ICLR
International Conference On Learning Representations

Meta-Adaptive Prompt Distillation for Few-Shot Visual Question Answering

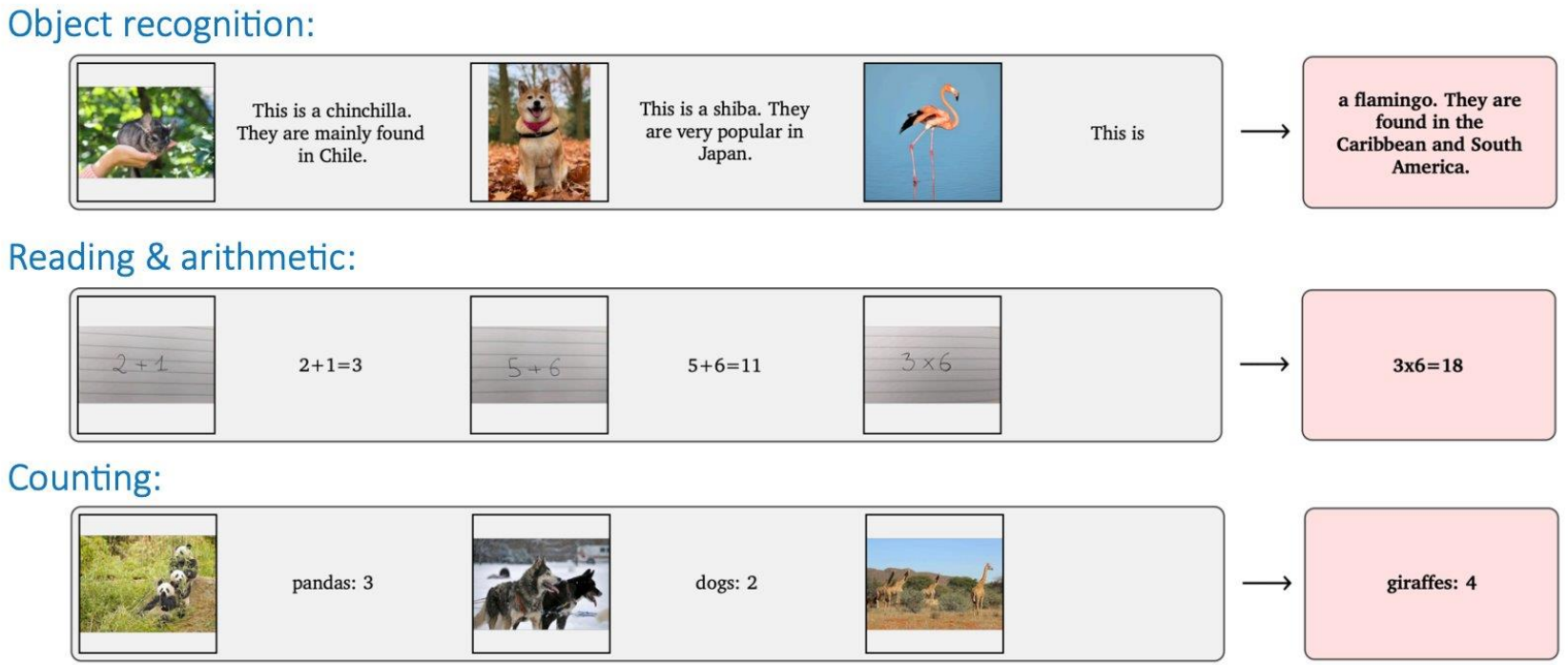
Akash Gupta, Amos Storkey, Mirella Lapata

