HyperET: Efficient Training in Hyperbolic Space for Multi-modal Large Language Models

Zelin Peng[1], **Zhengqin Xu**[2], **Qingyang Liu**[1], **Xiaokang Yang**[1], **Wei Shen**[1] (✉)
[1] MoE Key Lab of Artificial Intelligence, AI Institute, School of Computer Science, SJTU
[2] State Key Laboratory of Infrared Physics, Shanghai Institute of Technical Physics, CAS
{zelin.peng, fate311, narumimaria, xkyang, wei.shen}@sjtu.edu.cn

Abstract

Multi-modal large language models (MLLMs) have emerged as a transformative approach for aligning visual and textual understanding. They typically require extremely high computational resources (e.g., thousands of GPUs) for training to achieve cross-modal alignment at multi-granularity levels. We argue that a key source of this inefficiency lies in the vision encoders they widely equip with, e.g., CLIP and SAM, which lack the alignment with language at multi-granularity levels. To address this issue, in this paper, we leverage hyperbolic space, which inherently models hierarchical levels and thus provides a principled framework for bridging the granularity gap between visual and textual modalities at an arbitrary granularity level. Concretely, we propose an efficient training paradigm for MLLMs, dubbed as HyperET, which can optimize visual representations to align with their textual counterparts at an arbitrary granularity level through dynamic hyperbolic radius adjustment in hyperbolic space. HyperET employs learnable matrices with Möbius multiplication operations, implemented via three effective configurations: diagonal scaling matrices, block-diagonal matrices, and banded matrices, providing a flexible yet efficient parametrization strategy. Comprehensive experiments across multiple MLLM benchmarks demonstrate that HyperET consistently improves both existing pre-training and fine-tuning MLLMs clearly with less than 1% additional parameters. Code is available at https://github.com/godlin-sjtu/HyperET.

1 Introduction

Thanks to advancements in pre-trained foundation models in both computer vision and natural language processing [60, 61, 58, 1, 29, 50, 14, 18, 50, 15], researchers have been inspired to explore the alignment of visual and language models, leading to the development of multi-modal large language models (MLLMs). This alignment is often achieved through adapters, e.g., Q-former [16]. As a result, MLLMs [2, 10, 16, 40, 49, 4, 9, 65] rapidly develop in recent years, demonstrating strong performance in tasks that require both visual and textual understanding, e.g., image captioning and visual question answering (VQA).

Although modern multimodal large language models (MLLMs), e.g., Qwen-VL series [5, 64, 4] and Intern-VL series [13, 12, 11], achieve cross-modal alignment across a wide range of MLLM tasks that inherently involve multi-granularity levels, their success heavily depends on extensive data scaling and massive computational resources. For example, InternVL [13] requires training on hundreds of millions of image-text pairs using up to 640 GPUs. Such a resource-intensive training scheme raises serious concerns about efficiency, reproducibility, and long-term sustainability in the MLLM

✉ Corresponding Author: wei.shen@sjtu.edu.cn

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

community, especially for researchers and institutions constrained by limited computational resources. To address these concerns, it is essential to identify the underlying cause of this inefficiency. We argue that a key factor lies in the vision encoders that are commonly used, e.g., CLIP [54], SAM [29], and DINOv2 [46]. These encoders are typically aligned with language at a single granularity level, e.g., either pixel-level or object-level, and are thus insufficient to deal with tasks required for alignments at different granularity levels. This mismatch in granularity during training significantly impedes the optimization process, leading to inefficient cross-modal alignment and increased reliance on large-scale computational resources.

To solve this issue, we propose to directly quantify the granularity levels by leveraging hyperbolic space [7]. Empirical observations in prior works demonstrate that visual representations at different hierarchical levels (e.g., image-level and object-level) naturally stratified in hyperbolic space [20, 52, 47]. This property enables the use of hyperbolic radius [57]—defined as the distance from a point to the origin in hyperbolic space—to quantify granularity levels [47]. Specifically, points in hyperbolic space closer to the origin (smaller radius) encode low-level visual features (e.g., pixel-level information), while points near the boundary (larger radius) represent high-level visual semantics (e.g., image-level concepts). This hierarchical level suggests that adjusting the hyperbolic radius of visual representations can effectively align them with language models at arbitrary granularity levels.

Building on this insight, we propose an efficient training paradigm (HyperET) for MLLMs that is capable of optimizing visual representations to align with their textual counterparts at arbitrary granularity levels. This is achieved through learnable matrices equipped with Möbius multiplication operations [62], which enable direct and continuous adjustment of the hyperbolic radius of visual representations, thereby facilitating cross-modal alignment. In practice, we introduce three parameter-efficient forms of the learnable matrices for adjusting the hyperbolic radius: (1) diagonal scaling matrices, (2) block-diagonal scaling matrices, and (3) banded scaling matrices. These designs significantly reduce the number of trainable parameters while retaining the capacity to address granularity mismatch. To further enhance parameter flexibility, the matrices can be extended to a dense version with fully populated learnable elements. This expansion increases the expressive capacity of our proposed training paradigm, facilitating more effective alignment across diverse cross-modal scenarios.

To evaluate the generalization capability and effectiveness of HyperET, we conduct extensive experiments across multiple pre-trained MLLM benchmarks and various downstream MLLM tasks, e.g., ScienceQA [42]. Results demonstrate that the proposed HyperET can be easily plugged-and-play and consistently improve various MLLMs, including LLaVA-1.5 [39] and LLaVA-Next [32] for pre-training, as well as MemVP [26] and LaVIN [43] for fine-tuning on downstream tasks. Notably, HyperET introduces less than 1% of the total trainable parameters, ensuring high parameter efficiency.

2 Related Work

2.1 Multi-modal Large Language Models

Multi-modal large language models (MLLMs) [31, 10, 33, 38, 2, 55, 36, 70] make significant breakthroughs in recent advancements, aiming to equip large language models [73, 1, 60, 61, 3] with the capability to process and interpret visual information. Most MLLMs achieve this goal by integrating a CLIP's vision encoder [69, 54] into pre-trained large language models through adapters, e.g., MLPs [40, 39], Q-Former [16, 35], and attention mechanisms [2]. Despite its effectiveness, this straightforward connection between the visual and language modalities still fails to align pre-trained models from different modalities, thereby resulting in inferior performance, e.g., hallucinations, across various downstream tasks.

2.2 Towards Alignment in MLLM

This failure mainly stems from the fact that the CLIP's vision encoder [50] is designed for standard classification tasks and is not equipped to handle the more fine-grained visual understanding tasks required by the language modality [59, 51, 27]. To better align the vision and language modalities in MLLMs, recent works explore parameter-efficient fine-tuning methods [43, 56, 71, 26, 44]. For example, LaVIN [43] introduces adapters in both the vision encoder and LLaMA [60] to achieve better alignment of the modalities. Another line of research [66, 59, 27] seeks to overcome this

bottleneck by incorporating additional vision models, such as DINOv2 [46], to construct a more powerful and capable vision branch. Despite the advancements in the field, few studies focus on understanding the changes in the representation of vision encoders that enable MLLMs to tackle more complex vision tasks. In this work, we aim to bridge this gap by leveraging hyperbolic space to directly model the granularity levels and adapt the granularity of visual representations to an appropriate level.

