\mathbf{t}}_l\}_{l=1}^{L}$ that encode token-level linguistic information.

In the original CLIP model, the input text length is limited to 77 tokens due to the use of learned absolute positional embedding. Before being fed into f_t, text tokens are truncated to the first 77 tokens. To enable CLIP models to handle longer contexts, prior works [3, 13, 41, 60, 63, 67] extend the positional embedding via interpolation to support longer input sequences (*e.g.*, up to 248 tokens).

Our goal is to improve long-context understanding by explicitly modeling hierarchical semantic relationships within long descriptions in hyperbolic space, while preserving the well-established short text-image alignment learned by CLIP.

4.2 Short-text guided cross-manifold similarity distillation

In pre-trained CLIP, the similarity between short text–image pairs is well learned. Inspired by prior similarity-based distillation methods that transfer relational structure through pairwise similarity distributions [58, 62], we use this well-established geometry as a reference when transferring the learned embedding space of CLIP into hyperbolic space. Unlike prior methods that typically distill similarities within the same geometric space, we perform cross-manifold similarity distillation, which aligns the similarity distributions of the Euclidean teacher and the hyperbolic student. Formally, let $\tilde{\mathbf{v}}_i \in \mathbb{R}^n$ denote the image embedding, and let $\tilde{\mathbf{t}}_i^s, \tilde{\mathbf{t}}_i^l \in \mathbb{R}^n$ denote the embeddings of the short and long texts corresponding to the same image, obtained from a pre-trained CLIP model that serves as the Euclidean teacher. Since CLIP is originally trained on short text–image pairs, we perform the distillation using the short-text embeddings to preserve the well-learned similarity geometry. We project these Euclidean embeddings

into hyperbolic space via the exponential map at the origin: $\mathbf{v}_i = \exp_{\mathbf{o}}^{\kappa}(\tilde{\mathbf{v}}_i), \mathbf{t}_i^s = \exp_{\mathbf{o}}^{\kappa}(\tilde{\mathbf{t}}_i^s), \mathbf{t}_i^l = \exp_{\mathbf{o}}^{\kappa}(\tilde{\mathbf{t}}_i^l)$, where $\mathbf{v}_i, \mathbf{t}_i^s, \mathbf{t}_i^l \in \mathbb{L}^n$.

We define the Euclidean teacher similarity S^{E} as the cosine similarity between the short-text and image embeddings. In hyperbolic space, the student similarity S^{H} is defined as the negative geodesic distance in the Lorentz model (Eq. (4)). Formally,

$$S^{\mathrm{E}}(\tilde{\mathbf{t}}_i^s, \tilde{\mathbf{v}}_j) = \frac{\langle \tilde{\mathbf{t}}_i^s, \tilde{\mathbf{v}}_j \rangle}{\|\tilde{\mathbf{t}}_i^s\| \|\tilde{\mathbf{v}}_j\|}, \qquad S^{\mathrm{H}}(\mathbf{t}_i^s, \mathbf{v}_j) = -\, d_{\mathbb{L}}(\mathbf{t}_i^s, \mathbf{v}_j)\,. \tag{10}$$

We construct probability distributions over image candidates induced by the similarity scores. Given a similarity matrix S and temperature τ, the distribution is defined as:

$$P(S, \tau)_{ij} = \frac{\exp(S_{ij}/\tau)}{\sum_k \exp(S_{ik}/\tau)}. \tag{11}$$

The Euclidean teacher distribution and hyperbolic student distribution are then obtained as $P^{\mathrm{E}} = P(S^{\mathrm{E}}(\tilde{\mathbf{t}}^s, \tilde{\mathbf{v}}), \tau_E), P^{\mathrm{H}} = P(S^{\mathrm{H}}(\mathbf{t}^s, \mathbf{v}), \tau_H)$. The cross-manifold distillation loss is defined as the Kullback-Leibler divergence between the teacher and student distributions, given by:

$$\mathcal{L}_{\mathrm{distill}} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{KL}\left(P_{i\cdot}^{\mathrm{E}} \parallel P_{i\cdot}^{\mathrm{H}}\right), \tag{12}$$

where B denotes the batch size, i indexes the query samples in the mini-batch, and $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence.

