VL-ICL Bench: The Devil in the Details of Multimodal In-Context Learning

Yongshuo Zong*, Ondrej Bohdal*, and Timothy Hospedales

University of Edinburgh   * Co-first authors

@yongshuozong

yongshuo.zong@ed.ac.uk

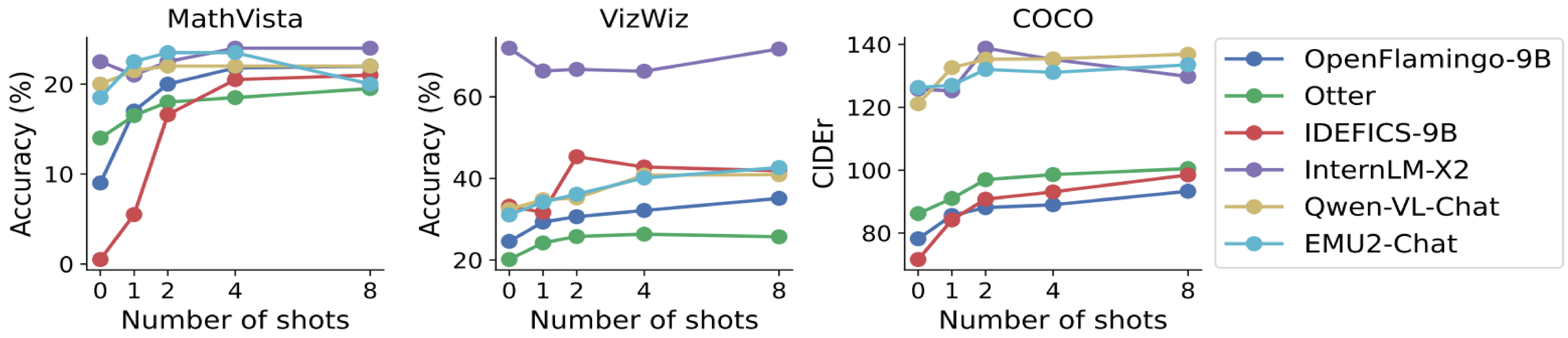
ICLR

Paper & Code