2.3 Learning in Hyperbolic Space

Unlike Euclidean space, hyperbolic space can be viewed as the continuous analog of a tree [6], making it inherently suitable for capturing hierarchical levels among various data types. Since visual and textual concepts are inherently hierarchical, recent works [17, 30, 52, 47, 48] show that hyperbolic space serves as a promising manifold for preserving granularity levels in vision-language model representations, leading to strong performance across downstream tasks. Most existing methods directly reshape the original hierarchical structure to align different modalities. In contrast, our method preserves the original hierarchical structure and specifically leverages the intrinsic property of hyperbolic radius to enable visual representations to adapt their hierarchical level, aligning them with the language modality. This approach provides a complementary perspective to existing efforts in the MLLM community.

3 Preliminary Concepts

Hyperbolic Geometry. In contrast to Euclidean or spherical geometries, hyperbolic geometry is characterized by a constant negative curvature, which fundamentally distinguishes its geometric properties and computational behaviors. Following prior studies [8, 57], we utilize the classical Poincaré ball model—one of five principal analytic models for constructing hyperbolic space [7]—due to its demonstrated efficacy in representing hierarchical levels [23, 19, 45]. The Poincaré ball model $(\mathbb{D}_c^n, g^{\mathbb{D}_c})$, characterized by a radius of $1/\sqrt{c}$ and constant negative curvature $-c$ $(c > 0)$ is formally defined as follows:

$$\begin{cases} \mathbb{D}_c^n := \{\mathbf{X} \in \mathbb{R}^n : c\|\mathbf{X}\| < 1\} \\ g^{\mathbb{D}_c} := \lambda_{c,\mathbf{X}}^2 g^E \end{cases}, \tag{1}$$

where $\lambda_{c,\mathbf{X}} = \frac{2}{1-c\|\mathbf{X}\|^2}$ and $g^E = \mathbf{I}_n$ denotes the Euclidean metric tensor, serving as the foundation for the hyperbolic space construction.

Hyperbolicity. Building upon the theoretical framework of gyrovector spaces [62, 63], we incorporate Möbius operations into hyperbolic space, specifically the Möbius addition operation "$\oplus_c$" and Möbius multiplication operation "$\otimes_c$". In hyperbolic geometry, the tangent space $\mathcal{T}_{\mathbf{X}}^c\mathbb{D}_c^n$ at any point $\mathbf{X} \in \mathbb{D}_c^n$ serves as a first-order approximation of $\mathbb{D}_c^n$, representing an n-dimensional Euclidean space that locally approximates the hyperbolic structure. The tangent space $\mathcal{T}_{\mathbf{X}}^c\mathbb{D}_c^n$ and $\mathbb{D}_c^n$ are mapped to each other by exponential ($\mathcal{T}_{\mathbf{X}}^c\mathbb{D}_c^n \mapsto \mathbb{D}_c^n : \exp_{\mathbf{X}}^{\mathbb{D},c}(\cdot)$) and logarithmic ($\mathbb{D}_c^n \mapsto \mathcal{T}_{\mathbf{X}}^c\mathbb{D}_c^n : \log_{\mathbf{X}}^{\mathbb{D},c}(\cdot)$) maps, respectively. Detailed mathematical definitions and derivations are provided in the supplementary material.

4 Methodology

This section first provides the necessary background on the conventional training paradigm from the perspective of parameter space tuning (Sec. 4.1) to contextualize our approach, followed by the detailed presentation of HyperET in Sec. 4.2. Then, Sec. 4.3 provides a theoretical analysis on adjusting the hyperbolic radius of visual representations.

4.1 Parameter Space Tuning

Parameter space tuning in multi-modal large language models (MLLMs), whether through full fine-tuning or parameter-efficient fine-tuning, aims to adapt pre-trained visual and language models to target multi-modal scenarios. However, these tuning methods, which rely solely on gradient updates, operate as constraint-free adjustments in Euclidean space. They often implicitly assume that visual representations can sufficiently adapt to the required granularity level (e.g., transitioning

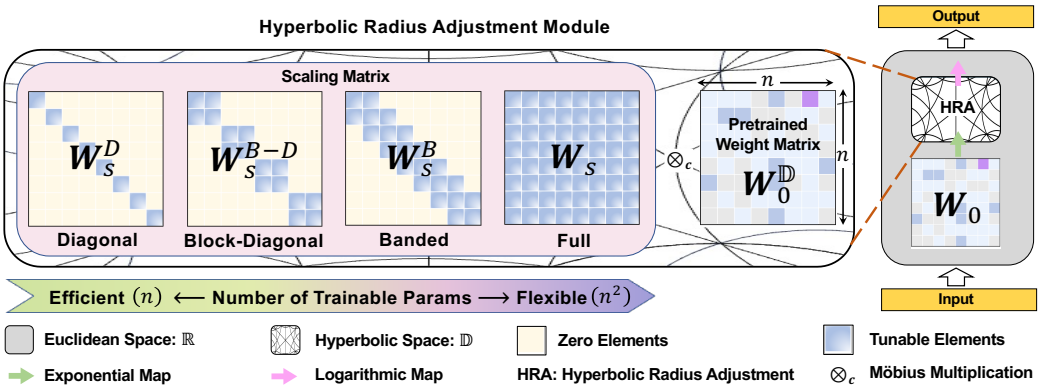

Figure 1: **The schematic representation of HyperET.** In HyperET, we efficiently train MLLMs in hyperbolic space by adjusting the hyperbolic radius using a tunable scaling matrix $\mathbf{W}_s$. Here, $\mathbf{W}_s$ can be configured into three parameter-efficient variants, i.e., Diagonal, Block-Diagonal and Banded.

from image-level to pixel-level), which may lead to inefficiency in alignment at a certain granularity level. In contrast, the core technique of our proposed HyperET involves hyperbolic radius adjustment, which explicitly adjusts the granularity level of visual representations in MLLMs. HyperET provides a simple yet effective solution to granularity mismatch challenges.

4.2 General Hyperbolic Radius Adjustment

As previously discussed, hyperbolic radius adjustment provides a direct mechanism for optimizing the granularity level of visual representations, effectively bridging the granularity gap between visual and language modalities. In practice, we introduce a radius adjustment constraint into the weight update process, generally defined as follows:

$$\mathrm{Rad}_{\mathbf{W}^{\mathbb{D}}}/\mathrm{Rad}_{\mathbf{W}_0^{\mathbb{D}}} = s \quad \Leftrightarrow \quad \mathrm{Rad}_{\mathbf{W}^{\mathbb{D}}} = s \cdot \mathrm{Rad}_{\mathbf{W}_0^{\mathbb{D}}}, \tag{2}$$

where $\mathrm{Rad}_{\mathbf{W}^{\mathbb{D}}}$ and $\mathrm{Rad}_{\mathbf{W}_0^{\mathbb{D}}}$ represent the hyperbolic radii of $\mathbf{W}^{\mathbb{D}}$ and $\mathbf{W}_0^{\mathbb{D}}$, respectively, quantifying their granularity levels. The hyperbolic weight matrices $\mathbf{W}^{\mathbb{D}}$ and $\mathbf{W}_0^{\mathbb{D}}$ are derived by projecting their Euclidean counterparts $\mathbf{W}$ and $\mathbf{W}_0$ into hyperbolic space through exponential mapping operations defined in Sec. 3. The scaling coefficient s is task-adaptive, dynamically adjusting to optimize performance. However, this modification requires a two-step procedure: (1) computing the hyperbolic radius of $\mathbf{W}^{\mathbb{D}}$ and (2) subsequently adjusting it. To streamline this process, we propose a more efficient approach that directly optimizes $\mathbf{W}^{\mathbb{D}}$ without intermediate radius computations.