4.3 Hyperbolic geodesic contrastive loss

After transferring Euclidean geometric relations through short text–image pairs, we further optimize the model using long text–image pairs through a geodesic contrastive objective on the Lorentz manifold $\mathbb{L}^n$. Following the hyperbolic InfoNCE formulation of [11], we define the image-text contrastive loss $\mathcal{L}_{\mathrm{itc}}$ in Lorentz space as follows:

$$L_{\mathrm{info}}(\mathbf{v}, \mathbf{t}; \tau_c) = -\sum_i \log \frac{\exp\left(-d_{\mathbb{L}}(\mathbf{v}_i, \mathbf{t}_i)/\tau_c\right)}{\sum_{k \neq i} \exp\left(-d_{\mathbb{L}}(\mathbf{v}_i, \mathbf{t}_k)/\tau_c\right)}, \tag{13}$$

where $(\mathbf{v}_i, \mathbf{t}_k)$ denotes the matched text-image pair in the batch, while the remaining text embeddings $\{\mathbf{t}_k\}_{k \neq i}$ serve as negative samples. The temperature parameter τ_c controls the sharpness of the softmax distribution.

We employ a bidirectional contrastive objective for both short text–image and long text–image pairs, defined as:

$$\mathcal{L}_{v \leftrightarrow t}^{s} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{info}}(\mathbf{v}, \mathbf{t}^s; \tau_c) + \mathcal{L}_{\mathrm{info}}(\mathbf{t}^s, \mathbf{v}; \tau_c)\right), \tag{14}$$

$$\mathcal{L}_{v \leftrightarrow t}^{\ell} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{info}}(\mathbf{v}, \mathbf{t}^{\ell}; \tau_c) + \mathcal{L}_{\mathrm{info}}(\mathbf{t}^{\ell}, \mathbf{v}; \tau_c)\right). \tag{15}$$

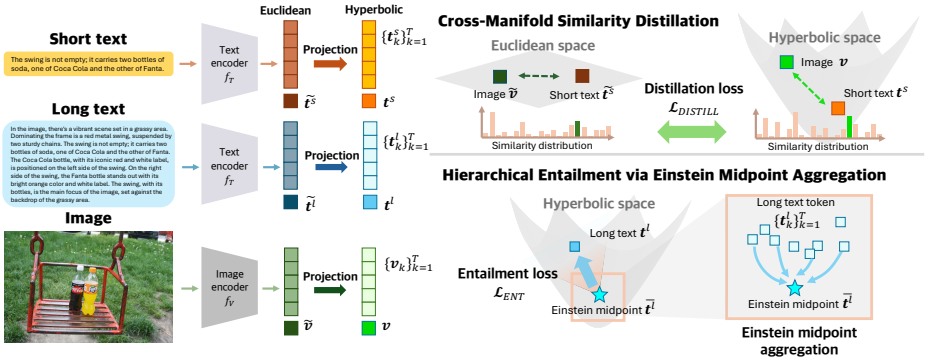

Fig. 2: Overview of our HyFL-CLIP framework. Our HyFL-CLIP transfers Euclidean text–image alignment into hyperbolic space via a short-text guided cross-manifold similarity distillation. Hierarchical entailment with Einstein midpoint aggregation abstracts token-wise information within each modality and aligns it with a global representation. Hyperbolic geodesic contrastive loss aligns both long texts and their semantic components with the corresponding image, while an entropy regularizer stabilizes the embedding distribution.

The final hyperbolic geodesic contrastive objective is given by:

$$\mathcal{L}_{\text{itc}} = \mathcal{L}^{\ell}_{v\leftrightarrow t} + \lambda_1 \mathcal{L}^{s}_{v\leftrightarrow t}, \tag{16}$$

where λ_1 is a hyperparameter.

4.4 Hierarchical entailment via Einstein midpoint aggregation

Long-context understanding requires capturing a global representation as well as modeling the individual components and their semantic inclusion within the same modality. To capture such hierarchical structure, we construct a parent representation by aggregating token-wise features via a similarity-weighted Einstein midpoint. The weights are determined by their semantic similarity to the global representation from the opposite modality. We then enforce a hyperbolic entailment constraint between the aggregated representation and the corresponding global embedding.

For clarity, we first describe the formulation for the long-text $\mathbf{t}^{\ell}$; the same procedure is applied to the image modality $\mathbf{v}$. Let $\{\mathbf{t}^{\ell}_{i,k}\}_{k=1}^{K} \subset \mathbb{L}^n$ denote the long-text token embeddings of the i-th sample. To measure the semantic importance of each token-level feature, we compute attention weights $\alpha_{i,k}$ as follows:

$$\alpha_{i,k} = \frac{\exp\left(-d_{\mathbb{L}}(\mathbf{t}^{\ell}_{i,k}, \mathbf{v}_i)/\tau_{\text{ent}}\right)}{\sum_m \exp\left(-d_{\mathbb{L}}(\mathbf{t}^{\ell}_{i,m}, \mathbf{v}_i)/\tau_{\text{ent}}\right)}, \qquad \sum_k \alpha_{i,k} = 1. \tag{17}$$