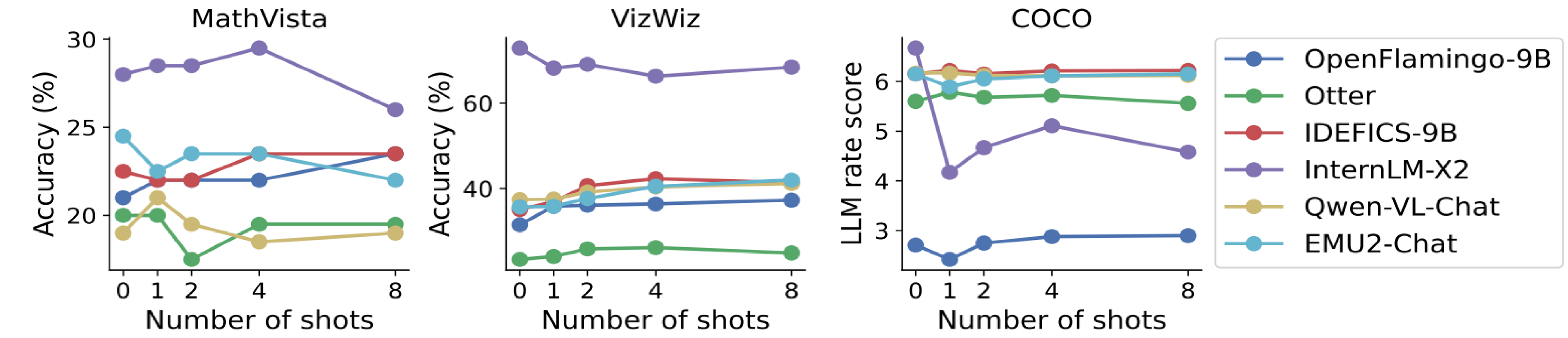
THE UNIVERSITY of EDINBURGH

VQA and Captioning are Poor Benchmarks for Multimodal ICL

😳 **ICL on these benchmarks primarily learns answer <u>style/format</u>**

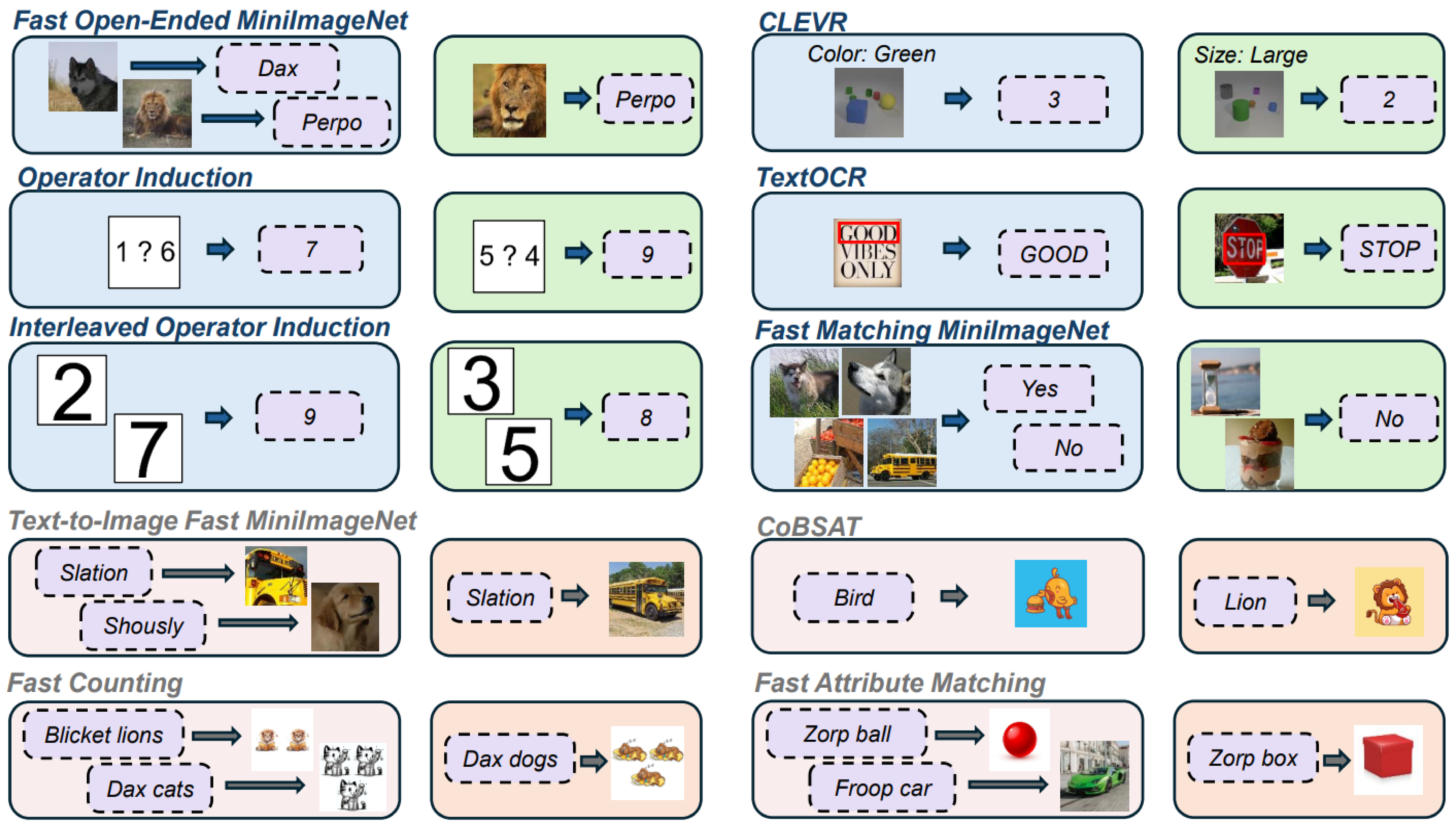


(a) Exact match evaluation.



