HyperET

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

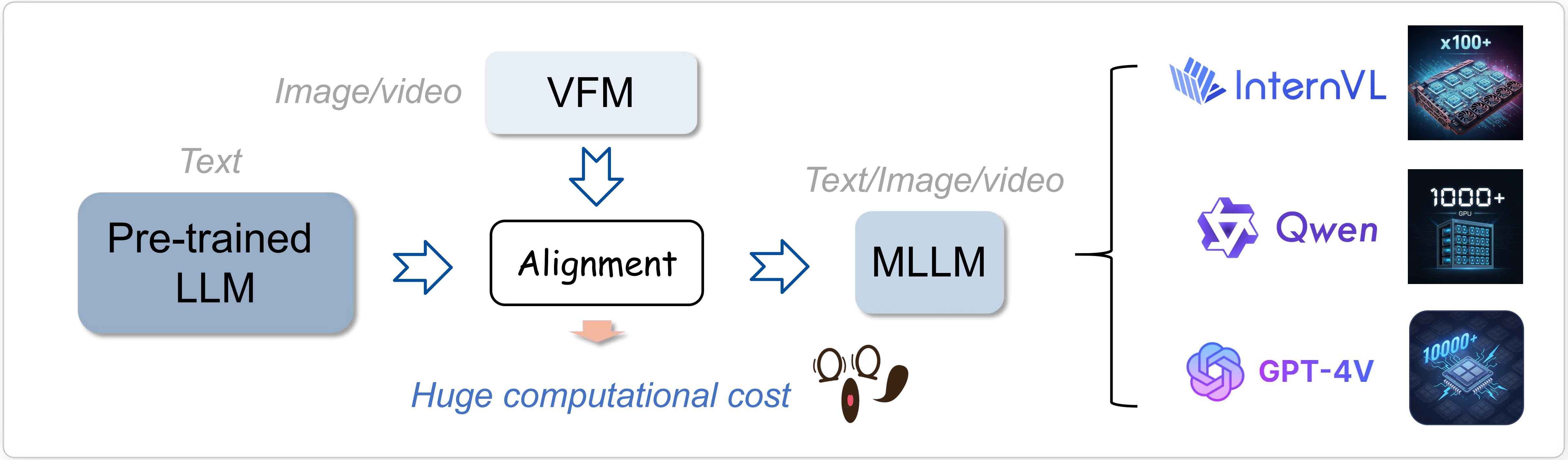
Zelin Peng[1], Zhengqin Xu[2], Qingyang Liu[1], Xiaokang Yang[1], Wei Shen[1]

[1]MoE Key Lab of Artificial Intelligence, School of Computer Science, Shanghai Jiao Tong University

[2]Shanghai Institute of Technical Physics, Chinese Academy of Sciences

Motivation-The cost

Background: MLLMs exhibit powerful capabilities, but their training is extremely costly, often requiring thousands or even tens of thousands of GPUs.

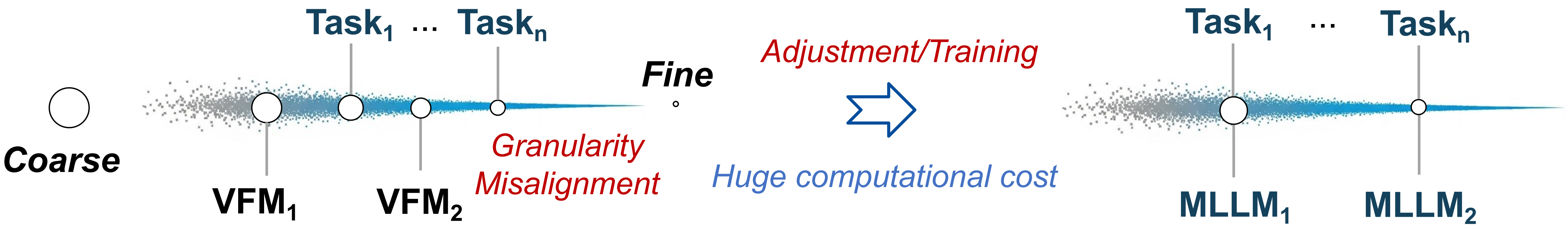


What causes the cost? The granularity misalignment

Underlying cause: A granularity misalignment exists between VFMs (e.g., CLIP) and the visual question answering tasks required by MLLMs.[1]

Status quo and Challenge: The granularity of visual embeddings is typically not finely adjustable during training. Consequently, alignment is inefficient and incurs huge computational cost.