Then we abstract token-wise feature by calculating weighted Einstein midpoint, multiplying $\alpha_{i,k}$ to the Lorentz factor in Eq. (9). The weighted Einstein midpoint

is given as:

$$\bar{\mathbf{t}}_i^{\ell} = \Pi_{\mathbb{K}\to\mathbb{L}}\left(\frac{\sum_k \alpha_{i,k}\gamma_{i,k}\Pi_{\mathbb{L}\to\mathbb{K}}(\mathbf{t}_{i,k}^{\ell})}{\sum_k \alpha_{i,k}\gamma_{i,k}}\right), \qquad \bar{\mathbf{t}}_i^{\ell} \in \mathbb{L}^n. \tag{18}$$

We interpret the aggregated representation $\bar{\mathbf{t}}_i^{\ell}$ as a more general semantic concept and enforce that the corresponding global embedding $\mathbf{t}_i^{\ell} \in \mathbb{L}^n$ lies within its entailment cone. Following the hyperbolic entailment formulation in [14, 31], the half-aperture of the cone centered at $\bar{\mathbf{t}}_i^{\ell}$ is defined as:

$$\omega(\bar{\mathbf{t}}_i^{\ell}) = \arcsin\left(\frac{2K}{\sqrt{\kappa}\,\|\bar{\mathbf{t}}_i^{\ell}\|_{\mathbb{L}}}\right), \tag{19}$$

where K is a constant controlling stability near the origin. We enforce the entailment constraint by penalizing violations of the cone boundary using the following loss (see Fig. 3):

$$\mathcal{L}_{\text{ent}}^{\mathbf{t}^{\ell}} = \frac{1}{B}\sum_{i=1}^{B} \max\left(0, \phi(\bar{\mathbf{t}}_i^{\ell}, \mathbf{t}_i^{\ell}) - \eta\,\omega(\bar{\mathbf{t}}_i^{\ell})\right), \tag{20}$$

where $\phi(\cdot,\cdot)$ denotes the hyperbolic angle between two embeddings in the Lorentz model. The above formulation is symmetrically applied to the image embeddings $\mathbf{v}$, yielding $\mathcal{L}_{\text{ent}}^{\mathbf{v}}$. The final hierarchical entailment loss is defined as:

$$\mathcal{L}_{\text{ent}} = \mathcal{L}_{\text{ent}}^{\mathbf{t}} + \mathcal{L}_{\text{ent}}^{\mathbf{v}}. \tag{21}$$

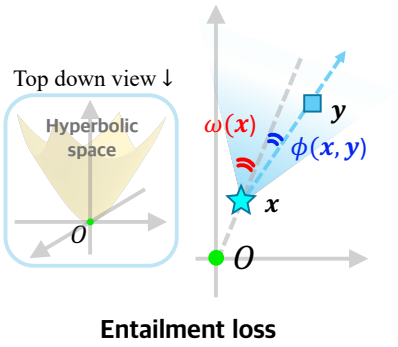

Fig. 3: Entailment loss in hyperbolic space. Adapted from the MERU [11], this figure illustrates the entailment loss in hyperbolic space. The $\phi(\mathbf{x}, \mathbf{y})$ measures the geodesic angle between the two embeddings, and $\omega(\mathbf{x})$ denotes the aperture of the entailment cone centered at $\mathbf{x}$. Geodesic angle between the two embeddings is used to determine if $\mathbf{y}$ lies within the entailment region of $\mathbf{x}$.

Thus, our final loss is given as:

$$\mathcal{L} = \lambda_2\mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{itc}} + \lambda_3\mathcal{L}_{\text{ent}} + \lambda_4\mathcal{L}_{\text{reg}}. \tag{22}$$

Inspired by entropy-based regularization strategies used in prior works [17, 28], we introduce a radius-entropy regularization term:

$$\mathcal{L}_{\text{reg}} = -H(\mathbf{p}), \qquad p_i = \frac{\exp\big(d_{\mathbb{L}}(\mathbf{t}_i^{\ell}, \mathbf{o})\big)}{\sum_j \exp\big(d_{\mathbb{L}}(\mathbf{t}_j^{\ell}, \mathbf{o})\big)}. \tag{23}$$

Here, $H(\mathbf{p}) = -\sum_i p_i \log p_i$ denotes the entropy of the hyperbolic radius distribution. This regularization promotes balanced utilization of hyperbolic radii. More details on the hyperparameter choices are in the supplementary material.