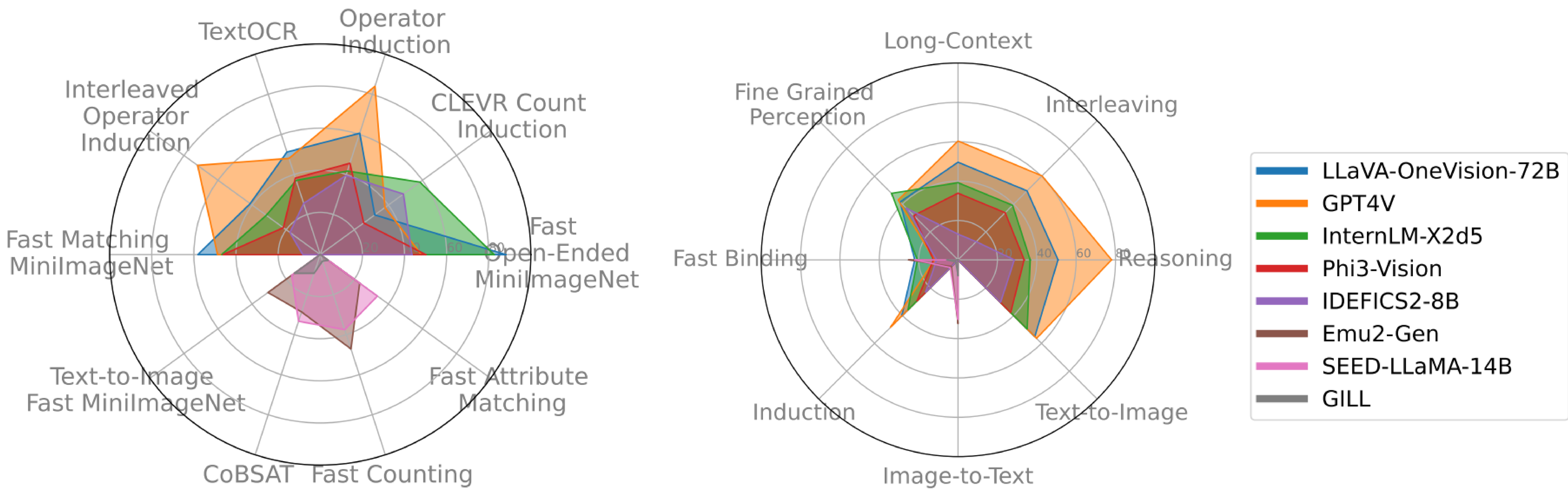
VL-ICL is a Better ICL Benchmark for Multimodal Input and Output

