When Parameter-efficient Tuning Meets General-purpose Vision-language Models

By: Yihang Zhai et al.

Presenter:

Hari Umesh

Example of Mutimodal Q/A

Instruction:

- What objects are in the picture?
- What is the relationship between the objects in the picture?
- What is the relationship between the objects in the picture ?

Ground Truth

A man wearing a red t-shirt sweeps the sidewalk in front of a brick building

Full finetune:

A man in a red shirt is sweeping the sidewalk

LoRA:

A man in a red shirt is sweeping the sidewalk

PETAL:

A man in a red shirt is sweeping the sidewalk in front of a brick building

Outline

- Introduction
- PETAL Architecture
- Evaluation
- Summary

Outline

- Introduction
- PETAL Architecture
- Evaluation
- Summary

Larger Models require larger compute

- As model size and accuracy increases, the demand for the amount of compute also increases
- Training Large models from scratch are expensive

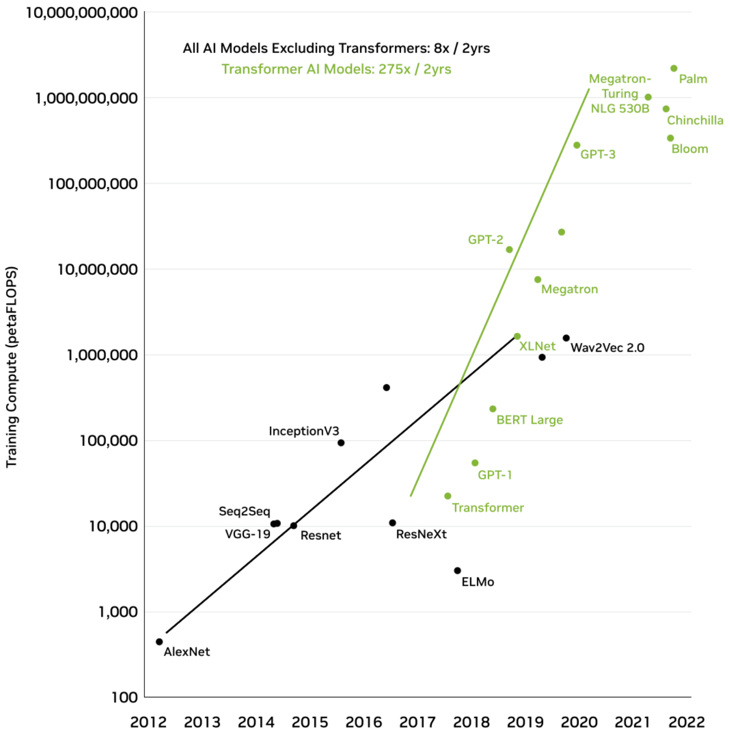


Figure 3. Compute required for training transformer models.

Instruction Tuning

- Training LLMs based on instructions
- Allows models adaptable to a wide-range of tasks without task-specific training

Instruction:
What fruit is typically added to the top of cereal?

Answer:
“banana”, “banana”, “banana”, “banana”, “banana”, “banana”, “strawberry”, “strawberry”, “blueberry”, “blueberry”