Intuitively, the entailment cone provides a geometric margin of tolerance. The global text embedding is trained to lie within the cone of the aggregated midpoint, which summarizes token-level features. When a perturbation removes or reorders tokens, the midpoint shifts, but the cone's nonzero aperture (Eq. 19) allows $\mathbf{t}_i^\ell$ to remain within the entailment region, preserving alignment. By contrast, Euclidean contrastive objectives enforce point-to-point matching with no such margin, making them sensitive to any shift in the aggregated representation.

5 Experiments

5.1 Experimental setup

Training details. Following Long-CLIP [67], we train HyFL-CLIP on dataset ShareGPT4V [8], which contains 1.2 million image–caption pairs with multi-sentence annotations and long captions averaging 143.6 words. The batch size is 1024 and we train our models for 2 epochs. We set the learning rate to 1×10^{-5} and the weight decay to 2.5×10^{-2}, and optimize the model using AdamW. Our model builds upon Open-CLIP, and we experiment with two CLIP architectures, CLIP-ViT-B/16 and CLIP-ViT-L/14.

Evaluation setup. We evaluate our model against other baselines [3, 13, 41, 60, 63, 64, 67] on four different tasks:
(1) Zero-shot long/short caption cross modality retrieval. We evaluate HyFL-CLIP on zero-shot long-caption text-image retrieval using DOCCI [42], DCI [59], Long-DCI [41], and Urban-1k [67], whose captions average 131.4 to 174.2 tokens. Performance was measured using Top-1 retrieval accuracy. To examine whether HyFL-CLIP successfully distills the text–image similarity geometry learned from short captions while adapting to long captions, we further evaluate short-text retrieval on COCO [34] and Flickr30K [45].
(2) Zero-shot long-caption cross-modal retrieval under caption perturbation. Using all datasets from Task (1), we perturb the captions to test whether models can robustly retrieve the correct images under semantic-preserving modifications. The perturbations include random word dropping ($p = 0.5$), random subsampling of $n = 2$ or $n = 3$ sentences, sentence order shuffling, and removal of the first sentence. Each model is evaluated five times per perturbation, and Top-1 retrieval accuracy is averaged across all datasets.
(3) Text-to-text intra modality retrieval. We perform text-to-text retrieval following the evaluation protocol of [40]. For COCO [34], Flickr30K [45], and nocaps [1] datasets, we ignore the images and use the first caption of each image as the query, while the remaining captions of the same image are treated as positives. We use the Karpathy split [26] for COCO and Flickr30K, and the validation split for nocaps. We further evaluate on purely textual datasets, including 20 Newsgroups [30] and IMDB Reviews [39].
(4) Text-to-image generation. To qualitatively evaluate how HyFL-CLIP can be integrated into different model frameworks, we apply it to Stable Diffusion

XL [46] (SDXL). Since SDXL is originally trained with Euclidean text prompts, we encode text prompts from Long-DCI [41], DrawBench [51] and project them to the Euclidean tangent space using Eq. (7) before image generation. We compare the results with Long-CLIP [67] integrated with SDXL.

5.2 Experimental results

Zero-shot long-caption cross-modal retrieval. Tab. 1 presents long caption cross-modal retrieval results. HyFL-CLIP (Ours) consistently outperforms Euclidean baselines across datasets and architectures. Comparisons with state-of-the-art hyperbolic VLMs [11, 28, 43] are in the supplementary material.

Table 1: Comparison of zero-shot long-caption cross-modal retrieval. HyFL-CLIP (Ours) consistently outperforms existing long-context CLIP baselines across all datasets and model architectures. The best and second-best results are highlighted in **bold** and underline, respectively. We report numbers from the original papers when available; otherwise, we evaluate using our own implementation (marked with ∗). Results marked with † are obtained using the checkpoints provided by the original authors.