😍 **VL-ICL is hard or impossible to solve in zero-shot**

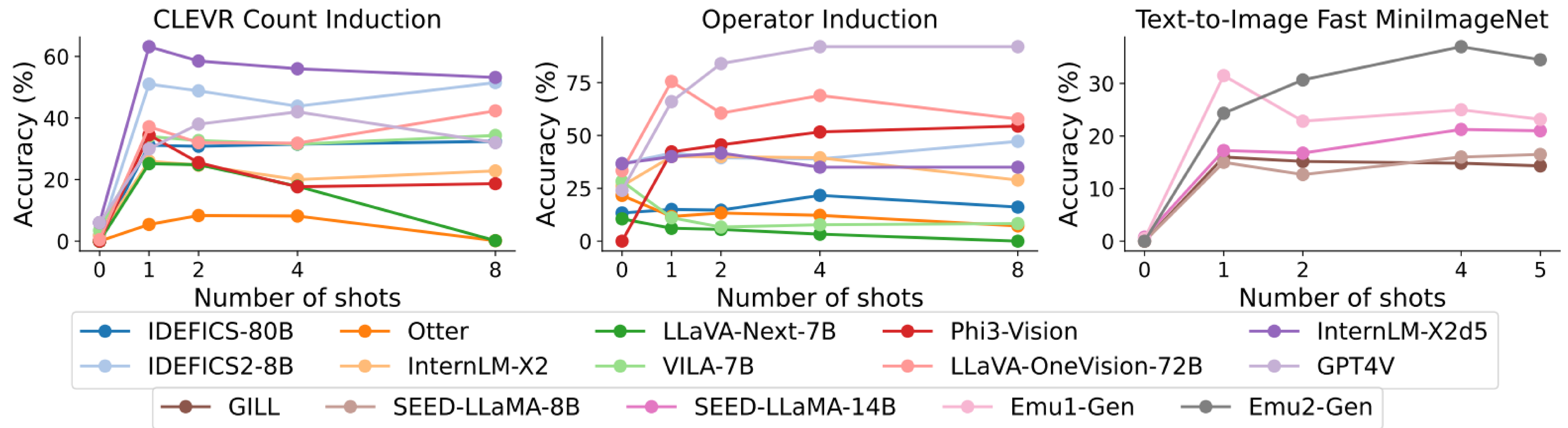


Dataset	Capabilities Tested	Train Set	Test Set
Fast Open MiniImageNet	I2T, Fast Binding	5,000	200
CLEVR Count Induction	I2T, Fine Grained Perception, Induction	800	200
Operator Induction	I2T, Induction, Reasoning	80	60
Interleaved Operator Induction	I2T, Induction, Reasoning, Interleaving, Long-Context	80	60
TextOCR	I2T, Fine Grained Perception, Induction	800	200
Matching MiniImageNet	I2T, Induction, Interleaving, Long-Context	1,600	400
Text-to-image MiniImageNet	T2I, Fast Binding	5,000	200
CoBSAT	T2I, Induction	800	200
Fast Counting	T2I, Fast Binding	800	40
Fast Attribute Matching	T2I, Fast Binding	300	200
Total	T2I, I2T, Binding, Perception, Long-Context, Interleaving, Induction, Reasoning	15,260	1,760

Main Results



Model	Avg. Rank		
	Z.s.	Pk.	Eff.
OpenFlamingo-9B	9.3	13.3	8.7
IDEFICS-9B	7.7	10.8	8.7
IDEFICS-80B	7.0	7.8	5.8
IDEFICS2-8B	3.0	5.5	8.5
Otter	8.0	14.5	13.0
InternLM-X2	5.5	9.3	10.0
Qwen-VL-Chat	7.7	10.8	8.3
LLaVA-Next-7B	7.3	13.7	13.2
Emu2-Chat	4.0	10.2	11.0
VILA-7B	3.7	8.0	10.8
Mantis-Idefics2	4.5	6.7	9.7
Phi3-Vision	9.2	6.3	5.3
LongVA-7B	6.3	10.5	9.3
LLaVA-OneVision-72B	2.5	**1.7**	**3.8**
InternLM-X2d5	**1.5**	3.7	5.8
GPT4V	2.8	2.3	**3.8**



Model	Avg. Rank		
	Z.s.	Pk.	Eff.
GILL	3.5	4.8	4.8
SEED-LLaMA-8B	3.5	2.5	**2.2**
SEED-LLaMA-14B	2.0	**2.0**	2.5
Emu1-Gen	3.5	3.8	2.8
Emu2-Gen	**1.5**	**2.0**	2.8

