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

Electrical and Computer Engineering Department

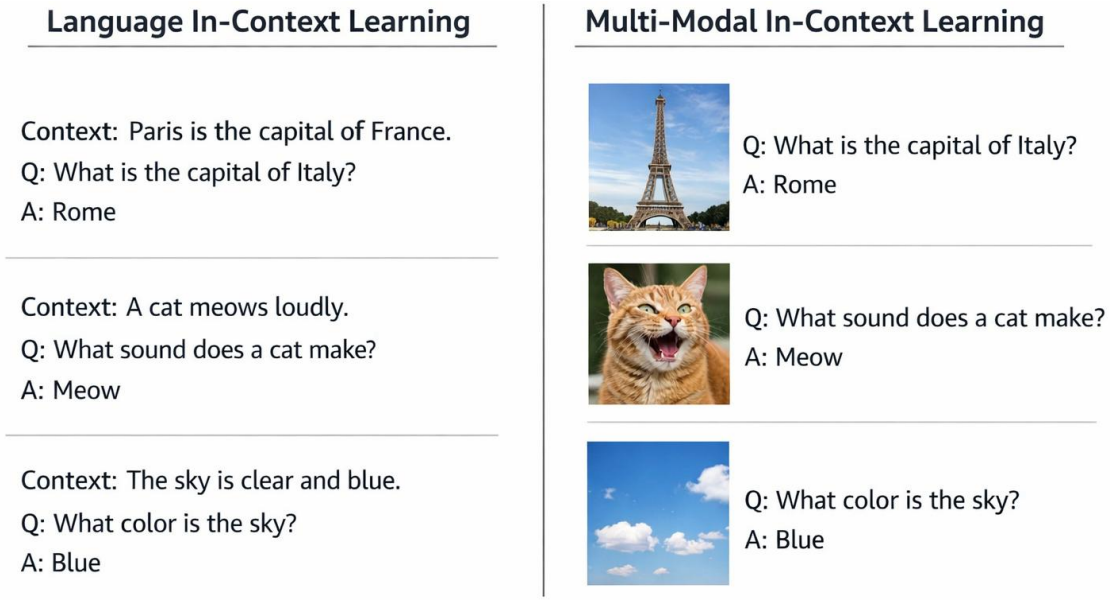
Clemson University

Niloufar Alipour Talemi, Hossein Kashiani, Fatemeh Afghah

Introduction and Motivation

What is In-Context Learning?

- It is a way for an LLM or MLLM to learn **from examples placed inside the input prompt**, instead of updating its weights.
- It does **not update model weights**, meaning the model is not retrained or fine-tuned.
- It uses **the pattern** in the examples to infer the task and answer a new input in the same way.



Challenges:

- Its performance can become unstable under **formatting changes** or **different example orderings**.
- Multimodal ICL is **expensive at inference time**.
- The problem is even worse in **multimodal settings**, where cross-modal alignment adds extra complexity and variance.
- In tasks like **VQA**, an MLLM may copy the **answer format** seen in demonstrations instead of learning the true input-output relationship.

Problem Setup:

For a query vector q ∈ Q, the single-head self-attention computation is given by:

$$\mathrm{SA}\left(q, \begin{bmatrix} K_D \\ K \end{bmatrix}, \begin{bmatrix} V_D \\ V \end{bmatrix}\right) = \mathrm{softmax}\left(\begin{bmatrix} qK_D^\top \\ qK^\top \end{bmatrix}\right)^\top \cdot \begin{bmatrix} V_D \\ V \end{bmatrix}$$

$$= \frac{\begin{bmatrix} \exp(qK_D^\top) \\ \exp(qK^\top) \end{bmatrix}^\top}{Z_1 + Z_2} \cdot \begin{bmatrix} V_D \\ V \end{bmatrix}$$