		DOCCI [42]		DCI [59]		Long-DCI [41]		Urban-1k [67]	
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I
ViT-B-16	Long-CLIP† [67]	63.10	71.49	59.88	61.28	42.21	48.38	79.40	79.60
	TULIP [41]	-	-	-	-	50.20	50.60	88.10	86.60
	HiMo-CLIP* [63]	77.37	79.35	71.09	69.93	58.59	57.00	89.20	89.20
	FineLIP* [3]	77.16	79.14	69.38	68.03	57.18	55.22	89.30	86.90
	LongD-CLIP [13]	-	-	-	-	-	-	87.20	87.30
	SmartCLIP [64]	77.40	78.00	64.90	64.00	53.40	52.80	90.00	87.40
	Fix-CLIP [60]	-	-	59.70	63.00	-	-	80.90	81.10
	HyFL-CLIP(Ours)	**78.41**	**81.12**	**71.54**	**71.79**	**59.00**	**58.75**	**91.80**	**91.10**
ViT-L-14	Long-CLIP† [67]	66.78	78.61	64.13	67.83	46.55	54.25	82.40	86.20
	TULIP [41]	77.90	79.10	-	-	55.70	56.40	90.10	91.10
	HiMo-CLIP† [63]	**82.35**	84.59	74.59	74.54	**62.06**	61.94	93.00	93.20
	FineLIP [3]	82.20	83.10	-	-	60.80	60.70	93.20	93.00
	LongD-CLIP [13]	-	-	-	-	-	-	91.90	90.80
	SmartCLIP [64]	81.60	82.50	68.20	69.80	57.60	58.50	93.30	90.10
	Fix-CLIP [60]	-	-	65.10	66.70	-	-	86.80	87.70
	HyFL-CLIP(Ours)	82.12	**85.39**	**74.74**	**76.19**	61.92	**63.93**	**94.60**	**94.30**

Zero-shot long-caption cross-modal retrieval under caption perturbation. Tab. 2 presents zero-shot long-caption cross-modal retrieval results under caption perturbations. The table reports the relative performance change (in %). HyFL-CLIP (Ours) consistently outperforms other baselines across all datasets and exhibits the smallest performance drop. This indicates that the hierarchical entailment loss enables the model to better capture part–whole semantic relationships and remain robust to perturbations that preserve the underlying semantic meaning. Full results for all datasets are in the supplementary material.

Table 2: Comparison of zero-shot long-caption cross-modal retrieval under caption perturbation. Under caption perturbations, other models exhibit large performance drops, while our method maintains strong performance across all types and degrees of perturbations. The small numbers shown beside each score indicate the percentage performance differences (%) relative to the original score. Notation follows Tab. 1 (∗: our implementation; †: author-provided checkpoints).

Model	Word Dropout	Sent. Removal	Order Shuffling	Random Subsampling	
	$p = 0.5$	first	random	$n = 2$	$n = 3$
Long-CLIP† [67]	48.88 ↓35.81	55.41 ↓19.45	61.79 ↓3.46	26.50 ↓91.89	31.29 ↓79.89
HiMo-CLIP* [63]	58.75 ↓27.82	64.44 ↓17.42	72.94 ↓1.88	28.24 ↓83.58	34.84 ↓71.52
FineLIP* [3]	58.70 ↓26.59	64.18 ↓16.25	73.41 ↑1.17	27.20 ↓86.05	33.41 ↓74.32
HyFL-CLIP (Ours)	**70.20** ↓9.21	**68.22** ↓12.69	**76.16** ↑1.27	**32.55** ↓75.36	**39.05** ↓63.94

Zero-shot short caption cross-modal retrieval. Tab. 3 presents the results of zero-shot short-caption cross-modal retrieval on COCO and Flickr30k [34, 45] datasets. These results indicate that HyFL-CLIP successfully distills the pre-trained text-image similarity geometry while being optimized in hyperbolic space, without sacrificing its base long-caption retrieval performance.

Table 3: Comparison of zero-shot short-caption cross-modal retrieval. HyFL-CLIP achieves comparable or better short-caption retrieval performance than Euclidean baselines. Notation follows Tab. 1 (∗: our implementation; †: author-provided checkpoints).

		COCO [34]		Flickr30k [45]	
		I2T	T2I	I2T	T2I
ViT-B/16	Long-CLIP† [67]	57.26	40.38	47.19	33.16
	TULIP [41]	56.80	40.70	46.10	**35.20**
	HiMo-CLIP* [63]	**60.84**	40.71	50.27	34.02
	FineLIP* [3]	58.32	40.05	**52.23**	33.85
	HyFL-CLIP (Ours)	58.20	**41.50**	50.60	**35.20**

Table 4: Comparison of text-to-text intra modality retrieval. Intra-modal retrieval performance across five datasets (Values in %). Bold indicates the best, and underlined indicates the second-best. Notation follows Tab. 1 (∗: our implementation; †: author-provided checkpoints).