Hyperbolic Radius. Without loss of generality, for any point $\mathbf{X} \in \mathbb{D}_c^n$ in hyperbolic space, its hyperbolic radius is formally defined as follows:

$$\mathrm{Rad}_{\mathbf{X}} := d_c^{\mathbb{D}}(\mathbf{X}, \mathbf{0}) = (\frac{2}{\sqrt{c}})\tanh^{-1}(\sqrt{c}\|\mathbf{X}\|), \tag{3}$$

where $\mathbf{0}$ denotes the origin point in hyperbolic space. In the subsequent analysis, we demonstrate that Mobiüs multiplication operations $\otimes_c$ defined in Sec. 3 can enable precise control over the hyperbolic radius. This critical property is formally stated in the following theorems:

Theorem 1 (Hyperbolic Radius Scaling) For a point $\mathbf{X} \in \mathbb{D}_c^n$ in hyperbolic space, the hyperbolic radius adjustment function is expanded as follows:

$$\begin{aligned} s \cdot \mathrm{Rad}_{\mathbf{X}} &= \frac{2}{\sqrt{c}}(s\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}}) \\ &= \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\|s \otimes_c \mathbf{X}\|) \\ &= \mathrm{Rad}_{s\otimes_c \mathbf{X}}. \end{aligned} \tag{4}$$

where $\otimes_c$ here is instantiated as a Mobiüs scalar multiplication operation. Therefore, hyperbolic radius adjustment can be precisely controlled through $\otimes_c$ between s and $\mathbf{X}$, with the scaling coefficient s serving as a primary learnable parameter. ■

According to Theorem 1, $\otimes_c$ can provide an equivalent mechanism during parameter space tuning. Consequently, hyperbolic radius adjustment in Eq. (2) can be easily achieved as follows:

$$\mathbf{W}^{\mathbb{D}} = s \otimes_c \mathbf{W}_0^{\mathbb{D}}. \tag{5}$$

Then, building upon the exponential mapping $\exp_{\mathbf{0}}^{\mathbb{D},c}(\cdot)$ and the logarithmic mapping $\log_{\mathbf{0}}^{\mathbb{D},c}(\cdot)$ defined in Sec. 3, we achieve general hyperbolic radius adjustment via the following reformulation of Eq. (5):

$$\mathbf{W} = \log_{\mathbf{0}}^{\mathbb{D},c}(s \otimes_c \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{W}_0)), \tag{6}$$

where s is a learnable parameter for adjustment. Given that a constrained parameter set—particularly when limited to a single learnable parameter—may ineffectively adjust hyperbolic radius during restricted training iterations, we propose a more flexible parameterization strategy through matrix-based formulations, termed flexible adjustment, which further enables precise control over hyperbolic radius optimization.

Flexible Adjustment for Hyperbolic Radius. To achieve this, we adopt a scaling matrix $\mathbf{W}_s$ to replace the scaling coefficient s, which is satisfied: $s = \|\mathbf{W}_s\mathbf{W}_0^{\mathbb{D}}\|/\|\mathbf{W}_0^{\mathbb{D}}\|$. Consequently, the right-hand side of Eq. (2) can be reformulated as follows:

$$\begin{aligned} \text{Rad}_{\mathbf{W}^{\mathbb{D}}} &= s \cdot \text{Rad}_{\mathbf{W}_0^{\mathbb{D}}} \\ &= \frac{\|\mathbf{W}_s\mathbf{W}_0^{\mathbb{D}}\|}{\|\mathbf{W}_0^{\mathbb{D}}\|} \cdot \text{Rad}_{\mathbf{W}_0^{\mathbb{D}}}. \end{aligned} \tag{7}$$

In this framework, the learnable parameters transition from scalar values to matrix-based formulations, significantly enhancing flexibility within constrained training iterations. The following theorem demonstrates that hyperbolic radius can be dynamically scaled through Mobiüs matrix multiplication operations, a generalized instantiation of $\otimes_c$.

Theorem 2 (Hyperbolic Radius Flexibility Scaling) For a point $\mathbf{X} \in \mathbb{D}_c^n$ in hyperbolic space and a scaling matrix $\mathbf{X}_s \in \mathbb{R}^{n\times n}$, the flexible radius adjustment function is formally defined through Eq. (7) and Mobiüs matrix multiplication operations as follows:

$$\begin{aligned} \frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|} \cdot \text{Rad}_{\mathbf{X}} &= \frac{2}{\sqrt{c}}\left(\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|}\frac{\sqrt{c}}{2}\text{Rad}_{\mathbf{X}}\right) \\ &= \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\|\mathbf{X}_s \otimes_c \mathbf{X}\|) \\ &= \text{Rad}_{\mathbf{X}_s \otimes_c \mathbf{X}}. \end{aligned} \tag{8}$$

Analogically, the scaling matrix $\mathbf{W}_s$ is able to direct adjust the hyperbolic radius $\text{Rad}_{\mathbf{X}}$ through Mobiüs matrix multiplication operations, providing precise control over representation granularity. ■

Building upon Theorem 2, we introduce a matrix-based formulation by replacing the scaling coefficient s with $\mathbf{W}_s$, resulting in the following reformulation of Eq. (6):

$$\mathbf{W} = \log_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{W}_s \otimes_c \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{W}_0)). \tag{9}$$

With up to $O(n^2)$ learnable elements, $\mathbf{W}_s$ in Eq. (9) offers significantly enhanced flexibility for hyperbolic radius adjustment. Notably, Eqs. (5) and (9) become equivalent when $\mathbf{W}_s$ is constrained to a diagonal matrix with uniform scaling factors, i.e., $\mathbf{W}_s = \text{diag}(\omega_1, \omega_2, \cdots, \omega_n) \in \mathbb{R}^{n\times n}$ and $\omega_i = \omega_j, i \neq j$. Moreover, in alignment with the current paradigm of parameter-efficient fine-tuning, we also introduce a parameter-efficient variant of $\mathbf{W}_s$, designed to maintain flexibility while reducing computational overhead.

Efficient Adjustment for Hyperbolic Radius. To realize this strategy, we define a *diagonal scaling matrix* $\mathbf{W}_s^D = \text{diag}(\omega_1, \omega_2, \ldots, \omega_n) \in \mathbb{R}^{n\times n}$, where each ω_i is a learnable scalar and $\omega_i \neq \omega_j$ when $i \neq j$, significantly reducing the number of learnable parameters. In this form, only the diagonal elements are learnable parameters, making the diagonal scaling matrix $\mathbf{W}_s^D$ the most parameter-efficient configuration for hyperbolic radius adjustment. Building upon this foundation, we extend $\mathbf{W}_s^D$ into two more flexible variants by incrementally introducing learnable off-diagonal elements, balancing enhanced adjustment capability with parameter efficiency: (1) *Block-diagonal scaling matrix* $\mathbf{W}_s^{B\text{-}D} = \text{diag}(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_r)$, where $\mathbf{R}_i \in \mathbb{R}^{\frac{n}{r}\times\frac{n}{r}}$. Here, r is the block size (we assume

n is divisible by r). This formulation allows intra-block interactions while maintaining sparsity across blocks; (2) *Banded scaling matrix* $\mathbf{W}_s^B = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{n1} & \cdots & \omega_{nn} \end{bmatrix} \in \mathbb{R}^{n\times n}$, where $\omega_{ij} = 0$ for all i, j such that $|i-j| > d$. Here, d denotes the bandwidth, indicating that nonzero elements are allowed within d entries above and below the main diagonal. These two variants provide mechanisms to capture localized interactions while preserving parameter efficiency. Notably, both $\mathbf{W}_s^{B\text{-}D}$ and $\mathbf{W}_s^B$ degenerate to $\mathbf{W}_s^D$ under specific configurations: when $r = n$ for the block-diagonal matrix and $d = 0$ for the banded matrix. This highlights the inherent hierarchical flexibility enabled by our parametrization strategy and reflects a insightful way of designing fine-tuning methods.