Findings

★ **VLLMs demonstrate non-trivial ICL on VL-ICL Bench tasks.**

★ **VLLMs often struggle to make use of a larger number of ICL examples.**

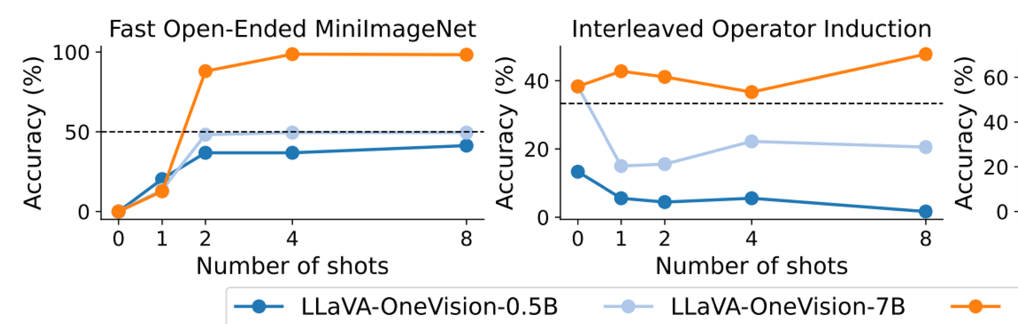
★ **LLaVA-OV-72B/GPT4V is the best overall I2T model.**

★ **No clear winner among text-to-image models.**

★ **Zero-shot performance ≠ ICL ability.**

Further Analysis

1. Emergent threshold of multimodal ICL



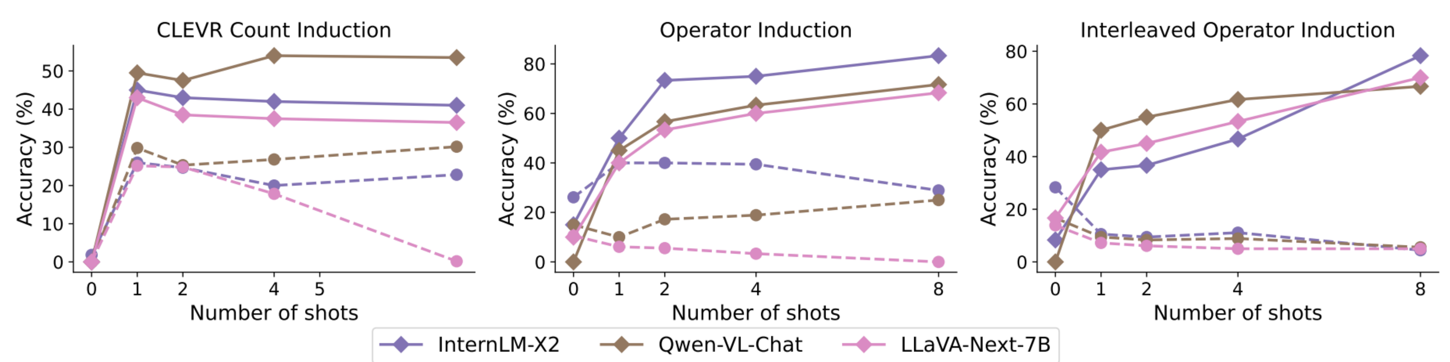
Comparison of different model sizes (dashed line indicates random chance). 72B model understands the tasks, while the smaller models fail, highlighting the impact of model size on ICL and the <u>presence of an emergent threshold</u>.