School of Informatics, University of Edinburgh

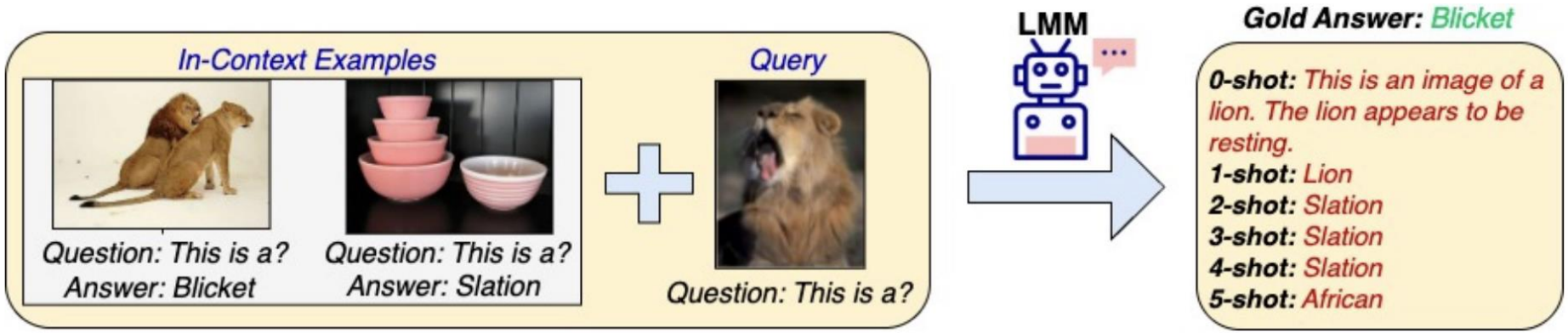
THE UNIVERSITY of EDINBURGH

Large Multimodal Models are few-shot learners

- Large Multimodal Models or LMMs are trained on large amounts of vision and language data.
- This allows them to develop general-purpose generation capabilities like in-context learning (ICL)
- ICL is highly beneficial, as given a new task, it allows models to adapt their generation by prompting with a few examples or shots.



But Image-text ICL can fail even for simple tasks

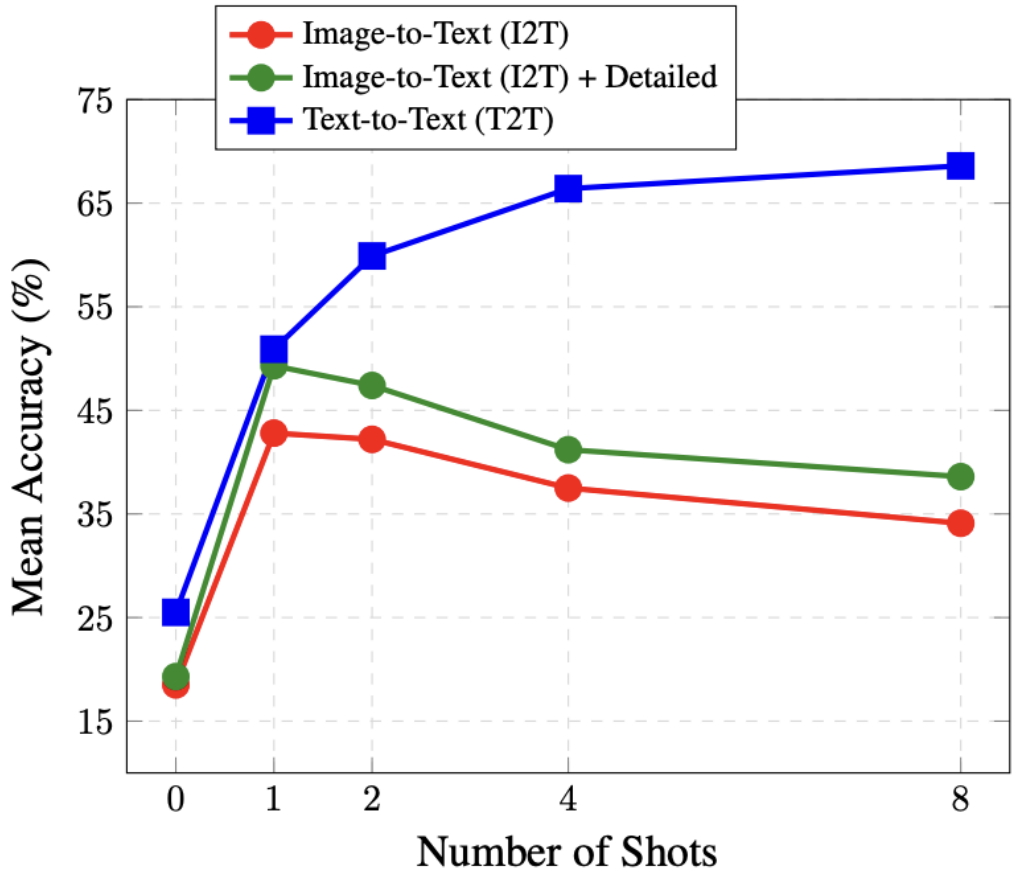


- We test LLaVA-OneVision-7B [1] LMM on the Fast Open-Ended MiniImageNet task from the VL-ICL Benchmark [2].
- This is an image classification task, where original image labels are replaced with non-sensical categorical names like "blicket" or "slation", to remove any dependency over LLM's prior knowledge.
- We see that increasing the number of examples (shots) does not help and the LMM keeps generating the wrong label for the query image