To summarize these fine-tuning strategies, Fig. 1 provides a unified illustration of different variants of the scaling matrix $\mathbf{W}_s$. Our introduced hyperbolic radius adjustment, i.e., HRA, adjusts visual representations by applying learnable scaling matrices to the frozen pre-trained weights $\mathbf{W}_0$ in hyperbolic space, enabling precise adjustment over arbitrary granularity levels.

4.3 Theoretical Analysis

To demonstrate the impact of hyperbolic radius adjustment, we provide a theoretical analysis of the Möbius multiplication operation. The following deduction shows that the proposed Möbius multiplication operation can directly adjust the granularity level of visual representations.

Deduction. For $\mathbf{Y}_0 = \mathbf{W}_0\mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^{d\times k}$ is the input embedding, $\mathbf{W}_0$ is the pre-trained weight and $\mathbf{Y}_0$ is the pre-trained visual representation, the forward pass of HyperET is as follows:

$$\mathbf{Y}_0 = \mathbf{W}\mathbf{X} = \log_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{W}_s \otimes_c \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{W}_0))\mathbf{X}. \tag{10}$$

Then, according to the definition of hyperbolic space, $\mathbf{Y}_0$ is firstly projected into the hyperbolic space $\mathbb{D}_c^n$ and our HyperET is applied to adjust the hyperbolic radius, resulting in the new visual representation $\mathbf{Y}$, expressed as:

$$\mathbf{Y}_0^{\mathbb{D}_c^n} = \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{Y}_0) = \mathbf{W}_0^{\mathbb{D}_c^n}\mathbf{X}^{\mathbb{D}_c^n} \tag{11}$$

$$\mathbf{Y}^{\mathbb{D}_c^n} = \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{Y}) = \mathbf{W}_s \otimes_c \mathbf{W}_0^{\mathbb{D}_c^n}\mathbf{X}^{\mathbb{D}_c^n}, \tag{12}$$

where $\mathbf{W}_0^{\mathbb{D}_c^n} = \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{W}_0)$ and $\mathbf{X}^{\mathbb{D}_c^n} = \exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{X})$. According to the Theorem 2, the hyperbolic radius of $\mathbf{Y}^{\mathbb{D}_c^n}$ is obtained as:

$$\begin{aligned} \mathrm{Rad}_{\mathbf{Y}} &= \mathrm{Rad}_{\mathbf{W}_s\otimes_c\mathbf{W}_0^{\mathbb{D}_c^n}\mathbf{X}^{\mathbb{D}_c^n}} \\ &= \frac{\|\mathbf{W}_s\mathbf{W}_0^{\mathbb{D}_c^n}\mathbf{X}^{\mathbb{D}_c^n}\|}{\|\mathbf{W}_0^{\mathbb{D}_c^n}\mathbf{X}^{\mathbb{D}_c^n}\|} \cdot \mathrm{Rad}_{\mathbf{W}_0^{\mathbb{D}_c^n}\mathbf{X}^{\mathbb{D}_c^n}} \\ &= \frac{\|\mathbf{W}_s\mathbf{W}_0^{\mathbb{D}_c^n}\mathbf{X}^{\mathbb{D}_c^n}\|}{\|\mathbf{W}_0^{\mathbb{D}_c^n}\mathbf{X}^{\mathbb{D}_c^n}\|}\mathrm{Rad}_{\mathbf{Y}_0} \\ &= s \cdot \mathrm{Rad}_{\mathbf{Y}_0}, \end{aligned} \tag{13}$$

where s is a scaling coefficient. Therefore, our hyperbolic radius adjustment method can directly adjust the hyperbolic radius of visual representations, thus is able to bridge the granularity gap between visual and textual modalities at an arbitrary granularity level.

5 Experiments

Our experimental evaluation encompasses two MLLM scenarios, (1) MLLM's fine-tuning (Sec. 5.1), and (2) MLLM's pre-training (Sec. 5.2). A detailed ablation study is presented in Sec. 5.3 and 5.4.

5.1 MLLM's Fine-tuning

Experimental Setting. We evaluate our method on ScienceQA [42], a challenging large-scale VQA benchmark encompassing diverse scientific domains. Our comparative analysis includes MemVP and other LLaMA-based models with input-space visual prompting: LLaVA [40], and LaVIN [43]. We

Table 1: **Comparision with SoTA fine-tuning methods** on ScienceQA test set [42]. Question categories: NAT = natural science, SOC = social science, LAN = language science, TXT = w/ text context, IMG = w/ image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. "Ours": we here realize the extra learnable parameters as diagonal matrices, i.e., $\mathbf{W}_s^D$. Vision encoder: CLIP.

Method	#Trainable Params	Language Model	Subject			Context Modality			Grade		Average
			NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Human	-	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
Fully Fine-Tuning											
LLaVA	13B	Vicuna-13B	90.36	**95.95**	88.00	89.49	88.00	90.66	90.93	**90.90**	90.92
Parameter-efficient Fine-Tuning											
LaVIN	3.8M	LLaMA-7B	89.25	94.94	85.24	88.51	87.46	88.08	90.16	88.07	89.41
LaVIN+Ours	3.85M (+0.05M)	LLaMA-7B	**89.35**	**96.06**	**86.54**	88.29	**88.01**	**89.33**	**91.36**	87.65	**90.03** (+0.62)
MemVP	3.9M	LLaMA-7B	94.45	95.05	88.64	93.99	92.36	90.94	93.10	93.01	93.07
MemVP+Ours	3.95M (+0.05M)	LLaMA-7B	**94.85**	**95.05**	**90.55**	**94.57**	**92.91**	**92.20**	**93.65**	**94.00**	**93.78** (+0.71)
LaVIN	5.4M	LLaMA-13B	90.32	94.38	87.73	89.44	87.65	90.31	91.19	89.26	90.50
LaVIN+Ours	5.45M (+0.05M)	LLaMA-13B	**90.57**	**95.63**	**89.89**	**89.61**	**88.75**	**92.02**	**91.95**	**90.58**	**91.46** (+0.96)
MemVP	5.5M	LLaMA-13B	95.07	95.15	90.00	94.43	92.86	92.47	93.61	94.07	93.78
MemVP+Ours	5.55M (+0.05M)	LLaMA-13B	**96.19**	**95.78**	**90.86**	**95.51**	**94.25**	**93.18**	**94.88**	**94.44**	**94.72** (+0.94)

follow the experiment setting in [43]. All models utilize a CLIP pre-trained ViT-L/14 visual encoder. The weights of HyperET in this task are implemented using the three parameter-efficient scaling matrices, i.e., $\mathbf{W}_s^D$, $\mathbf{W}_s^{B-D}$ and $\mathbf{W}_s^B$, and are adapted in the attention layer, consistent with most parameter-efficient tuning methods, e.g., LoRA [24]. The curvature c is 0.01. All experiments are conducted using a maximum of 8 NVIDIA H800 GPUs.