2. Disentangling context length and in-context learning

Dataset	Fast Open-Ended MiniImageNet			CLEVR Count Induction			Operator Induction			TextOCR		
	Z.s.	Pk.	Eff.	Z.s.	Pk.	Eff.	Z.s.	Pk.	Eff.	Z.s.	Pk.	Eff.
LLaVA-Next-7B (w/o SelfExtend)	0.0	37.2	29.4	0.0	25.2	19.3	10.6	10.6	-6.8	24.7	24.7	**-23.0**
LLaVA-Next-7B (w/ SelfExtend)	0.0	**51.0**	**38.9**	0.0	**29.0**	**25.4**	**11.7**	**11.7**	**-5.8**	**26.0**	**26.0**	-23.7
VILA-7B (w/o SelfExtend)	0.0	38.2	32.3	3.5	34.3	**27.5**	**28.3**	**28.3**	**-18.9**	**28.0**	**30.2**	**-3.7**
VILA-7B (w/ SelfExtend)	0.0	**54.0**	**40.0**	**4.0**	**34.8**	**27.5**	**28.3**	**28.3**	-20.3	**28.0**	29.7	-4.5

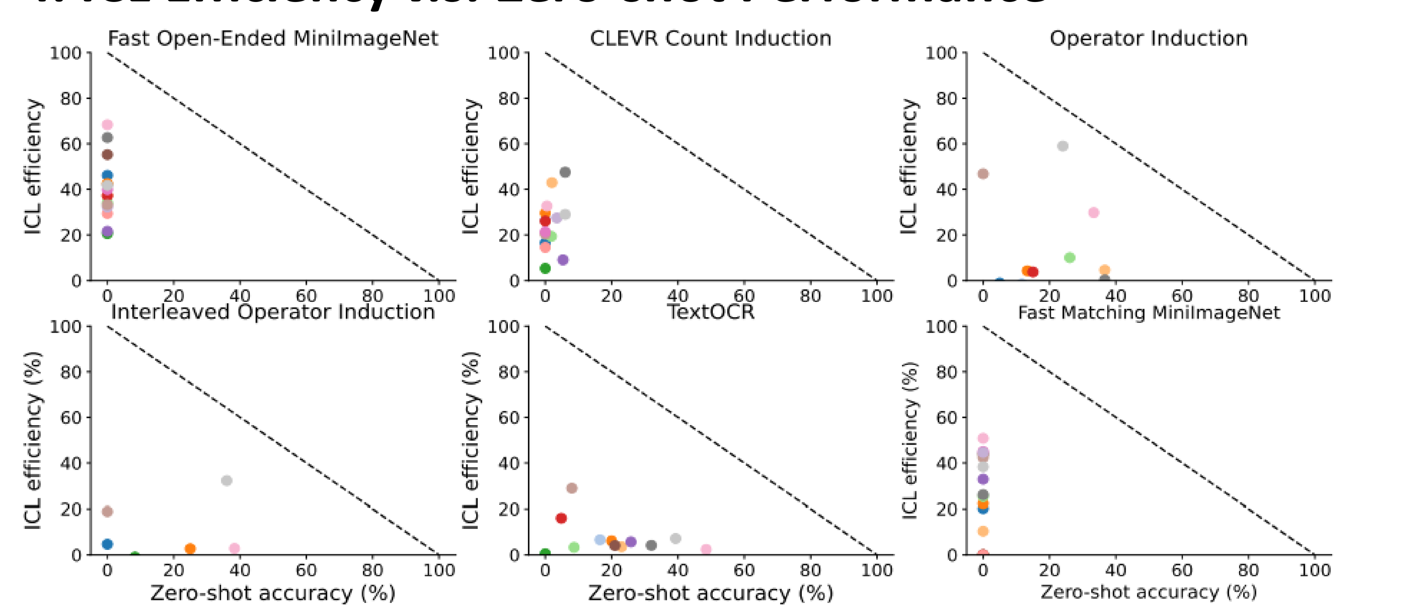
Comparison of models with and without context extension strategy (SelfExtend). While it is helpful in some cases, context extension does **not** necessarily improve the performance of ICL.

3. Text *v.s.* Multimodal ICL



Comparison of multimodal (dashed line) and text (solid line) ICL: Text shows sharper, steadier gains, underscoring multimodal ICL's difficulty.

4. ICL Efficiency *v.s.* Zero-shot Performance



Zero-shot performance + ICL efficiency = 100%