Granularity level: coarse to fine (image-level, object-level, pixel-level, etc.)

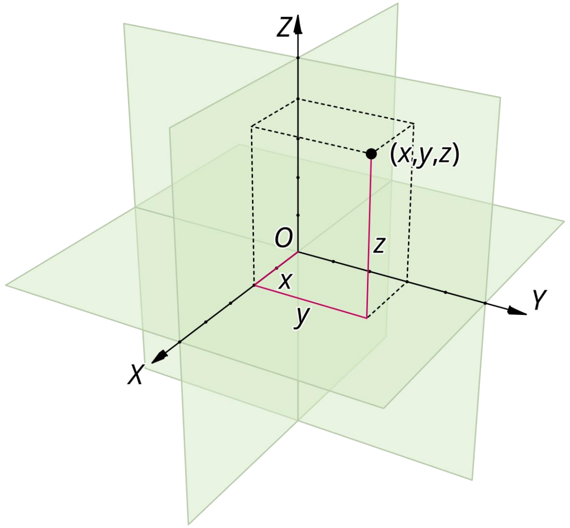


[1] Shengbang Tong, Saining Xie et al. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. CVPR2024.

Adjustment Bottleneck: Euclidean Space Limitations

Euclidean space is isotropic: All directions are equivalent, so moving a point does not effective change its level of granularity. Consequently, adjustment methods in Euclidean space are inefficient for aligning VFMs to the granularity required by MLLMs.

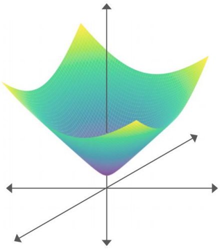
Can captures similarity between points, i.e., so-called Euclidean distance.



Euclidean space

Cannot capture any intrinsic properties of points, e.g., their granularity.

Solution – Hyperbolic geometry



Hyperbolic geometry -> Classical Poincaré ball model

⇩

Get inspiration: In the Poincaré ball model, the hyperbolic radius (i.e., the distance to the origin) can often be used to indicate the hierarchical level, e.g., the granularity of concepts.[1]

⇩

Core idea of HyperET: By tuning the hyperbolic radius of the visual embeddings for a given target task, we can effectively adjust their granularity level.

[1] Maximilian Nickel, Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. Neurips 2017.

Method - HyperET framework

Visual embeddings

Step 1

Euclidean space to Hyperbolic space

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

Step 2

Learnable Parameters & Möbius multiplication -> Adjusting hyperbolic radius

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

Step 3

Hyperbolic space to Euclidean space

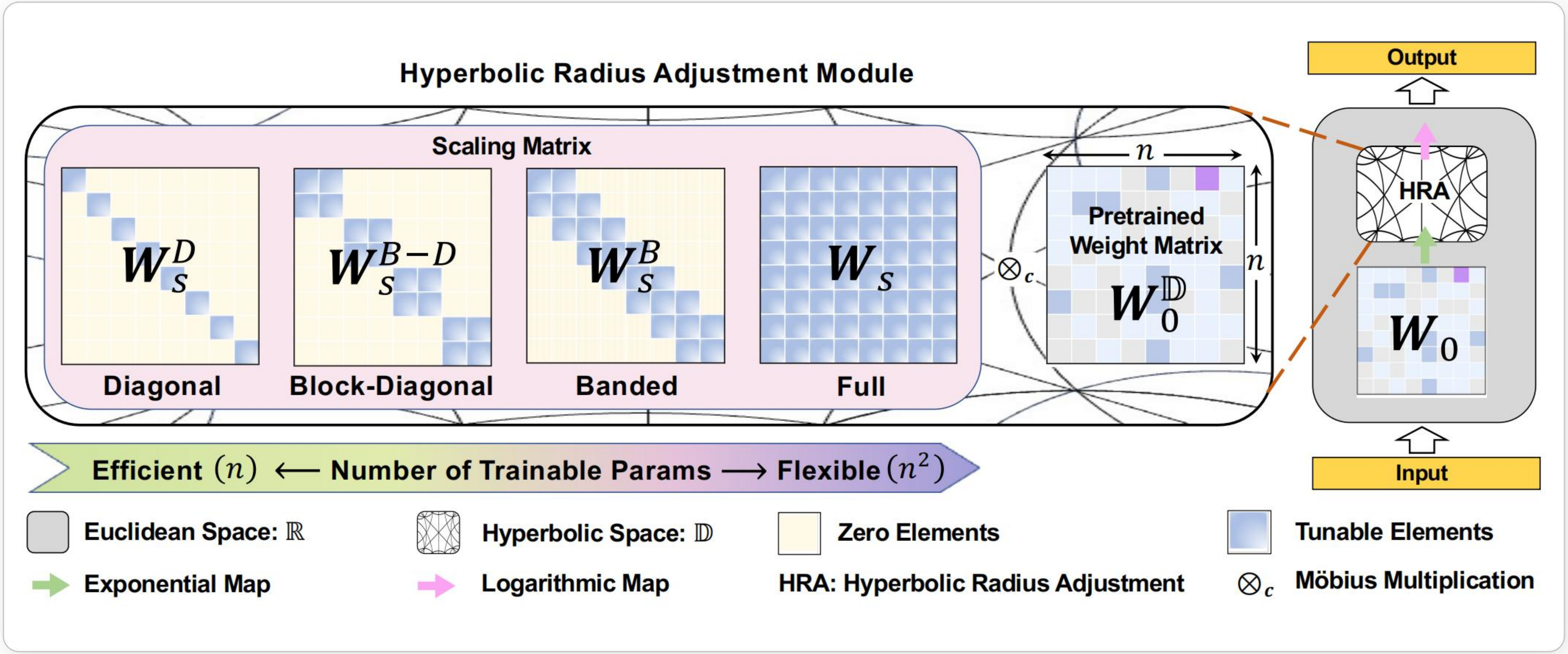
$$\log_{\mathbf{X}}^{\mathbb{D},c}(\mathbf{Y}) = \frac{2}{\sqrt{c}\lambda_{c,\mathbf{X}}}\tanh^{-1}\left(\sqrt{c}\|-\mathbf{X}\oplus_c \mathbf{Y}\|\right)\frac{-\mathbf{X}\oplus_c \mathbf{Y}}{\|-\mathbf{X}\oplus_c \mathbf{Y}\|}$$