$$= \left[\frac{Z_1}{Z_1 + Z_2} \cdot \frac{\exp(qK_D^\top)}{Z_1} V_D + \frac{Z_2}{Z_1 + Z_2} \cdot \frac{\exp(qK^\top)}{Z_2} V\right]$$

$$= (1-\mu)\mathrm{SA}(q, K, V) + \mu \mathrm{SA}(q, K_D, V_D)$$

$$f_i(x) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

Independent of the ICDs

Shift effects caused by the ICDs

where $Z_1(q, K_D) = \sum_{i=1}^{l_D} \exp(qK_D^{i\top})$, and $Z_2(q, K) = \sum_{j=1}^{l_q} \exp(qK^{j\top})$.

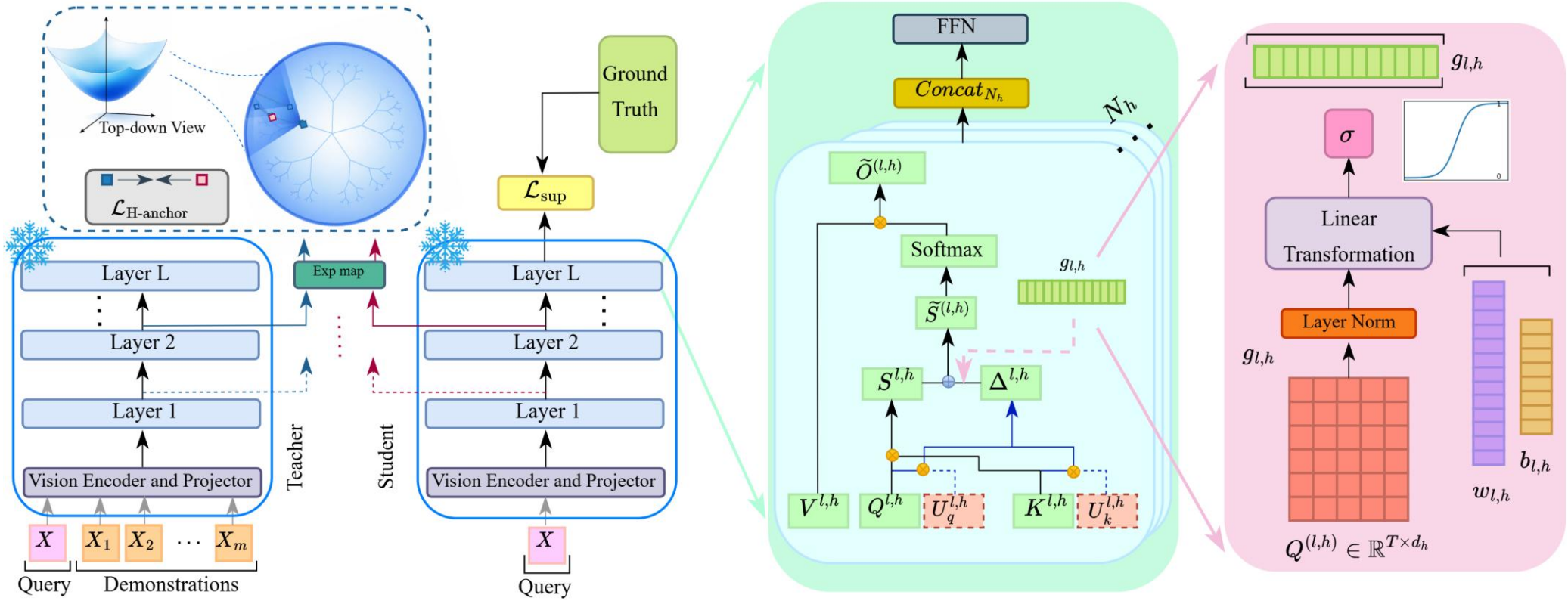
- In in-context settings, demonstrations primarily affect **where** the model attends by injecting additional key-value evidence.
- We propose an intervention that directly **calibrates the attention** distribution via a learnable logit-level bias.