[1] Li et al. LLaVA-OneVision: Easy Visual Task Transfer. 2024
[2] Zong et al. VL-ICL BENCH: THE DEVIL IN THE DETAILS OF MULTIMODAL IN-CONTEXT LEARNING. 2025

But text-only ICL performs much better

- We further test whether increasing the number of shots also affect unimodal text-only ICL performance for tasks in the VL-ICL Benchmark [1].
- Text-only (T2T) ICL (**blue**) significantly outperforms image-text (I2T) ICL (**red**) with detailed instructions (**green**), and improves monotonically with additional shots.
- This strongly suggests that naively increasing the number of image embeddings, impairs the model's inherent ICL ability.



[1] Zong et al. VL-ICL BENCH: THE DEVIL IN THE DETAILS OF MULTIMODAL IN-CONTEXT LEARNING. 2025

Our approach

- We provide an alternative to ICL for LMMs, by learning a fixed set of new embeddings that can be finetuned at test-time for any task.

- This approach is inspired by the previous work on prompt tuning [1], which shows that frozen language models can be conditioned with learnable soft prompt embeddings to perform a downstream task.

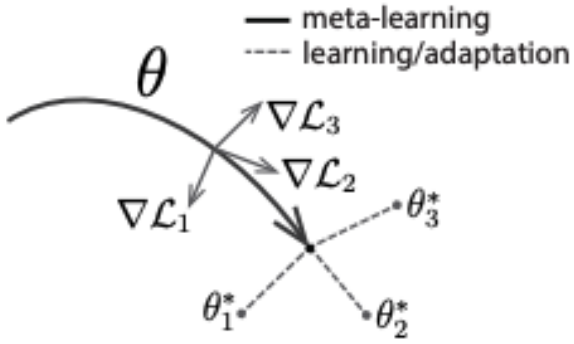
- Soft prompts can be easily trained using backpropagation and receive task information in the form of gradient updates.

- We term this class of method as **prompt distillation**, given that we are essentially distilling relevant image features from the ICL examples into soft prompts.

[1] Lester et al. The Power of Scale for Parameter-Efficient Prompt Tuning. 2021

Meta learning for few-shot adaptation

- Traditional finetuning methods are not suitable for gradient-based adaptation in few-shot data regimes and could lead to overfitting.
- Prior work [1,2] has addressed this challenge by training a meta-learner that can infer an optimal learning strategy for a new task after being exposed to a distribution of tasks.



- We apply this procedure to our multimodal prompt distillation setting by employing the widely known MAML algorithm [1] and use its lightweight first-order approximation to train the soft prompts.

[1] Ravi et al. Optimization as a model for few-shot learning. 2017
[2] Finn et al. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. 2017

LMM Architecture with Attention-Mapper

- Our LMM architecture is inspired from LLaVA v1.5 [1], due to its simplicity and ease of use.