Visual embeddings

A detailed theoretical analysis is provided in Section 4 of the main manuscript and in the supplementary material.

Parameter Efficiency Design - The Matrix Variants

Four Flexible Parametrization Params: diagonal <<block-diagonal≈banded<<full



Quantitative Results

Table 1: **Comparision with SoTA fine-tuning methods** on ScienceQA test set [40]. 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)

Table 2: **Comparison with SoTA pre-trained methods** on 12 MLLM benchmarks, including VQAv2 [20], GQA [24], VW: VisWiZ [21], SQA: ScienceQA-IMG [40], TVQA: TextVQA [53], PE: POPE [35], ME: MME [39], MB: MMBench [41], MB^{CN}: MMBench-Chinese [41], SD: SEED-Bench [32], LVA^{W}: LLaVA-Bench (In-the-Wild) [38] and M-Vet [66]. 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**

Key Ablation Study: Outperforming Euclidean Space Adjustment

Table 3: **Comparative analysis of fine-tuning spaces and flexibility levels** on ScienceQA test set [41]. 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

Qualitative Results

Spatial understanding

User	Is the man riding the bicycle?
LLaVA-1.5	Yes, the man is riding the bicycle.
LLaVA-1.5+Ours	The man is actually sitting beside the bicycle, not riding the bicycle.

Fine-grained perception

User	What's going on in this image?
LLaVA-1.5	In this image, many purple and white flowers are blooming.
LLaVA-1.5+Ours	Many purple and white flowers are blooming, and a bee is on a flower.