Comparing to SOTA. Our experimental evaluation compares the proposed approach with state-of-the-art parameter-efficient fine-tuning (PEFT) methods, including LaVIN [43] and MemVP [26]. As demonstrated in Table 4, which presents both baseline and HyperET enhanced results, our method establishes new state-of-the-art performance. HyperET achieves this breakthrough with minimal parameter overhead (fewer than 1%), delivering substantial improvements to both LaVIN and MemVP frameworks. Particularly noteworthy are the gains observed with LLaMA-13B as the backbone language model, where HyperET enhances average cross-domain accuracy by 0.96% for LaVIN and 0.94% for MemVP. Notably, when integrated with MemVP using LLaMA-7B, HyperET achieves performance on par with the more computationally intensive MemVP-LLaMA-13B configuration, i.e., 93.78. This result demonstrates that HyperET enhances visual representation and achieves comparable performance improvements while utilizing 100, 000 $\times$ fewer parameters (0.05M vs 6B) than MemVP using LLaMA-13B, highlighting its remarkable parameter efficiency.

5.2 MLLM's Pre-training

Experimental Setting. Our pre-training evaluation framework builds upon LLaVA-1.5 [39], employing identical datasets to evaluate HyperET's effectiveness in MLLM pre-training. Our comparative analysis includes LLaVA-1.5 [39] and LLaVA-Next [39] and train and fine-tune our HyperET with the same experiment setting. The weights of HyperET in this task are implemented using the highest flexible scaling matrices, i.e., $\mathbf{W}_s$, and are adapted in the attention layer, consistent with most parameter-efficient training methods, e.g., LoRA [24].

Comparing to SOTA. Our experimental framework evaluates the proposed method against state-of-the-art pre-trained MLLMs across 12 standard visual language benchmarks. We implement HyperET using the most flexible matrix configuration, i.e., $\mathbf{W}_s$, maximizing the model's adaptability. While this approach resembles full fine-tuning in structure, the actual parameter count remains remarkably efficient at approximately 50M—less than 1% of the language model's 13B parameters—maintaining parameter efficiency. As shown in Table 2, our approach demonstrates significant performance improvements over LLaVA-1.5 [39], particularly in mitigating the limitations of CLIP-based encoders. Notably, on the POPE benchmark [37] for object hallucination detection, HyperET substantially reduces visual hallucinations, providing empirical evidence that hyperbolic radius optimization effectively enhances the cross-modal alignment at an arbitrary level.

Table 2: **Comparison with SoTA pre-trained methods** on 12 MLLM benchmarks, including VQAv2 [21], GQA [25], VW: VisWiZ [22], SQA: ScienceQA-IMG [42], TVQA: TextVQA [53], PE: POPE [37], ME: MME [67], MB: MMBench [41], MB^{CN}: MMBench-Chinese [41], SD: SEED-Bench [34], LVA^{W}: LLaVA-Bench (In-the-Wild) [40] and M-Vet [68]. Top-1 accuracy is reported (Best in **bold**, second best is underlined). Lan. Model: Language model. Benchmark names are abbreviated due to space limits. "Ours": we here realize the extra learnable parameters as full matrices, i.e., $\mathbf{W}_s$. Vision encoder: CLIP.

Method	Lan. Model	VQAv2	GQA	VW	SQA	TVQA	PE	ME	MB	MB^{CN}	SD	LVA^{W}	M-Vet
LLaVA-1.5	Vicuna-7B	78.5	62.0	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5+Ours	Vicuna-7B	**80.3**	**63.7**	**51.9**	**69.1**	**60.8**	**87.7**	**1536.2**	**66.8**	**60.5**	**60.2**	**65.6**	**32.4**
LLaVA-1.5	Vicuna-13B	80.0	63.3	53.6	71.6	61.3	85.9	1531.3	67.7	63.6	61.6	70.7	35.4
LLaVA-1.5+Ours	Vicuna-13B	**82.3**	**65.7**	**55.2**	**73.7**	**63.9**	**88.7**	**1584.7**	**69.8**	**65.2**	**63.4**	**72.6**	**38.3**
LLaVA-Next	Vicuna-7B	81.8	64.2	57.6	70.1	64.9	86.5	1519	67.4	60.6	70.2	81.6	43.9
LLaVA-Next+Ours	Vicuna-7B	**82.9**	**65.4**	**58.9**	**70.8**	**65.1**	**88.9**	**1551**	**69.9**	**62.5**	**71.0**	**82.9**	**44.8**

Table 3: **Comparative analysis of fine-tuning spaces and flexibility levels** on ScienceQA test set [42]. All experiments utilize MemVP [26] with LLaMA-13B as the backbone language model. The notation is defined as follows: $\mathbf{W}_s^D$ represents diagonal scaling matrices, $\mathbf{W}_s^{B-D}$ denotes block-diagonal scaling matrices. $\mathbf{W}_s^B$ indicates banded scaling matrices, and $\mathbf{W}_{se}^*$ corresponds to Euclidean space fine-tuning matrices. Key parameters include d for banded size and $\frac{n}{r}$ for block size. $\otimes_c$: Möbius matrix multiplication.

Method	#Trainable Params (M)	d	$\frac{n}{r}$	$\otimes_c$	Average
MemVP	5.5	-	-	-	93.78
Efficient training					
$+\mathbf{W}_{se}^D$	5.55 (+0.05)	0	1	-	93.81 (+0.03)
$+\mathbf{W}_{se}^{B-D}$	5.64 (+0.14)	-	2	-	93.70 (−0.08)
$+\mathbf{W}_{se}^B$	5.71 (+0.21)	1	-	-	93.65 (−0.13)
Efficient training in hyperbolic space					
$+\mathbf{W}_s^D$	5.55 (+0.05)	0	1	✗	93.91
$+\mathbf{W}_s^D$	5.55 (+0.05)	0	1	✓	94.72 (+0.94)
$+\mathbf{W}_s^{B-D}$	5.64 (+0.14)	-	2	✓	94.79 (+1.01)
	5.78 (+0.28)	-	4	✓	94.84
	6.08 (+0.58)	-	8	✓	94.82
$+\mathbf{W}_s^B$	5.71 (+0.21)	1	-	✓	**94.89** (+1.11)
	5.86 (+0.36)	2	-	✓	94.82
	6.15 (+0.65)	4	-	✓	94.83

Table 4: **Ablation studies of HyperET across vision encoders with varying granularity levels** on ScienceQA test set.

Method	Lang. Model	Vision Encoder	Average
MemVP	LLaMA-13B	DINOV2	91.47
MemVP	LLaMA-13B	SAM	91.16
Efficient training			
$+\mathbf{W}_{se}^D$	LLaMA-13B	DINOV2	91.98 (+0.51)
$+\mathbf{W}_{se}^D$	LLaMA-13B	SAM	92.05 (+0.89)
Efficient training in hyperbolic space			
$+\mathbf{W}_s^D$	LLaMA-13B	DINOV2	93.38 (+1.91)
$+\mathbf{W}_s^D$	LLaMA-13B	SAM	93.74 (+2.58)

Table 5: **Ablation study on the key components of HyperET** on selected five MLLM benchmarks. We here realize the extra learnable parameters as full matrices, i.e., $\mathbf{W}_s$. $\otimes_c$: Möbius matrix multiplication. $\mathbf{W}_{se}$ corresponds to Euclidean space fine-tuning matrices with the same number of parameters.