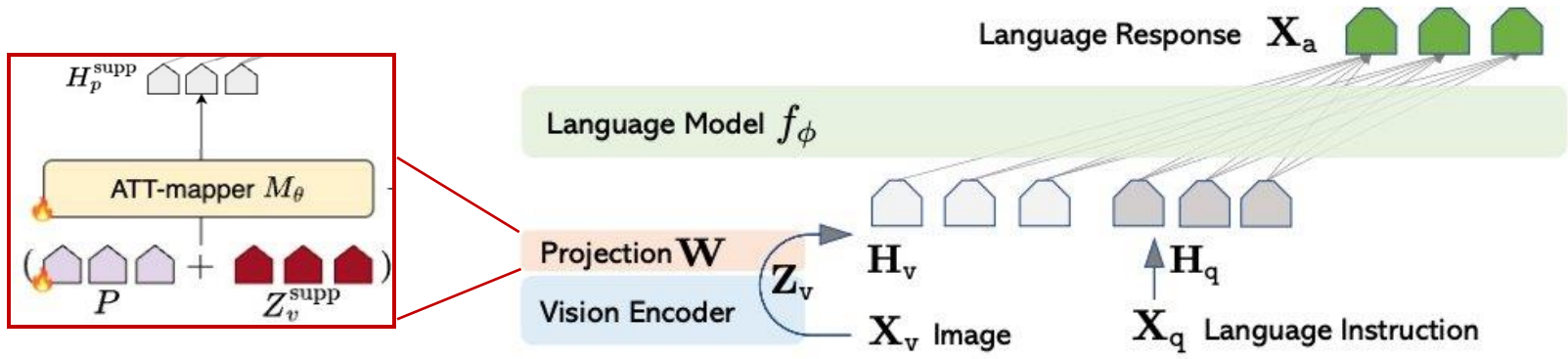


Figure 1: LLaVA network architecture.

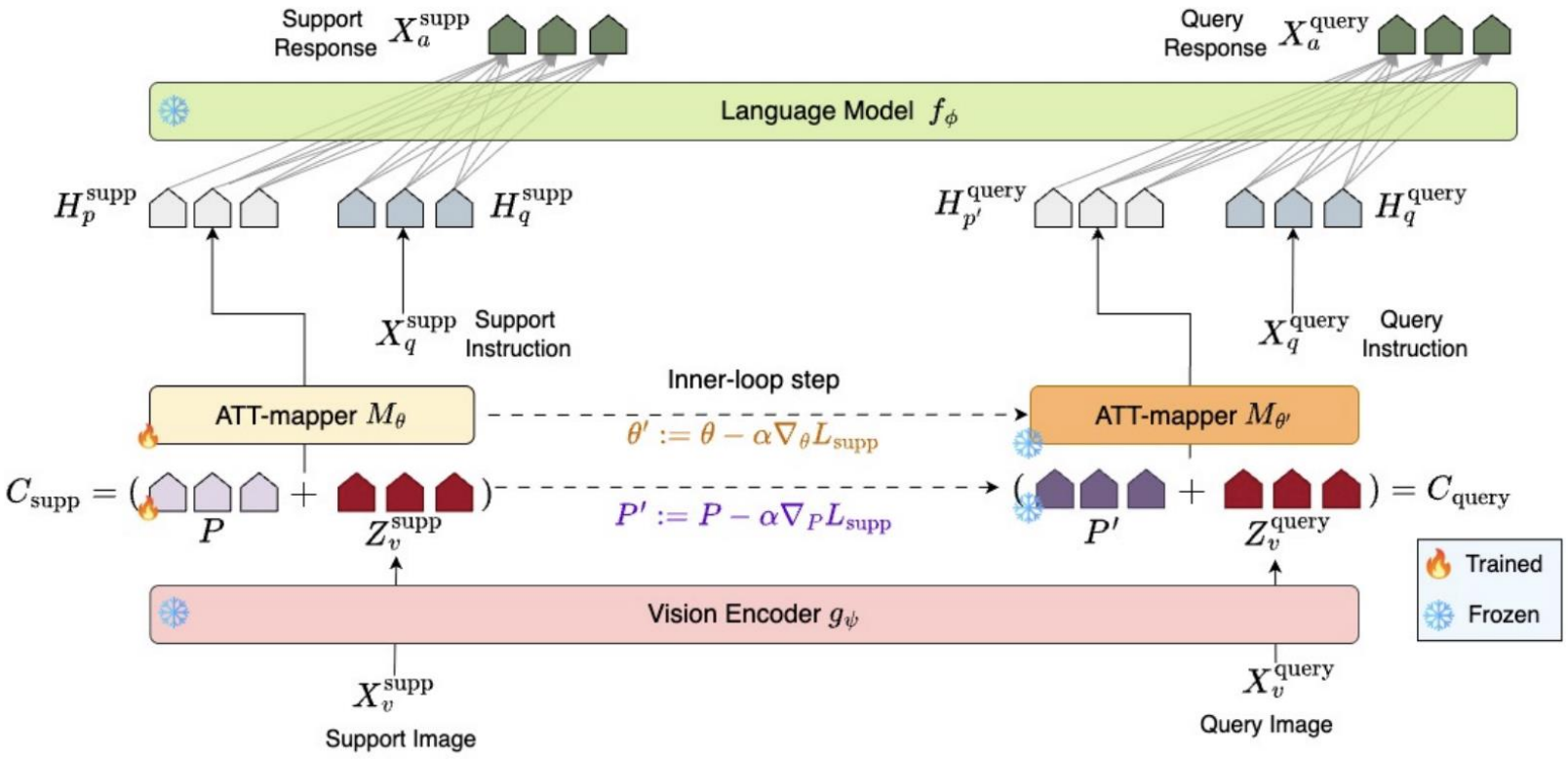
- To enable efficient prompt distillation, we introduce an attention-mapper (ATT-mapper) module in the projection layer of the LMM.