Analysis – Change of hyperbolic radius

Adapt CLIP with HyperET on VQA datasets

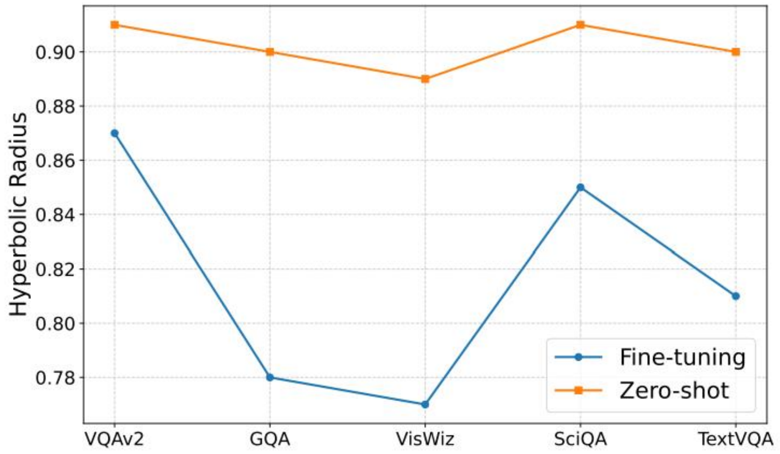
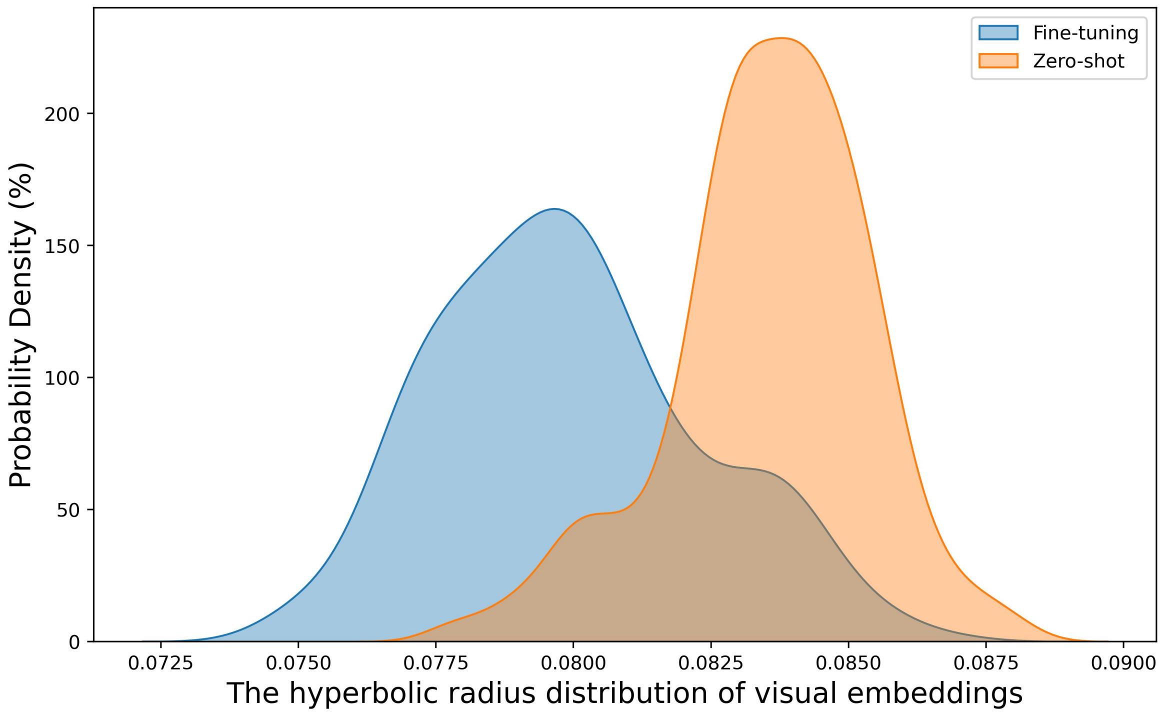


Figure 2: **Visualization of hyperbolic radius changes in visual representation after training** across different MLLM benchmarks. Normalizing the hyperbolic radius to a range of 0–1 facilitates comparison. A smaller hyperbolic radius corresponds to a more low granularity level of visual representation.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

Adapt CLIP with HyperET on segmentation datasets



More Inspiring Findings

Hyperbolic radius of visual embedding varies with ***model size*** and ***image resolution*.**

DINOV3 VITS16	DINOV3 VITB16	DINOV3 VITL16		
Range:0.70~0.85	Range:0.67~0.82	Range:0.65~0.80	Mean:0.72	Mean:0.70

Granularity level: **SAM** series models < **DINO** series models <**CLIP** series models.

Range:0.15~0.30

Range:0.55~0.85

Range:0.77~0.92

Task$_1$··· Task$_n$: Dense Prediction | Visual Grounding | Visual Question Answering | Classification

Conclusion & Take-home Messages

A New Perspective on Training MLLM

Identified **granularity mismatch** as one of the key bottlenecks in efficient MLLM training.

Proposed hyperbolic space as the ideal manifold to model granularity levels.

HyperET Framework

Introduces hyperbolic radius adjustment via learnable matrices and Möbius multiplication.

Enables **arbitrary alignment** of granularity level with target tasks.

Efficiency & Effectiveness

Achieves clear improvements with **< 1% additional parameters.**

Provides interpretability: The hyperbolic radius correlates with model size, resolution, and etc.

Thanks!

https://github.com/godlin-sjtu/HyperET

Please consider citing our paper if it is helpful in your research and development.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

中国科学院
CHINESE ACADEMY OF SCIENCES