Method	VQAv2	GQA	VW	SQA	TVQA
Baseline	80.0	63.3	53.6	71.6	61.3
Efficient training					
$+\mathbf{W}_{se}$	80.8	63.8	53.8	71.7	61.8
Efficient training in hyperbolic Space					
$+\mathbf{W}_s$	**82.3**	**65.7**	**55.2**	**73.7**	**63.9**
$-\otimes_c$	81.1	64.0	53.9	71.9	62.1

5.3 Ablation Study on Fine-tuning

Here, we do an ablation study on the ScienceQA [42], employing MemVP [26] as the backbone.

Introducing Different Flexibility into HyperET. We systematically investigate our proposed three distinct parameterization strategies for HyperET 's learnable components. These strategies include diagonal scaling matrices $\mathbf{W}_s^D$ and their two extended variants: banded scaling matrices $\mathbf{W}_s^B$ and block-diagonal scaling matrices $\mathbf{W}_s^{B-D}$, which systematically increase parameter flexibility through progressively relaxed structural constraints. As evidenced in Table 3, enhanced flexibility—achieved through $\mathbf{W}_s^B$ fine-tuning—provides only marginal performance improvements (0.17% accuracy gain over $\mathbf{W}_s^D$) while introducing over-fitting risks when further increasing learnable parameters via expanded banded sizes. These findings demonstrate that our fine-tuning method effectively accomplishes two key objectives: (1) efficient adaptation to optimal hyperbolic radii and (2) robust cross-modal alignment for downstream tasks, all while maintaining parameter efficiency.

Effectiveness of Training in Hyperbolic Space. To isolate the performance gains attributable to hyperbolic space rather than parameter increases, we conduct controlled experiments by adding an equivalent number of parameters through matrix multiplication in Euclidean space, as shown in Table 3. The results demonstrate that such parameter augmentation either yields no significant performance improvement or even degrades model performance. This empirical evidence strongly validates the necessity of employing hyperbolic space for adjusting vision encoder in MLLMs, as the observed improvements in Table 3 cannot be explained by mere parameter increases.

Necessity of Mobiüs matrix multiplication. We conduct ablation studies to evaluate the impact of the Möbius matrix multiplication operation. As shown in Table 3, a comparison between rows 6 and 7 reveals that standard matrix multiplication (denoted by" ✗") underperforms 0.81% compared to the Möbius matrix multiplication, (represented by "✓"). This performance gap underscores the essential role of Möbius operations in precisely adjusting the hyperbolic radius of visual representations.

Discussion of Different Vision Encoder. To further clearly demonstrate the primary contributing factor behind the performance improvements, we present additional experimental results designed to isolate and eliminate the influence of merely increasing the number of trainable parameters. As shown in Table 4, fine-tuning SAM [29] and DINOv2 [46] in Euclidean space with the same number of additional parameters as HyperET yields only marginal gains. In contrast, fine-tuning with HyperET results in substantial performance improvements. These findings clearly illustrate that our main contribution arises specifically from the proposed hyperbolic radius adjustment mechanism, rather than simply from introducing more learnable parameters.

5.4 Ablation Study on Pre-training

To further assess the contribution of each key component in the proposed HyperET, we isolate each component in separate experiments, and the results are presented in Table 5. Comparing row 2 and row 3, training in Euclidean space instead of hyperbolic space leads to significantly lower performance, indicating that the granularity gap mismatch cannot be effectively addressed in Euclidean space. The observed improvement over the baseline is primarily attributable to the increased number of parameters rather than the alignment of granularity levels. Intriguingly, comparing row 2 and row 4, simply transitioning from Euclidean space to hyperbolic space yields a slight improvement, which we attribute to the inherent ability of hyperbolic space to capture granularity levels. However, this strategy lacks explicit mechanisms for hyperbolic radius adjustment, limiting its effectiveness in addressing granularity mismatches. Finally, by adjusting the hyperbolic radius through Möbius matrix multiplication and integrating it with learnable matrices $\mathbf{W}_s$, the results show significant performance improvements, e.g., a 2.1% gain on the TextVQA [53], demonstrating the effectiveness of HyperET. Additionally, we visualize the changes in the hyperbolic radius in Fig. 2 after training, and the observed change in the hyperbolic radius indicates the MLLM's requirement for multi-granularity levels, demonstrating the necessity of our proposed HyperET.

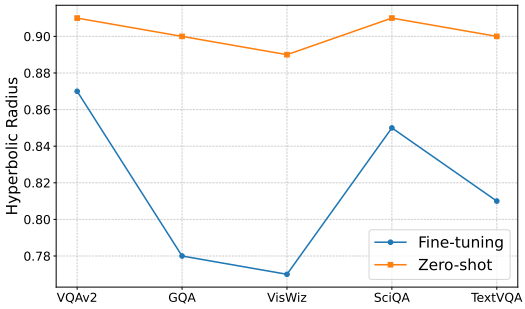

Figure 2: **Visualization of hyperbolic radius changes in visual representation after training** across different MLLM benchmarks. Normalizing the hyperbolic radius to a range of 0–1 facilitates comparison. A smaller hyperbolic radius corresponds to a more low granularity level of visual representation. "Zero-shot": maintaining the pre-trained weights of the vision encoder, i.e., CLIP, without additional training.

6 Conclusion

This work proposes an efficient training paradigm (HyperET) for multi-modal large language models in hyperbolic space. By dynamically adjusting the hyperbolic radius of visual representations through learnable matrices and Möbius multiplication operations, HyperET effectively bridges the granularity gap between visual and textual modalities at an arbitrary granularity level. Our experiments across multiple MLLM benchmarks demonstrate that HyperET consistently improve existing pre-training and fine-tuning baselines by large margins with less than 1% additional parameters.

Acknowledgment. This work was supported by the NSFC under Grant 62322604, 62176159, and in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[6] Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

[7] James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic geometry. *Flavors of geometry*, 31(59-115):2, 1997.

[8] Jiaxin Chen, Jie Qin, Yuming Shen, Li Liu, Fan Zhu, and Ling Shao. Learning attentive and hierarchical representations for 3d shape recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 105–122. Springer, 2020.

[9] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[10] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[15] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

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

[18] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[19] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

[20] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6840–6849, 2023.

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[22] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[23] Chengyang Hu, Ke-Yue Zhang, Taiping Yao, Shouhong Ding, and Lizhuang Ma. Rethinking generalizable face anti-spoofing via hierarchical prototype-guided distribution refinement in hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1032–1041, 2024.

[24] 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. *arXiv preprint arXiv:2106.09685*, 2021.

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[26] Shibo Jie, Yehui Tang, Ning Ding, Zhi-Hong Deng, Kai Han, and Yunhe Wang. Memory-space visual prompting for efficient vision-language fine-tuning. *arXiv preprint arXiv:2405.05615*, 2024.

[27] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pages 113–132. Springer, 2024.

[28] 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*, pages 6418–6428, 2020.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[30] Fanjie Kong, Yanbei Chen, Jiarui Cai, and Davide Modolo. Hyperbolic learning with synthetic captions for open-world detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16762–16771, 2024.

[31] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[32] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild. 2024.

[33] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

[34] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[36] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[37] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[38] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024.

[39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023.

[41] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[42] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[43] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36, 2023.