	Flickr30K [45]		COCO [34]		nocaps [1]		IMDB [39]		20News [30]	
Model	mAP	Pr@R	mAP	Pr@R	mAP	Pr@R	mAP	Pr@R	mAP	Pr@R
Long-CLIP† [67]	52.31	47.56	27.51	24.58	37.16	35.47	51.88	50.41	26.78	30.40
HiMo-CLIP* [63]	58.02	52.66	29.36	26.03	40.24	38.08	**52.65**	**50.59**	35.49	37.61
FineLIP* [3]	52.55	47.78	24.87	22.26	34.10	33.15	51.78	50.40	17.65	21.24
HyFL-CLIP(Ours)	**60.63**	**54.98**	**30.69**	**27.12**	**41.16**	**38.56**	52.60	**50.59**	**36.69**	**38.32**

Fig. 4: Embeddings and token weight visualization. We visualize the embedding distribution of text summary token (Einstein midpoint), text, and image using HoroPCA [6]. We also compare text token contribution weights based on their similarity to the image.

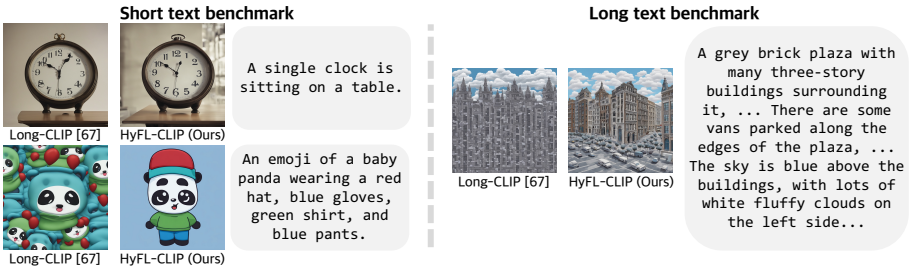

Fig. 5: SDXL text-to-image generation results. We replace the original text encoder with ours. Hyperbolic embeddings are mapped to the Euclidean space using the logarithmic map transport. Our model captures finer details compared to the baseline.

Text-to-text intra modality retrieval. Tab. 4 presents the results of text-to-text intra-modality retrieval across five datasets. HyFL-CLIP (Ours) consistently achieves superior performance, demonstrating its strong capability in capturing semantic relationships within textual representations. By leveraging hierarchical entailment within the same modality, HyFL-CLIP better captures semantic abstraction between local textual components and global representations, leading to improved retrieval performance across diverse datasets.

Text-to-image generation. Fig. 5 presents qualitative text-to-image generation results using SDXL [46] conditioned on both short captions from DrawBench [51] and long captions from Long-DCI [41]. As shown in Fig. 5, our model captures finer-grained details, such as the number of clocks and subtle attributes of the panda, compared to the baseline model. When long prompts are used, Long-CLIP [67] often ignores or corrupts parts of the description, whereas our model better preserves the full semantic context and generates more faithful results. These results indicate that our framework successfully distills the Euclidean similarity structure to produce more expressive hyperbolic embeddings.

5.3 Ablation studies

We conduct an ablation study in ViT-L/14 to evaluate the contribution of each component in our framework. First, removing $\mathcal{L}_{\text{ent}}$ reduces retrieval accuracy from 69.8% to 69.3%, performance averaged across all benchmarks [34,41,42,45, 59,67]. Next, we further remove $\mathcal{L}_{\text{distill}}$, which results in 68.3% performance. Full results are in the supplementary material. Fig. 4 shows embedding visualizations of our final model and a comparison of token contribution weights. HyFL-CLIP (Ours) assigns higher weights to semantically meaningful tokens such as street, bicycles, and city, which directly correspond to the key visual elements in the image.

6 Conclusion

We propose HyFL-CLIP, a framework that enables robust long-context understanding in CLIP by distilling the well-established text–image similarity geometry of a pre-trained Euclidean CLIP model into hyperbolic space. By leveraging hyperbolic geometry, HyFL-CLIP addresses the limitations of Euclidean contrastive learning by modeling hierarchical and part–whole relationships between global captions and their constituent elements. This design encourages the model to capture semantic relationships between whole captions and their constituent parts, leading to more robust representations under caption perturbations. Extensive experiments on long- and short-caption cross-modal retrieval as well as text-to-text intra-modality retrieval demonstrate state-of-the-art performance, highlighting the effectiveness of hyperbolic hierarchical representations for long-context understanding.