Overview of Existing PET Multimodal Tasks

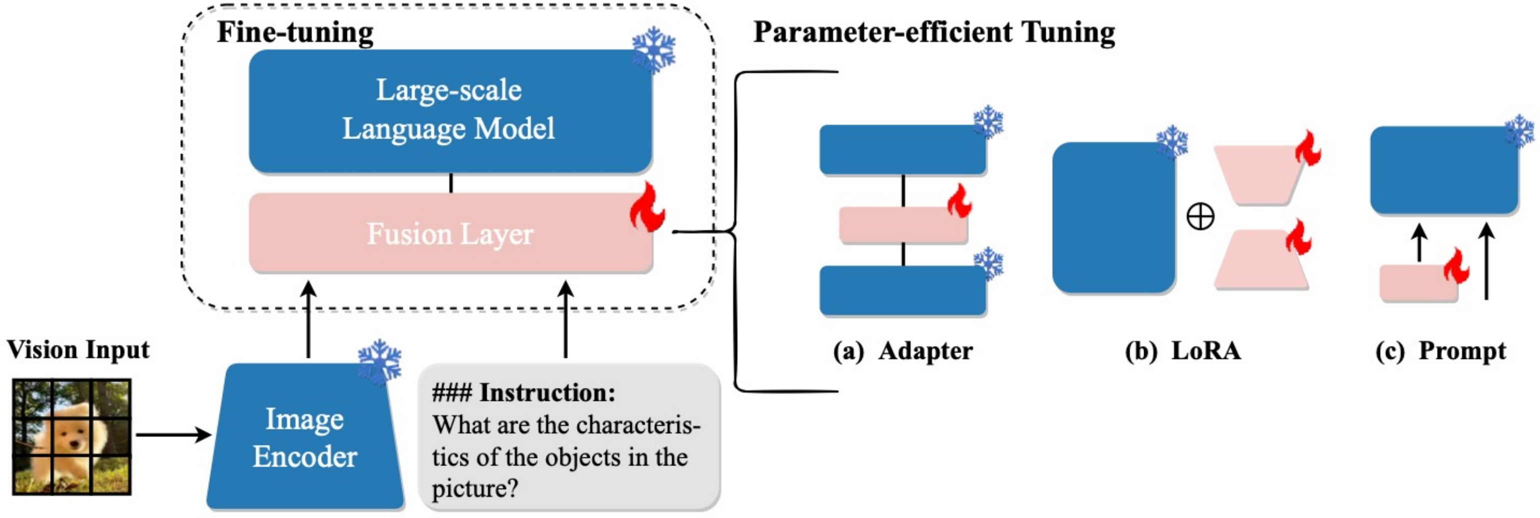
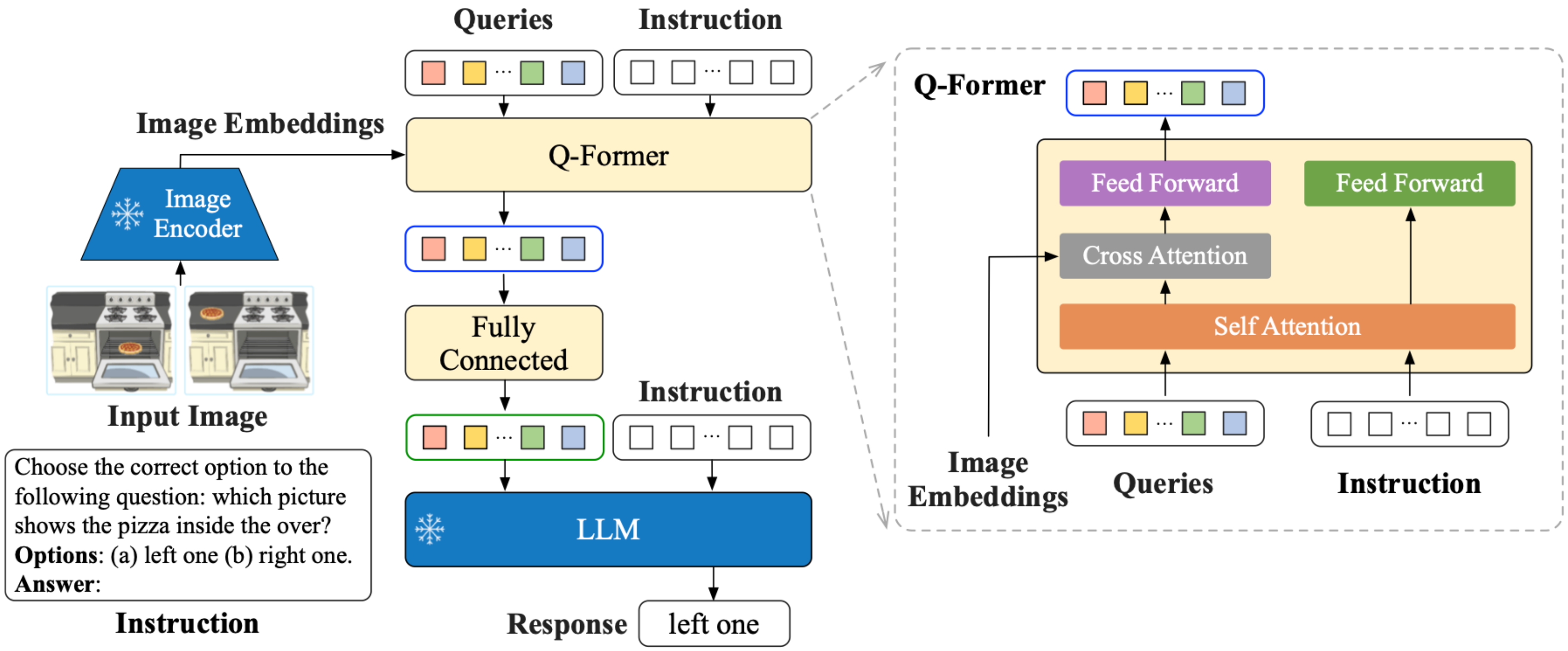


Figure 1. Overview of existing parameter-efficient tuning methods applied in multimodal tasks.