- ATT-mapper is a single multi-head attention block that takes the input image embeddings (Z) and soft prompts (P) and outputs distilled image features (H_p).

[1] Liu et al. Improved Baselines with Visual Instruction Tuning. 2024

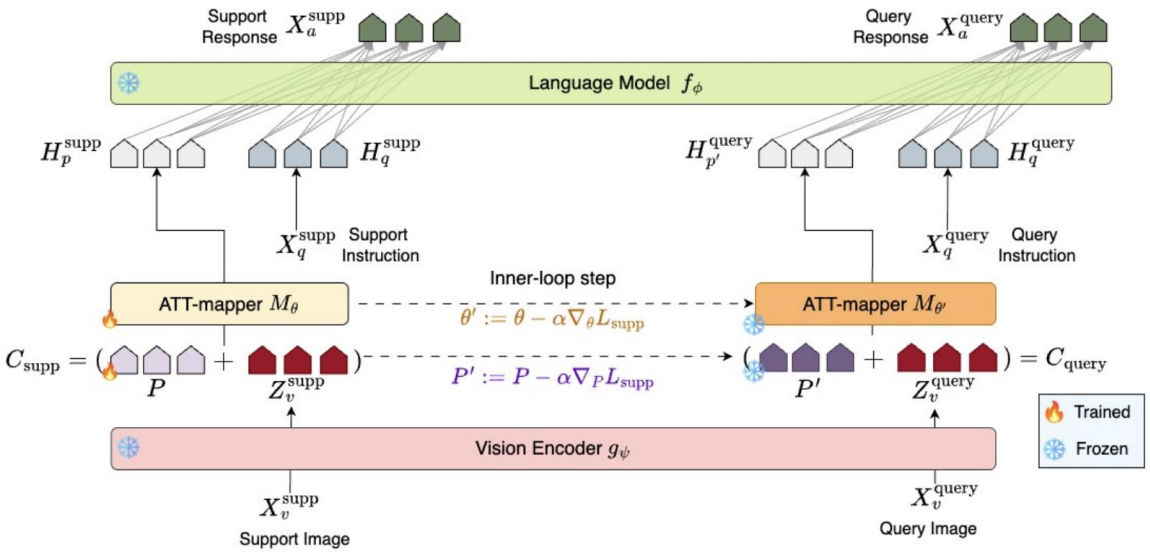
Meta-Adaptive Prompt Distillation

- We call our approach **Meta-Adaptive Prompt Distillation** (MAPD), which provides an alternative to ICL for LMMs for few-shot adaptation.



Meta-Adaptive Prompt Distillation

- MAPD meta-learns a fixed set of soft prompts with the help of an attention-mapper module over the few-shot examples and can easily be finetuned for any new task with a few number of gradient updates.
- MAPD learns task-specific information from the meta-tasks that optimize support set in the **inner-loop** and query sets in the **outer-loop**
- Note, that we only finetune the attention-mapper and soft prompts (<= 24M parameters) for any task which makes our approach highly parameter efficient.



Training and Evaluation

- We train MAPD over the well-defined dataset mixture of LLaVA v1.5 [1] and evaluate our approach on the VL-ICL benchmark which test abilties like fast concept binding, multimodal reasoning, and fine-grained perception.

- Specifically, we only test on the single-image VQA tasks from VL-ICL in this work – Fast Open-Ended Mini-ImageNet (OPEN_MI), Operator Induction (OP_IND), CLEVR Count Induction and TextOCR.