Acknowledgements

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government(MSIT) [No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University) / No.RS-2025-02314125, Effective Human-Machine Teaming With Multimodal Hazy Oracle Models], the National Research Foundation of Korea(NRF) grants funded by the Korea government(MSIT) (Nos. RS-2022-NR067592, RS-2025-02263628), the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT), Republic of Korea (No. RQT-25-120066), the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University, AI-Bio Research Grant through Seoul National University and the AI Seoul Tech Research Support Program of the Seoul Future Foundation.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: CVPR (2019)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. NeurIPS (2022)
3. Asokan, M., Wu, K., Albreiki, F.: Finelip: Extending clip's reach via fine-grained alignment with longer text inputs. In: CVPR (2025)
4. Atigh, M.G., Schoep, J., Acar, E., van Noord, N., Mettes, P.: Hyperbolic image segmentation. In: CVPR (2022)
5. Byun, J., Kim, D., Moon, T.: Mafa: Managing false negatives for vision-language pre-training. In: CVPR (2024)
6. Chami, I., Gu, A., Nguyen, D.P., Ré, C.: Horopca: Hyperbolic dimensionality reduction via horospherical projections. In: ICML (2021)
7. Chami, I., Ying, Z., Ré, C., Leskovec, J.: Hyperbolic graph convolutional neural networks. NeurIPS **32** (2019)
8. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: ECCV (2024)
9. Chen, T.S., Hung, W.C., Tseng, H.Y., Chien, S.Y., Yang, M.H.: Incremental false negative detection for contrastive learning. ICLR (2022)
10. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR (2023)
11. Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, S.R.: Hyperbolic image-text representations. In: ICML (2023)
12. Dhingra, B., Shallue, C., Norouzi, M., Dai, A., Dahl, G.: Embedding text in hyperbolic spaces. In: TextGraphs-12 (2018)
13. Feng, Y., Wen, C., Peng, Z., Zhu, S., et al.: Retaining knowledge and enhancing long-text representations in clip through dual-teacher distillation. In: CVPR (2025)
14. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic entailment cones for learning hierarchical embeddings. In: ICML. PMLR (2018)
15. Ge, Y., Ren, J., Gallagher, A., Wang, Y., Yang, M.H., Adam, H., Itti, L., Lakshminarayanan, B., Zhao, J.: Improving zero-shot generalization and robustness of multi-modal models. In: CVPR (2023)
16. Geng, S., Yuan, J., Tian, Y., Chen, Y., Zhang, Y.: Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. ICLR (2023)
17. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. NeurIPS (2004)
18. He, N., Anand, R., Madhu, H., Maatouk, A., Krishnaswamy, S., Tassiulas, L., Yang, M., Ying, R.: Helm: Hyperbolic large language models via mixture-of-curvature experts. NeurIPS (2025)
19. He, N., Yang, M., Ying, R.: Hypercore: The core framework for building hyperbolic foundation models with comprehensive modules. arXiv preprint arXiv:2504.08912 (2025)
20. He, X., Peng, Y.: Fine-grained image classification via combining vision and language. In: CVPR (2017)

21. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020)
22. Huynh, T., Kornblith, S., Walter, M.R., Maire, M., Khademi, M.: Boosting contrastive self-supervised learning with false negative cancellation. In: CVPR (2022)
23. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
24. Jin, L., Luo, G., Zhou, Y., Sun, X., Jiang, G., Shu, A., Ji, R.: Refclip: A universal teacher for weakly supervised referring expression comprehension. In: CVPR (2023)
25. Kang, D.U., Kim, H., Chun, S.Y.: Cdam: Class distribution-induced attention map for open-vocabulary semantic segmentations. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025), iCLR
26. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
27. Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., Lempitsky, V.: Hyperbolic image embeddings. In: CVPR (2020)
28. Kim, H., Jang, J.H., Kim, J.J., Chun, S.Y.: Uncertainty-guided compositional alignment with part-to-whole semantic representativeness in hyperbolic vision-language models. In: CVPR (2026)
29. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
30. Lang, K.: Newsweeder: Learning to filter netnews. In: Machine learning proceedings 1995 (1995)
31. Le, M., Roller, S., Papaxanthos, L., Kiela, D., Nickel, M.: Inferring concept hierarchies from text corpora via hyperbolic embeddings. In: ACL (2019)
32. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
33. Li, L., Li, J., Lei, J., Xiao, J., Shao, F., Chen, L.: Learning hierarchical hyperbolic embeddings for compositional zero-shot learning. arXiv preprint arXiv:2512.20029 (2025)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2023)
36. Liu, Q., Nickel, M., Kiela, D.: Hyperbolic graph neural networks. NeurIPS **32** (2019)
37. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS **32** (2019)
38. Lv, X., Zhao, Y., Yin, H., Chen, Y., Liu, J.: Msg-clip: Enhancing clip's ability to learn fine-grained structural associations through multi-modal scene graph alignment. Pattern Recognition
39. Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (2011)
40. Mistretta, M., Baldrati, A., Agnolucci, L., Bertini, M., Bagdanov, A.D.: Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. ICLR (2025)
41. Najdenkoska, I., Derakhshani, M.M., Asano, Y.M., Van Noord, N., Worring, M., Snoek, C.G.: Tulip: Token-length upgraded clip. ICLR (2025)
42. Onoe, Y., Rane, S., Berger, Z., Bitton, Y., Cho, J., Garg, R., Ku, A., Parekh, Z., Pont-Tuset, J., Tanzer, G., et al.: Docci: Descriptions of connected and contrasting images. In: ECCV (2024)