Proposed Method

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

Main Contributions:

- Low-rank Logit-level Adapter
- Query-adaptive Token-wise Modulation
- Layer-wise Hyperbolic Anchor Distillation



Low-rank Logit-level Adapter:

At layer l, and attention head h, the standard attention logits are computed using the scaled dot product between queries and key:

$$S^{(l,h)} = \frac{Q^{(l,h)}\left(K^{(l,h)}\right)^{\top}}{\sqrt{d_h}} \quad , \qquad Q^{(l,h)}, K^{(l,h)} \in \mathbb{R}^{T\times d_h} \text{ and } S^{(l,h)} \in \mathbb{R}^{T\times T} \quad , \qquad O^{(l,h)} = \text{softmax}\left(S^{(l,h)}\right) V^{(l,h)}$$

We intervene directly on the pre-softmax logits to steer the attention distribution:

$$\widetilde{S}^{(l,h)} = S^{(l,h)} + \text{Diag}\left(g_{l,h}\right) \Delta^{(l,h)}$$

$$A^{(l,h)} = Q^{(l,h)} U_q^{(l,h)} \in \mathbb{R}^{T\times r} \quad , \qquad r \ll d_h \quad , \qquad \Delta^{(l,h)} = \frac{A^{(l,h)} B^{(l,h)T}}{\sqrt{r}} \in \mathbb{R}^{T\times T}$$
$$B^{(l,h)} = K^{(l,h)} U_k^{(l,h)} \in \mathbb{R}^{T\times r}$$

Then: $$\widetilde{O}^{(l,h)} = \text{softmax}\left(\widetilde{S}^{(l,h)}\right) V^{(l,h)}$$

Query-adaptive Token-wise Modulation:

We propose a token-wise modulation mechanism that adaptively scales the intervention strength for each layer and head based on the current query representation.

For layer l and head h:

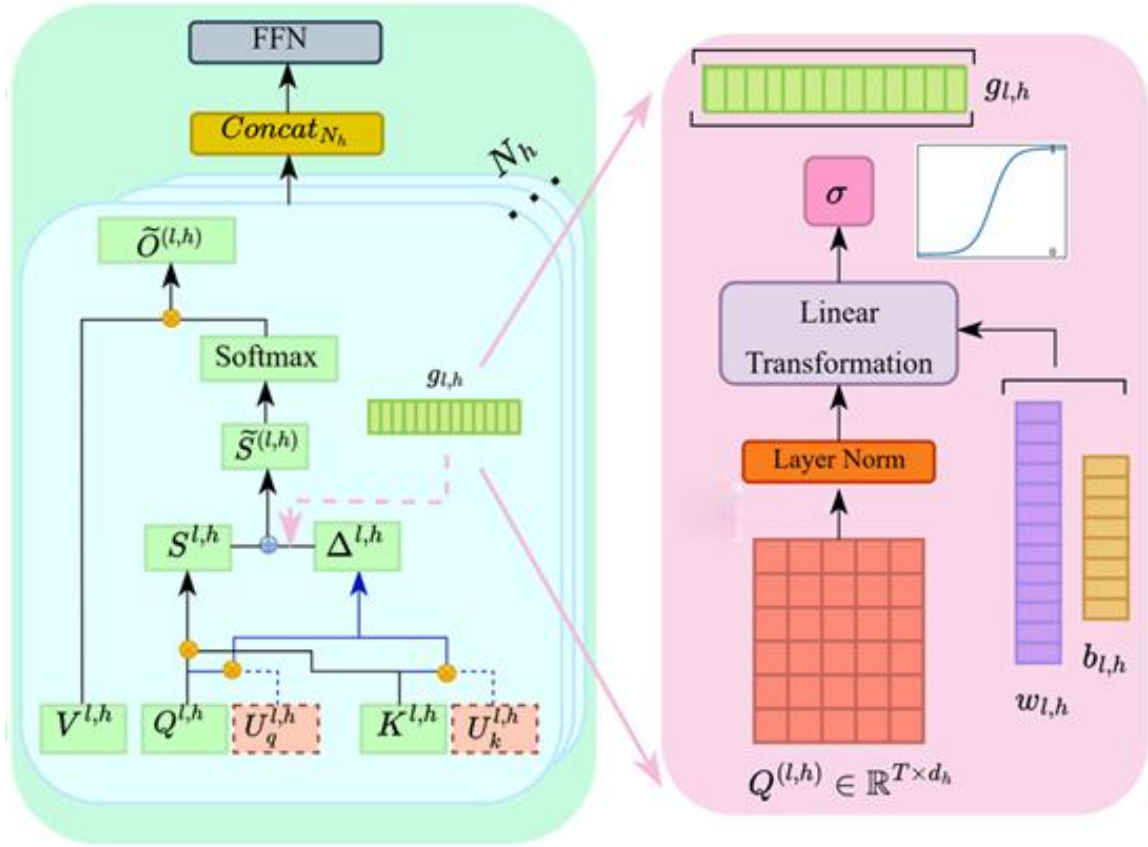
$$g_{l,h} = \sigma\Big(\mathrm{LN}\big(Q^{(l,h)}\big)\, w_{l,h} + b_{l,h}\mathbf{1}\Big)$$