- For a fair comparison, we propose other baselines based on multi-task learning (Multi-Task^{PD}), In-context tuning [2] (In-Context^{PD}), Model Averaging [3] (Model-Avg^{PD}) and original LLaVA v1.5 finetuning with no meta-tasks (NoMetaTask^{PD})

[1] Liu et al. Improved Baselines with Visual Instruction Tuning. 2024

[2] Chen et al. Meta-learning via language model in-context tuning. 2022

[3] Choshen et al. Fusing finetuned models for better pretraining. 2022

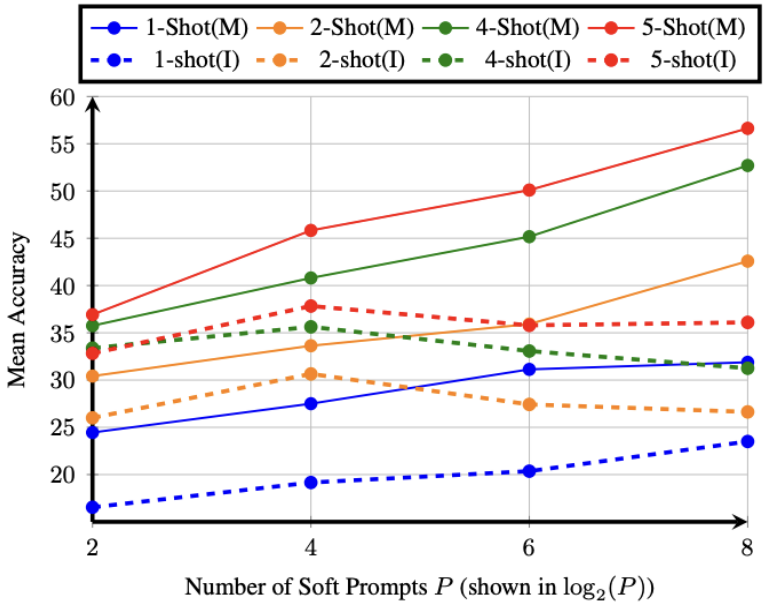
Results

- Prompt distillation improves task induction in LMMs at test-time and finetuning-based adaptation surpasses ICL by **21.2%.**
- Methods trained with meta-tasks are indeed superior as it allows training on a batch with a more balanced task distribution.
- Meta-learning outperforms all the other parameter initializations and even surpasses the next best baseline – Multi-Task^{PD} by **10%** on Operator Induction.

Methods	MT	Open-MI	OP_IND	CLEVR	TextOCR
TTA with ICL					
NoMeta-task^{PD}	✗	43.8 ± 0.9	12.1 ± 0.6	18.0 ± 0.2	6.8 ± 0.4
Model-Avg^{PD}	✗	26.6 ± 0.7	9.2 ± 0.5	7.6 ± 0.1	2.8 ± 0.3
In-Context^{PD}	✓	51.1 ± 0.9	20.6 ± 0.8	24.1 ± 0.2	23.8 ± 0.3
Multi-Task^{PD}	✓	48.6 ± 0.9	10.0 ± 0.6	12.5 ± 0.1	6.9 ± 0.4
MAPD	✓	53.3 ± 0.9	9.60 ± 0.5	12.3 ± 0.1	7.30 ± 0.4
TTA with FT $\leq$ 30					
NoMeta-task^{PD}	✗	68.0 ± 0.8	38.8 ± 0.6	25.8 ± 0.2	22.5 ± 0.3
Model-Avg^{PD}	✗	63.1 ± 0.8	40.0 ± 0.6	29.1 ± 0.2	21.5 ± 0.3
In-Context^{PD}	✓	64.5 ± 0.8	30.9 ± 0.5	30.9 ± 0.2	18.9 ± 0.3
Multi-Task^{PD}	✓	74.6 ± 0.7	45.1 ± 0.5	29.9 ± 0.2	22.9 ± 0.4
MAPD	✓	$\mathbf{77.9 \pm 0.7}$	$\mathbf{47.7 \pm 0.5}$	$\mathbf{31.4 \pm 0.2}$	$\mathbf{26.4 \pm 0.5}$

TTA:Test-Time Adaptation, FT: Finetuning with K ≤ 30 gradient steps, ICL: In-Context Learning, MT: Meta-Tasks used (✓) or not (✗) during training

