Hyper-ICL: Attention Calibration with Hyperbolic Anchor Distillation for Multimodal In-Context Learning

Niloufar Alipour Talemi, Hossein Kashiani, Fatemeh Afghah
Clemson University

Holcombe Department of ELECTRICAL AND COMPUTER ENGINEERING
Clemson University

ICML International Conference On Machine Learning

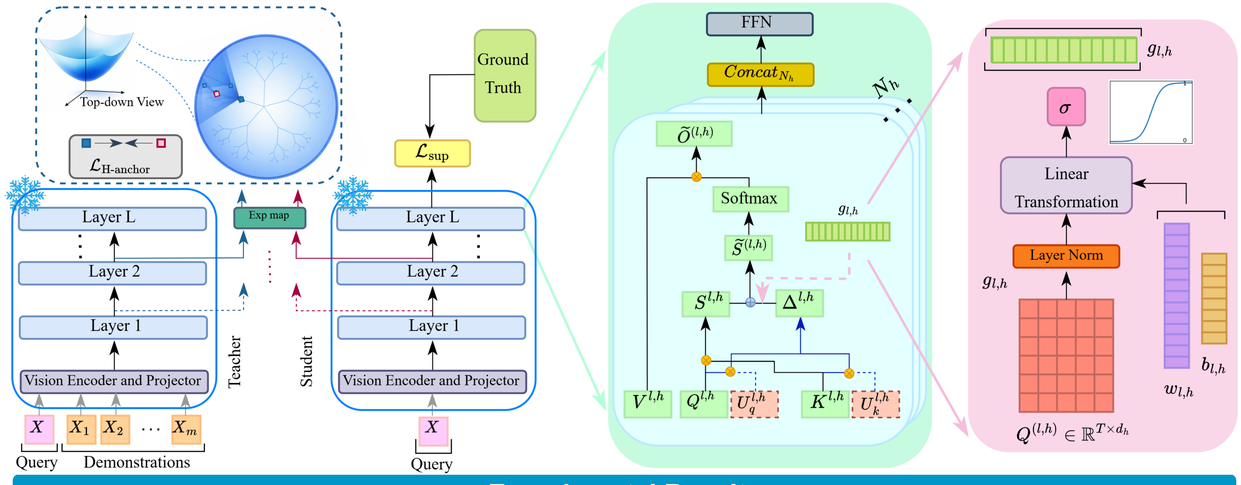
Challenges

- Multimodal ICL relies on interleaved image-text ICDs, which substantially increase input length and inference latency.
- MLLMs are sensitive to ICD formatting, ordering, and content, often overfitting to superficial demonstration cues instead of recovering robust demonstration-query relationships.

Contributions

- We introduce Hyper-ICL, an efficient multimodal ICL framework that decomposes demonstration effects within self-attention and proposes a logit-level attention intervention that directly calibrates attention distributions rather than approximating demonstration-induced shifts only at the output level.
- We propose a layer-wise hyperbolic anchor distillation loss that aligns intermediate student features to a demonstration-conditioned teacher via Lorentz geodesic distance, helping preserve relative similarity ordering under dense demonstration–query relationships and enabling demonstration-free inference.
- Experiments on two large-scale MLLMs across six widely used, challenging benchmarks (including VQAv2, OK-VQA, and COCO Caption) show that Hyper-ICL consistently improves performance and stability over direct ICD prompting, vector-based ICV baselines, and training-based alternatives.

Hyper-ICL Framework

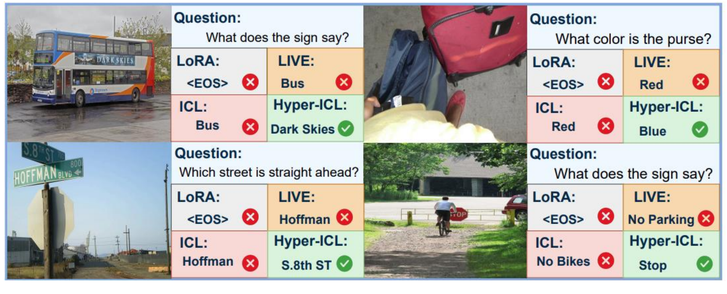


Experimental Results

Results on Idefics-9B and Idefics2-8B-base.

Model	Type	Method	# Params (M)	VQAv2	OK-VQA	COCO
Idefics-9b	Direct ICDs	Zero-shot	-	29.25	30.54	63.06
		32-shot ICL	-	56.18	48.48	105.89
		RICES	-	58.07	51.11	110.64
	Non-Learnable	FV	-	30.21	31.02	74.01
		TV	-	43.68	32.68	84.72
	Learnable	LoRA	25.0 (×21.2)	55.60	47.06	97.75
		LIVE	0.13 (×0.11)	53.71	46.05	112.76
		MimIC	0.26 (×0.22)	59.64	52.05	114.89
		Hyper-ICL	1.18 (×1)	**62.08**	**55.31**	**117.44**
Idefics2-8b	Direct ICDs	Zero-shot	-	55.39	43.08	40.00
		8-shot ICL	-	66.20	57.68	122.51
		RICES	-	66.44	55.73	111.44
	Non-Learnable	FV	-	36.47	34.58	75.24
		TV	-	47.12	38.27	87.61
	Learnable	LoRA	17.6 (×14.9)	66.54	55.05	116.69
		LIVE	0.13 (×0.11)	67.60	54.86	126.04
		MimIC	0.26 (×0.22)	69.29	58.74	132.87
		Hyper-ICL	1.18 (×1)	**71.17**	**62.24**	**135.66**

Qualitative comparison of hallucination behavior across methods on VQA.



Analysis of inference FLOPs and runtime.

Metric	Hyper-ICL	0-Shot	8-shot	16-shot	32-shot
FLOPs (T)	0.955	0.935	6.375	12.341	23.364
Runtime (ms)	59.32	56.69	158.21	266.13	468.55

Methodology

- **In-Context Attention Calibration**: Hyper-ICL reconstructs ICD effects within self-attention by adding a parameter-efficient low-rank logit-level adapter to the attention logits:

$$\widetilde{S}^{(l,h)} = S^{(l,h)} + \mathrm{Diag}\big(g_{l,h}\big)\,\Delta^{(l,h)}$$

This directly calibrates the attention distribution rather than approximating ICD effects as output-level shifts.

- **Query-adaptive Token-wise Modulation**: A token-wise modulation vector controls the intervention strength for each query token across layers and heads:

$$g_{l,h} = \sigma\Big(\mathrm{LN}\big(Q^{(l,h)}\big)\,w_{l,h} + b_{l,h}\mathbf{1}\Big)$$

- **Layer-wise Hyperbolic Anchor Distillation:** A query-only student is aligned with a demonstration-conditioned teacher using Lorentz geodesic distance:

$$\mathcal{L}_{\text{H-anchor}} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{T}\sum_{i=1}^{T} d_{\mathcal{L}}^2\Big(P_i'^{(l)},\, P_i^{(l)}\Big)$$

This preserves layer-wise demonstration-query relationships for demonstration-free inference.

Ablation Studies

- Compares Hyper-ICL with LoRA, LIVE, and MimIC.
- Evaluated across training sample sizes on VQAv2 and OK-VQA.
- Uses Idefics-9B and Idefics2-8B-base backbones.