and

$$g_{l,h} \in (0,1)^T ,$$

where $w_{l,h} \in \mathbb{R}^{d_h}$ and $b_{l,h} \in \mathbb{R}$ are learnable parameters .

This design produces one modulation coefficient per query token, enabling fine-grained and **input-dependent** control over the intervention magnitude within **each attention head**.



Layer-wise Hyperbolic Anchor Distillation

We introduce an intermediate alignment regularizer that matches the student's internal representations to those of a frozen teacher model that observes full demonstrations.

Given the hyperbolic embeddings, $P_i^{(l)}$ and $P_i'^{(l)}$, we measure alignment using the Lorentz geodesic distance:

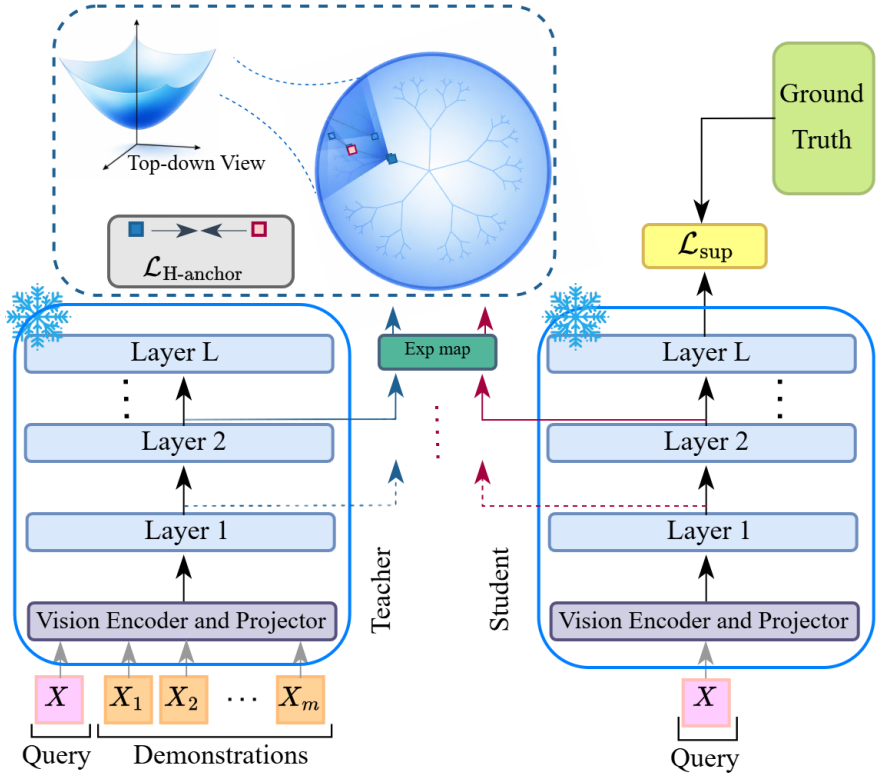
$$d_L(P', P) = \sqrt{\frac{1}{\kappa}} \operatorname{arcosh}\Big(-\kappa \langle P', P \rangle_L \Big)$$