43. Pal, A., van Spengler, M., di Melendugno, G.M.D., Flaborea, A., Galasso, F., Mettes, P.: Compositional entailment learning for hyperbolic vision-language models. ICLR (2024)
44. Peng, Z., Xu, Z., Zeng, Z., Wen, C., Huang, Y., Yang, M., Tang, F., Shen, W.: Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In: CVPR (2025)
45. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: CVPR (2015)
46. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
47. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: CVPR (2023)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
49. Ramasinghe, S., Shevchenko, V., Avraham, G., Thalaiyasingam, A.: Accept the modality gap: An exploration in the hyperbolic space. In: CVPR (2024)
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
51. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeuriPS (2022)
52. Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z.: Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In: CVPR (2023)
53. Sala, F., De Sa, C., Gu, A., Ré, C.: Representation tradeoffs for hyperbolic embeddings. In: ICML (2018)
54. Sarkar, S.D., Miksik, O., Pollefeys, M., Barath, D., Armeni, I.: Crossover: 3d scene cross-modal alignment. In: CVPR (2025)
55. Sinha, A., Zeng, S., Yamada, M., Zhao, H.: Learning structured representations with hyperbolic embeddings. NeurIPS (2024)
56. Srivastava, S., Wu, K.: Hypervlm: Hyperbolic space guided vision language modeling for hierarchical multi-modal understanding. In: ICCV (2025)
57. Tifrea, A., Bécigneul, G., Ganea, O.E.: Poincar\'e glove: Hyperbolic word embeddings. arXiv preprint arXiv:1810.06546 (2018)
58. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: ICCV (2019)
59. Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., Romero-Soriano, A.: A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In: CVPR (2024)
60. Wang, B., Ning, Z., Ding, J., Gao, X., Li, Y., Jiang, D., Yang, J., Liu, W.: Fixclip: Dual-branch hierarchical contrastive learning via synthetic captions for better understanding of long text. In: CVPR (2025)
61. Wang, Z., Ramasinghe, S., Xu, C., Monteil, J., Bazzani, L., Ajanthan, T.: Learning visual hierarchies in hyperbolic space for image retrieval. In: CVPR (2025)
62. Wu, K., Peng, H., Zhou, Z., Xiao, B., Liu, M., Yuan, L., Xuan, H., Valenzuela, M., Chen, X.S., Wang, X., et al.: Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In: ICCV (2023)

63. Wu, R., Chen, P., Shen, F., Zhao, S., Hui, Q., Gao, H., Lu, T., Liu, Z., Zhao, F., Wang, K., et al.: Himo-clip: Modeling semantic hierarchy and monotonicity in vision-language alignment. AAAI (2025)
64. Xie, S., Lingjing, L., Zheng, Y., Yao, Y., Tang, Z., Xing, E.P., Chen, G., Zhang, K.: Smartclip: Modular vision-language alignment with identification guarantees. In: CVPR (2025)
65. Yan, S., Liu, Z., Xu, L.: Hyp-uml: Hyperbolic image retrieval with uncertainty-aware metric learning. arXiv preprint arXiv:2310.08390 (2023)
66. Yang, M., Feng, A., Xiong, B., Liu, J., King, I., Ying, R.: Hyperbolic fine-tuning for large language models. NeurIPS (2025)
67. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip. In: ECCV (2024)
68. Zhang, B., Tao, J., Zeng, Z., He, N., Maatouk, A., Yang, M., Ying, R.: Parameter-efficient fine-tuning of llms with mixture of space experts. arXiv preprint arXiv:2602.14490 (2026)
69. Zhao, Y., Jiang, B., Ding, Y., Wang, X., Tang, J., Luo, B.: Fine-grained vlm fine-tuning via latent hierarchical adapter learning. arXiv preprint arXiv:2508.11176 (2025)
70. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: CVPR (2022)
71. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)