[44] Feipeng Ma, Hongwei Xue, Yizhou Zhou, Guangting Wang, Fengyun Rao, Shilin Yan, Yueyi Zhang, Siying Wu, Mike Zheng Shou, and Xiaoyan Sun. Visual perception by large language model's weights. *arXiv preprint arXiv:2405.20339*, 2024.

[45] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018.

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[47] 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. *arXiv preprint arXiv:2410.06912*, 2024.

[48] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Changsong Wen, Yu Huang, Menglin Yang, Feilong Tang, and Wei Shen. Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4562–4572, 2025.

[49] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[51] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34, 2024.

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

[53] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[54] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[55] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[56] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237, 2022.

[57] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12617, 2021.

[58] Yehui Tang, Fangcheng Liu, Yunsheng Ni, Yuchuan Tian, Zheyuan Bai, Yi-Qi Hu, Sichao Liu, Shangling Jui, Kai Han, and Yunhe Wang. Rethinking optimization and architecture for tiny language models. *arXiv preprint arXiv:2402.02791*, 2024.

[59] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[61] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[62] Abraham A Ungar. Hyperbolic trigonometry and its application in the poincaré ball model of hyperbolic geometry. *Computers & Mathematics with Applications*, 41(1-2):135–147, 2001.

[63] Abraham Albert Ungar. A gyrovector space approach to hyperbolic geometry. *Synthesis Lectures on Mathematics and Statistics*, 1(1):1–194, 2008.

[64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[65] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2023.

[66] Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. Diffusion feedback helps clip see better. *arXiv preprint arXiv:2407.20171*, 2024.

[67] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[68] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

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

[70] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[71] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[72] Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10421–10429, 2025.

[73] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract clearly states the paper's main contributions and scope. We propose an efficient training paradigm for MLLMs, dubbed as HyperET, which can optimize visual representations to align with their textual counterparts at an arbitrary granularity level through dynamic hyperbolic radius adjustment in hyperbolic space. Our method consistently improves both existing pre-training and fine-tuning MLLMs by large margins with less than 1% additional parameters.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: Those are not discussed in the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are proved clearly in the main paper or the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Detailed experimental settings and information are provided in Sec. 4 and Sec. 5, which are sufficient to reproduce the results reported in this paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be made public after acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the experimental settings are clarified in Sec. 5 and supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars or statistical significance measures mainly because the experiments were conducted in a stable and controlled environment, where repeated trials showed minimal variability. Additionally, due to computational resources and time constraints, extensive repeated runs were not feasible. Moreover, omitting such measures aligns with common practice in this research area.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Computing resource requirements are detailed in Sec. 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We ensure our research adheres to the guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits all original papers, states their versions, and respects the copyright and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Appendix of HyperET

This appendix provides additional theoretical and empirical details to support our work. Section A introduces the formal definition of hyperbolicity. Section B presents the full derivation of the core theorem introduced in the main manuscript. Section D offers an extended empirical comparison between HyperET and representative training methods in hyperbolic space, i.e., MERU, further validating the effectiveness of HyperET.

A The Definition of Hyperbolicity

This section introduces the concepts of hyperbolicity utilized in this work, including Möbius addition operation, Möbius multiplication operation, and the tangent space.

Möbius addition operation. For $\mathbf{X}, \mathbf{Y} \in \mathbb{D}_c^n$, the Möbius addition operation is defined as:

$$\mathbf{X} \oplus_c \mathbf{Y} := \frac{(1 + 2c\langle \mathbf{X}, \mathbf{Y} \rangle + c\|\mathbf{Y}\|^2)\mathbf{X} + (1 - c\|\mathbf{X}\|^2)\mathbf{Y}}{1 + 2c\langle \mathbf{X}, \mathbf{Y} \rangle + c^2\|\mathbf{X}\|^2\|\mathbf{Y}\|^2}. \tag{14}$$

Möbius scalar multiplication operation. For a scalar $r \in \mathbb{R}$ and a vector $\mathbf{X} \in \mathbb{D}_c^n$, the Möbius scalar multiplication operation is defined as:

$$r \otimes_c \mathbf{X} := (1/\sqrt{c})\tanh(r \cdot \tanh^{-1}(\sqrt{c}\|\mathbf{X}\|))\frac{\mathbf{X}}{\|\mathbf{X}\|}. \tag{15}$$

Möbius matrix multiplication operation. Refer to the relation definition in [28], for a matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ and and a point $\mathbf{X} \in \mathbb{D}_c^n$, if $\mathbf{MX} \neq \mathbf{0}$, the Möbius matrix multiplication operation is defined as:

$$\mathbf{M} \otimes_c \mathbf{X} := (\frac{1}{\sqrt{c}})\tanh\left(\frac{\|\mathbf{MX}\|}{\|\mathbf{X}\|} \cdot \tanh^{-1}(\sqrt{c}\|\mathbf{X}\|)\right)\frac{\mathbf{MX}}{\|\mathbf{MX}\|}.$$

Tangent Space. The tangent space $\mathcal{T}_{\mathbf{X}}^c\mathbb{D}_c^n$ at a point $\mathbf{X} \in \mathbb{D}_c^n$ is the first order approximation of $\mathbb{D}_c^n$, which is an n-dimensional Euclidean space. The tangent space $\mathcal{T}_{\mathbf{X}}^c\mathbb{D}_c^n$ and $\mathbb{D}_c^n$ are mapped to each other by exponential ($\mathcal{T}_{\mathbf{x}}^c\mathbb{D}_c^n \mapsto \mathbb{D}_c^n : \exp_{\mathbf{X}}^{\mathbb{D},c}(\cdot)$) and logarithmic ($\mathbb{D}_c^n \mapsto \mathcal{T}_{\mathbf{X}}^c\mathbb{D}_c^n : \log_{\mathbf{X}}^{\mathbb{D},c}(\cdot)$) maps, respectively. For any $\mathbf{X}, \mathbf{Y} \in \mathbb{D}_c^n$ and $\mathbf{V} \in \mathcal{T}_{\mathbf{X}}^c\mathbb{D}_c^n$, the mapping functions are given for $\mathbf{V} \neq \mathbf{0}$ and $\mathbf{Y} \neq \mathbf{X}$ by:

$$\exp_{\mathbf{X}}^{\mathbb{D},c}(\mathbf{V}) = \mathbf{X} \oplus_c \left(\tanh\left(\sqrt{c}\frac{\lambda_{c,\mathbf{X}}\|\mathbf{V}\|}{2}\right)\frac{\mathbf{V}}{\sqrt{c}\|\mathbf{V}\|}\right), \tag{16}$$

$$\log_{\mathbf{X}}^{\mathbb{D},c}(\mathbf{Y}) = \frac{2}{\sqrt{c}\lambda_{c,\mathbf{X}}}\tanh^{-1}\left(\sqrt{c}\|-\mathbf{X} \oplus_c \mathbf{Y}\|\right)\frac{-\mathbf{X} \oplus_c \mathbf{Y}}{\|-\mathbf{X} \oplus_c \mathbf{Y}\|}. \tag{17}$$

B Derivation of the Theorem

Theorem 1 (Hyperbolic Radius Scaling) For a point $\mathbf{X} \in \mathbb{D}_c^n$ in hyperbolic space, the hyperbolic radius adjustment function is expanded as follows:

$$\begin{aligned} s \cdot \mathrm{Rad}_{\mathbf{X}} &= \frac{2}{\sqrt{c}}(s\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}}) \\ &= \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\|s \otimes_c \mathbf{X}\|) \\ &= \mathrm{Rad}_{s\otimes_c\mathbf{X}}. \end{aligned} \tag{18}$$

where $\otimes_c$ here is instantiated as a Mobiüs scalar multiplication operation. Therefore, hyperbolic radius adjustment can be precisely controlled through $\otimes_c$ between s and $\mathbf{X}$, with the scaling coefficient s serving as a primary learnable parameter.