Layer-wise hyperbolic anchor distillation loss that aligns student and teacher representations across all transformer layers:

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

Therefore, the final training objective combines these losses as a weighted sum:

$$\mathcal{L} = \mathcal{L}_{\text{H-anchor}} + \lambda \mathcal{L}_{\text{sup}}$$



Experiments

- Comparison with Existing Methods:
  - Direct use of ICDs
  - Non-learnable vector-based ICDs
  - Learnable use of ICDs
- Comprehensive Ablation Studies and Discussions
  - Inference Efficiency Analysis
  - Hallucination Analysis
  - Analysis of Curvature and Loss-Weight Sensitivity

Benchmark:

- **VQAv2** (204,721 images, 1,105,904 questions, and 10 human answers per question).
- **OK-VQA** (14,055 open-ended questions, 5 ground-truth answers per question).
- **COCO Caption** (5 captions per image).
- **Flickr30k** (31,783 images and 158,915 English captions, with 5 captions per image).
- **MME** (It measures both perception and cognition across 14 subtasks).
- **SEED-Bench** (19K multiple-choice questions with human annotations and covers 12 evaluation dimensions, including understanding of both images and videos).

MLLMs

- **Idefics-9B** (cross-attention architecture)
- **Idefics2-8B-base** (fully autoregressive architecture)

COCO Images

OK-VQA Examples

Q: What sort of vehicle uses this item?
A: firetruck

Q: How many chromosomes do these creatures have?
A: 23

Comparison with a broad range of baselines across three datasets:

- **Non-learnable ICV methods** such as **FV** and **TV** remain clearly weaker than few-shot ICL and trainable alternatives.
- **Hyper-ICL achieves the best overall results** across both backbones and all three benchmarks.
- Hyper-ICL delivers these gains with only **1.18M trainable parameters**, making it far more parameter-efficient than **LoRA.**