Ablation studies



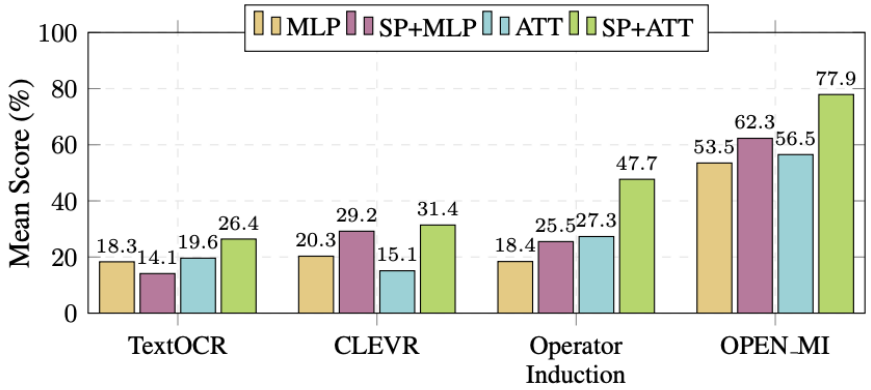
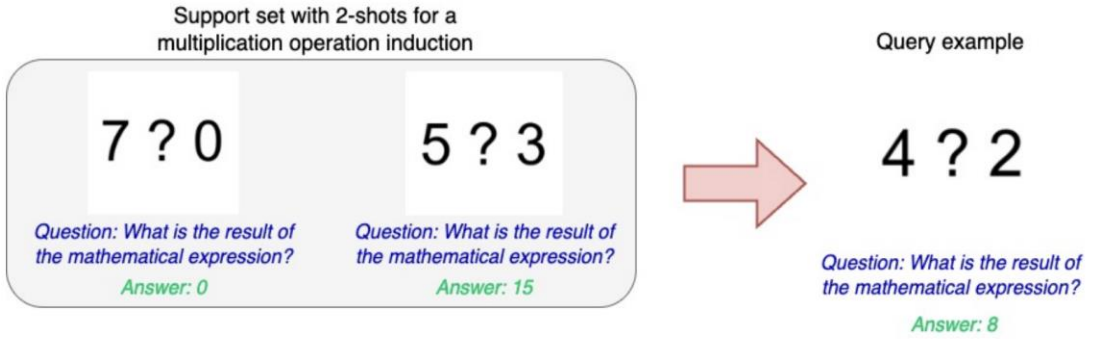


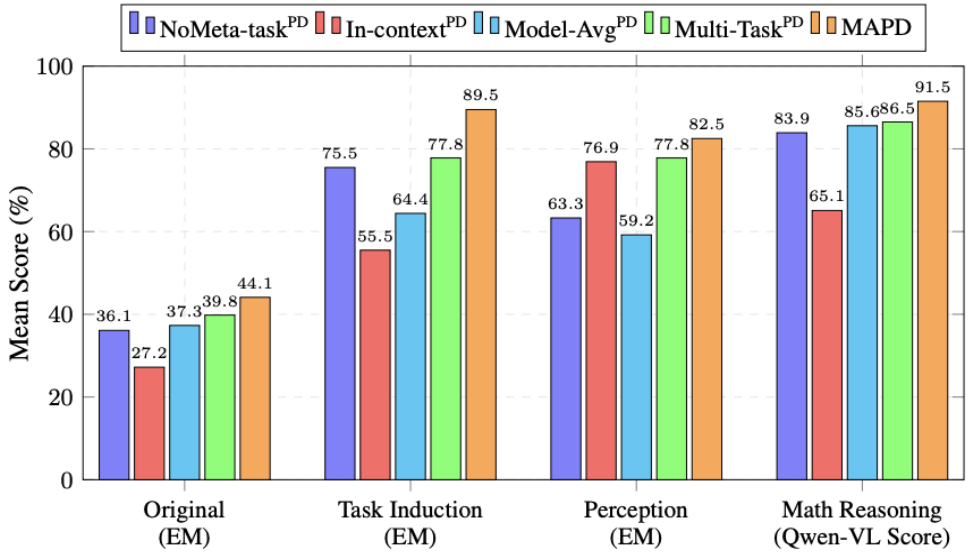
Figure 5: Projection layer architectures in the base LMM. SP: Soft Prompts, ATT: Attention-Mapper, MLP: 2-layer MLP (originally used in LLaVA v1.5).