Proof. In a hyperbolic space, considering a point $\mathbf{X} \in \mathbb{D}_c^n$ with hyperbolic radius $\mathrm{Rad}_{\mathbf{X}} := (2/\sqrt{c})\tanh^{-1}(\sqrt{c}\|\mathbf{X}\|)$, the detailed expansion of the hyperbolic radius adjustment function is

formulated as:

$$
\begin{aligned}
s \cdot \mathrm{Rad}_{\mathbf{X}} &= \frac{2}{\sqrt{c}}(s\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}}) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}(\tanh(s\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}})) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}(\frac{\sqrt{c}}{\sqrt{c}}\tanh(s\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}})\frac{\|\mathbf{X}\|}{\|\mathbf{X}\|}) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\left\|\frac{\tanh(s\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}})}{\sqrt{c}\|\mathbf{X}\|}\mathbf{X}\right\|) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\left\|\frac{\tanh(s\tanh^{-1}(\sqrt{c}\|\mathbf{X}\|))}{\sqrt{c}\|\mathbf{X}\|}\mathbf{X}\right\|) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\|s \otimes_c \mathbf{X}\|) \\
&= \mathrm{Rad}_{s\otimes_c\mathbf{X}}.
\end{aligned} \tag{19}
$$

Consequently, according to Eq. (19), the scaling scalar s can adjust the hyperbolic radius $\mathrm{Rad}_{\mathbf{X}}$ via the Möbius scalar multiplication operation. ■

Theorem 2 (Hyperbolic Radius Flexibility Scaling) For a point $\mathbf{X} \in \mathbb{D}_c^n$ in hyperbolic space and a scaling matrix $\mathbf{X}_s \in \mathbb{R}^{n\times n}$, the flexible radius adjustment function is formally defined through Eq. (7) and Mobiüs matrix multiplication operations as follows:

$$
\begin{aligned}
\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|} \cdot \mathrm{Rad}_{\mathbf{X}} &= \frac{2}{\sqrt{c}}(\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|}\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}}) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\|\mathbf{X}_s \otimes_c \mathbf{X}\|) \\
&= \mathrm{Rad}_{\mathbf{X}_s\otimes_c\mathbf{X}}.
\end{aligned} \tag{20}
$$

Proof. In a hyperbolic space, considering a point $\mathbf{X} \in \mathbb{D}_c^n$ and a scaling matrix $\mathbf{X}_s \in \mathbb{R}^{n\times n}$, the detailed expansion of the hyperbolic radius flexibility adjustment function can be formulated as:

$$
\begin{aligned}
&\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|} \cdot \mathrm{Rad}_{\mathbf{X}} \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}(\tanh(\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|}\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}})\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}_s\mathbf{X}\|}) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}\left(\left\|\frac{\tanh(\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|}\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}})}{\|\mathbf{X}_s\mathbf{X}\|}\mathbf{X}_s\mathbf{X}\right\|\right) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}\left(\sqrt{c}\left\|\frac{\tanh(\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|}\frac{\sqrt{c}}{2}\mathrm{Rad}_{\mathbf{X}})}{\sqrt{c}\|\mathbf{X}_s\mathbf{X}\|}\mathbf{X}_s\mathbf{X}\right\|\right) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}\left(\sqrt{c}\left\|(1/\sqrt{c})\tanh\left(\frac{\|\mathbf{X}_s\mathbf{X}\|}{\|\mathbf{X}\|}\tanh^{-1}\left(\sqrt{c}\|\mathbf{X}\|\right)\right)\frac{\mathbf{X}_s\mathbf{X}}{\|\mathbf{X}_s\mathbf{X}\|}\right\|\right) \\
&= \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\|\mathbf{X}_s \otimes_c \mathbf{X}\|) \\
&= \mathrm{Rad}_{\mathbf{X}_s\otimes_c\mathbf{X}}.
\end{aligned} \tag{21}
$$

Consequently, according to Eq. (21), the scaling matrix $\mathbf{X}_s$ can adjust the hyperbolic radius $\mathrm{Rad}_{\mathbf{X}}$ via the Möbius matrix multiplication operation. ■

C Validating HyperET via Comparison with Hyperbolic Training Methods

To further validate the effectiveness of our approach, we also compare HyperET with representative training methods that utilize hyperbolic space, i.e., MERU [17]. MERU [17] and its variants establish an asymmetric visual-semantic hierarchy that emphasizes a subordinate relationship wherein textual features dominate visual ones, inherently intensifying the granularity gap between text and visual modalites, leading to inferior performance. In contrast, HyperET attempts to align visual and textual modalities at arbitrary granularity levels. As shown in Table 6, our experimental results demonstrate that HyperET achieves superior performance compared to MERU, empirically validating the effectiveness and contribution of HyperET in the context of hyperbolic geometry.

Table 6: **Comparison with MERU [17]** on 12 MLLM benchmarks, including VQAv2 [21], GQA [25], VW: VisWiZ [22], SQA: ScienceQA-IMG [42], TVQA: TextVQA [53], PE: POPE [37], ME: MME [67], MB: MMBench [41], MB^{CN}: MMBench-Chinese [41], SD: SEED-Bench [34], LVA^{W}: LLaVA-Bench (In-the-Wild) [40] and M-Vet [68]. Lan. Model: Language model. Benchmark names are abbreviated to consistent with the main manuscript. "Ours": we here realize the extra learnable parameters as full matrices, i.e., $\mathbf{W}_s$. Vision encoder: CLIP.

Method	Lan. Model	VQAv2	GQA	VW	SQA	TVQA	PE	ME	MB	MB^{CN}	SD	LVA^{W}	M-Vet
LLaVA-1.5	Vicuna-13B	80.0	63.3	53.6	71.6	61.3	85.9	1531.3	67.7	63.6	61.6	70.7	35.4
LLaVA-1.5+MERU	Vicuna-13B	78.9	61.0	48.0	65.7	60.2	84.7	1456.8	62.3	59.1	56.6	68.4	32.5
LLaVA-1.5+Ours	Vicuna-13B	**82.3**	**65.7**	**55.2**	**73.7**	**63.9**	**88.7**	**1584.7**	**69.8**	**65.2**	**63.4**	**72.6**	**38.3**

D Validation on Non-Transformer MLLMs

To demonstrating our method's effectiveness on different architectures, we experiment HyperET on a non-Transformer architecture [72], i.e., Mamba. As shown in Table 7, our method yields consistent gains across all MLLM benchmarks, further demonstrating the robustness and generalizability of our approach.

Table 7: **Comparison with Cobra [72]** on 5 MLLM benchmarks, including VQAv2 [21], GQA [25], VW: VisWiZ [22], TVQA: TextVQA [53], PE: POPE [37]. Benchmark names are abbreviated to consistent with the main manuscript. "Ours": we here realize the extra learnable parameters as full matrices, i.e., $\mathbf{W}_s$.

Method	VQAv2	GQA	VW	TVQA	PE
Cobra	79.2	63.9	56.2	59.5	87.6
Cobra+Ours	**80.9**	**65.2**	**56.9**	**60.6**	**88.8**