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**

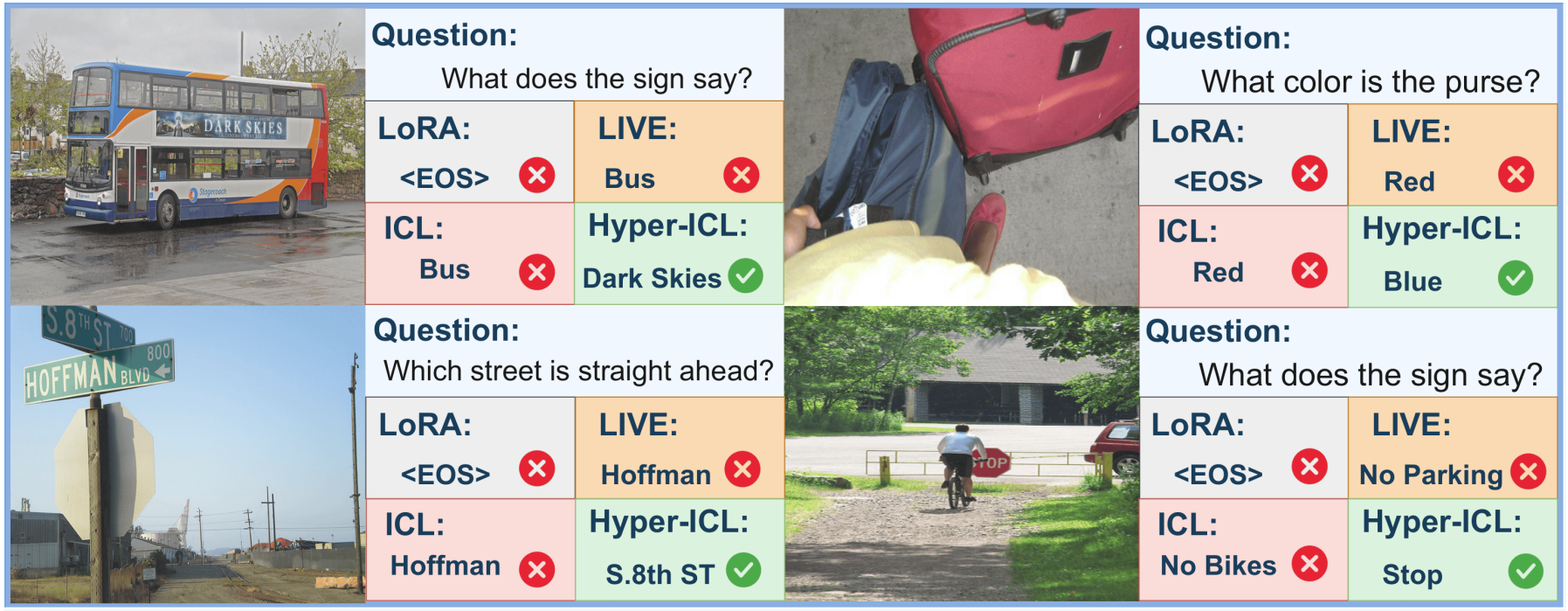
Generalize to More Challenging Benchmarks:

- Across **both backbones**, Hyper-ICL **consistently outperforms zero-shot and standard ICL** on all three tasks.
- On **Idefics-9B**, Hyper-ICL improves performance from **63.41 to 75.96** on Flickr30k, **52.11 to 65.46** on MME, and **28.30 to 31.87** on SEED-Bench over ICL.
- On **Idefics2-8B-base**, Hyper-ICL reaches the **best scores overall**, with **93.79** on Flickr30k, **82.13** on MME, and **48.31** on SEED-Bench.

Model	Method	Flickr30k	MME	SEED
Idefics-9b	Zero-shot	49.17	55.36	27.56
	ICL	63.41	52.11	28.30
	Hyper-ICL	**75.96**	**65.46**	**31.87**
Idefics2-8b-base	Zero-shot	53.04	74.80	12.91
	ICL	84.57	71.10	47.90
	Hyper-ICL	**93.79**	**82.13**	**48.31**

- The gains suggest that **logit-level attention calibration and layer-wise distillation transfer effectively across diverse multimodal tasks**, not just standard VQA settings.