- We first compare MAPD (**M**) with our best ICL baseline – In-Context^{PD} (**I**) and find that MAPD scales better as shots are increased with greater improvement for each added soft prompt token.
- On the right, we show that distilling task-relevant information from image embeddings into soft prompts yields an improvement of **5.2%** and attention-mapper further enhances this by **13.1%.**

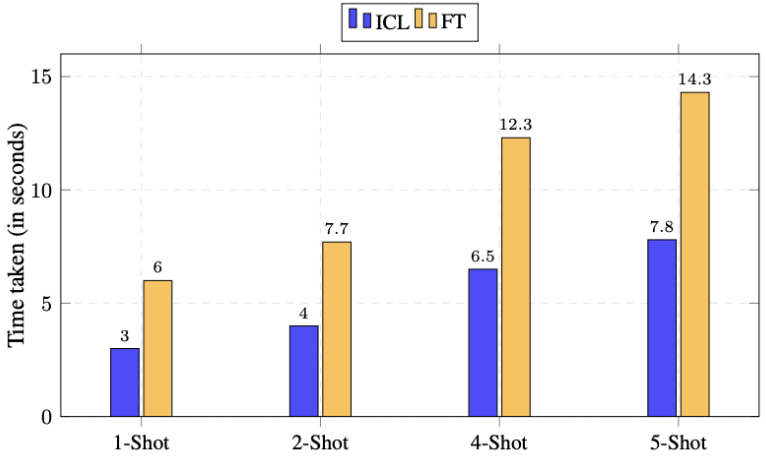
Additional analysis on Operator Induction

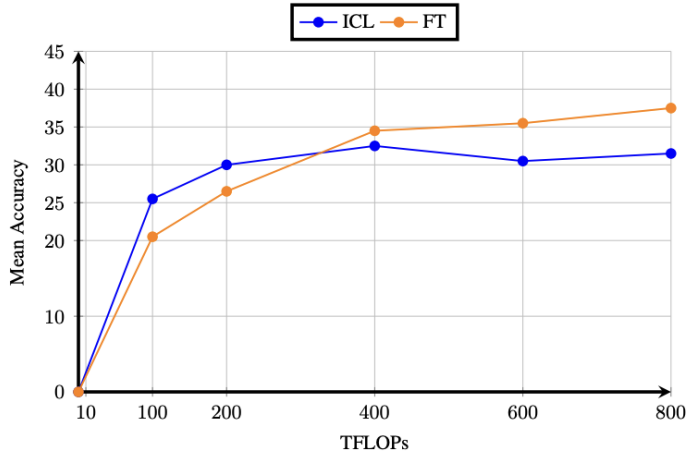
- We evaluate the task understanding of MAPD at test time by defining sub-tasks -
  - Perception - identifying the operands in the query example
  - Task Induction - identifying the operation from few-shot examples.
  - Mathematical reasoning – using the LMM's mathematical knowledge over the identified elements to reason towards the answer.
- MAPD outperforms all the other baselines on all 3 subtasks and shows a major improvement on task induction with an increase of **11.7%** compared to Multi-Task^{PD}





Limitations and Future Work





- We note that state-of-the-art performance of MAPD comes at a test-time computational cost where MAPD being dependent on finetuning-based (FT) adaptation takes twice as much inference time than ICL, due to gradient computation (**left figure**).
- But we also note that FT-based adaptation is more data-efficient (**right figure**): at 400 TFLOPs, the ICL performance of In-Context^{PD} begins to decline after 32 shots whereas MAPD with FT achieves comparable performance with 8 shots and 20 gradient steps
- Future work could focus on improving MAPD's computational efficiency for resource constrained scenarios, multi-image tasks and more complex reasoning problems.

Main Takeways

- ICL performance in LMMs can be inconsistent as the number of shots are increased due to sudden increase in the image embeddings in the context.
- Distilling task-relevant information from image features and prompting the underlying LLM improves performance over multimodal ICL benchmarks.
- Architectural modifications like the addition of soft prompts and substituting the MLP projection layer with a multi-head attention block further facilitates this distillation.
- Meta-learning serves as the best method to initialize the parameters of the soft prompts and attention mapper and allows for rapid finetuning over new few-shot VQA tasks.
- Finetuning-based test time adaptation is data-efficient but also requires more computational resources for gradient computation.