Qualitative comparison of hallucination behavior across methods on VQA

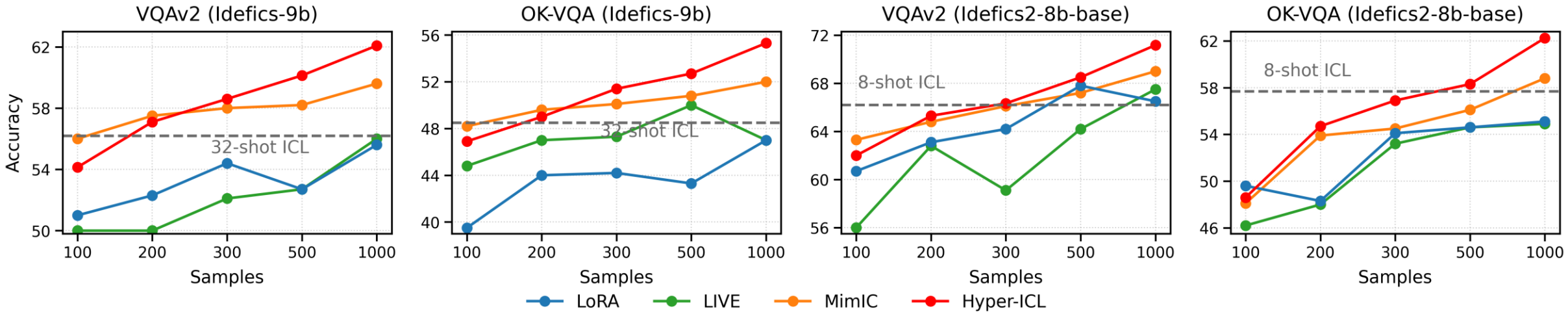


- Hyper-ICL stays aligned with visible evidence on representative VQA examples.
- ICL, LoRA, and LIVE more often produce ungrounded or incorrect answers.
- The figure highlights a key advantage of Hyper-ICL: more reliable multimodal reasoning.

Ablation and Analysis

- We assess the individual contributions of each component of the proposed frameworks.
- The ablations show that these components complement each other and jointly improve the performance.

Comparison of Hyper-ICL with SOTA Works Across Training Sample Sizes

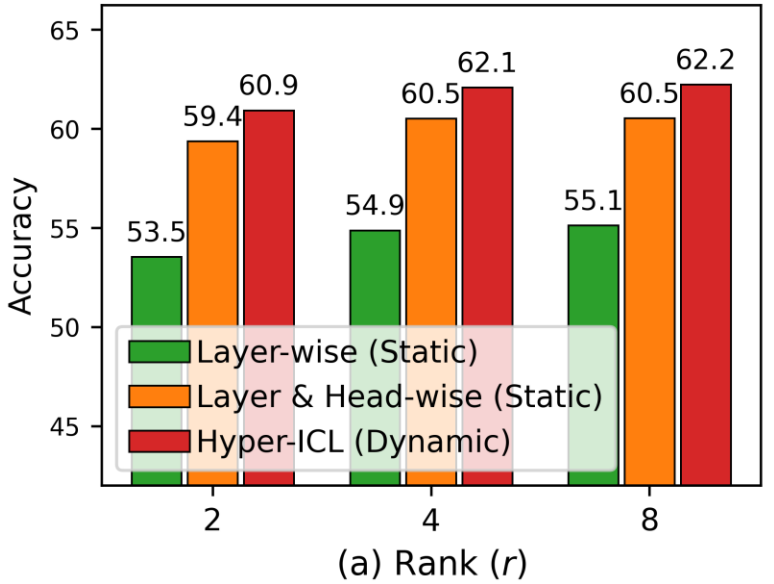


- As the number of training samples increases, **all methods improve**, showing that **supervised adaptation remains effective even in the low-data regime**.
- Hyper-ICL **surpass the standard few-shot ICL baseline with far fewer training examples**, highlighting its **data efficiency**.
- The advantage is larger on **OK-VQA**, where success depends more on **knowledge-based reasoning** and **better modeling of demonstration-query relations**.

Analysis of Adapter Rank and Query-Adaptive Tokenwise Modulation:

- **Accuracy improves as rank increases** for all methods.
- **Layer-and-head-wise** static calibration is better than simple layer-wise calibration.
- The gains become **small** after **r = 4**.
- A moderate rank already **captures most of the useful** improvement without adding much extra cost.

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



Analysis of inference FLOPs and Runtime:

- **Hyper-ICL is close to zero-shot efficiency** (0.955T FLOPs, 59.32 ms vs. 0.935T, 56.69 ms).
- **Standard ICL cost increases rapidly** as the number of demonstrations grows.
- Hyper-ICL **keeps demonstration benefits** while maintaining near zero-shot efficiency.

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

Thank you!

 nalipou@clemson.edu

 https://nilouap.github.io

 https://github.com/NilouAP