InstructBLIP



Multimodal Instruction Challenges

1. Finetuning full models is expensive

2. Lack in semantic information in instructions, which hinders multimodal alignment

Outline

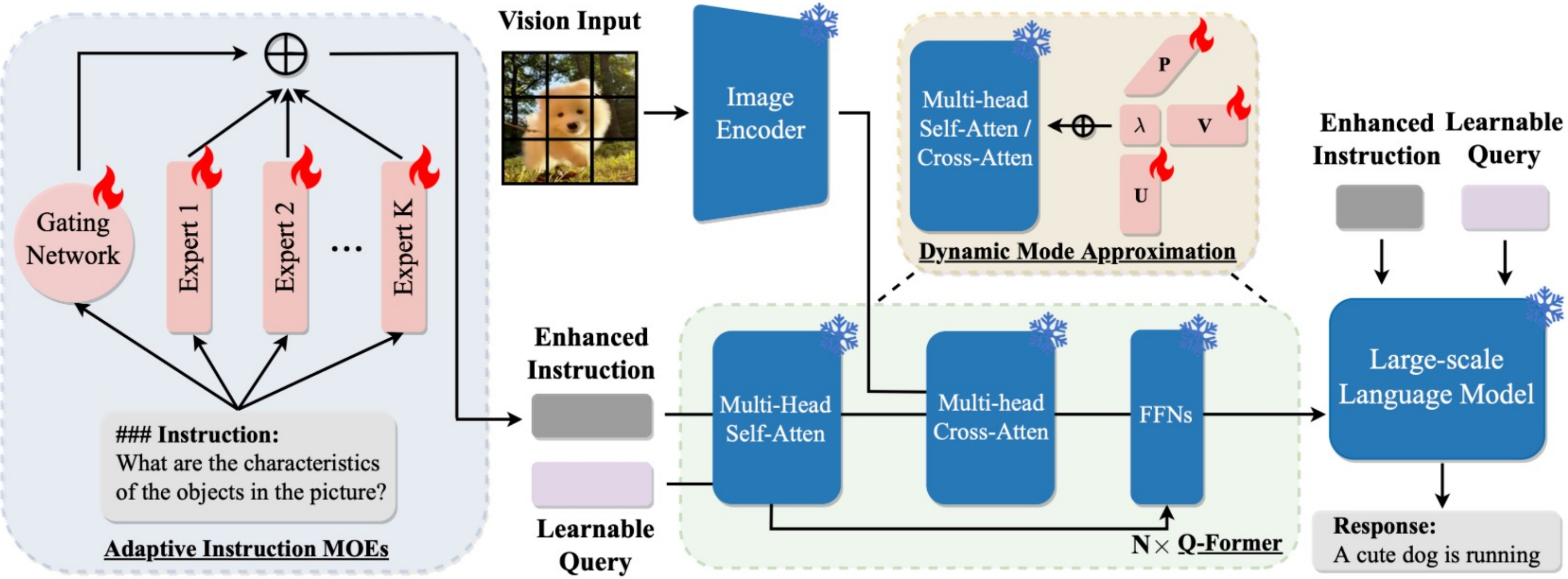
- Introduction
- PETAL Architecture
- Evaluation
- Summary

PETAL Main Contributions

1. Novel **Dynamic Mode Approximation** for efficient tuning

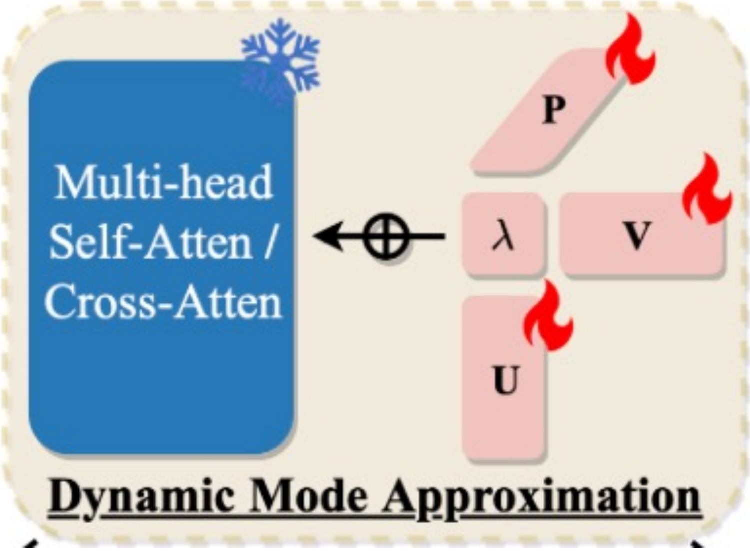
2. Enhanced Instruction Semantics
   - **Adaptive Instruction MOEs module**
   - **Score-based information bottleneck**

Model Overview Diagram



Dynamic Mode Approximation for Efficient Tuning

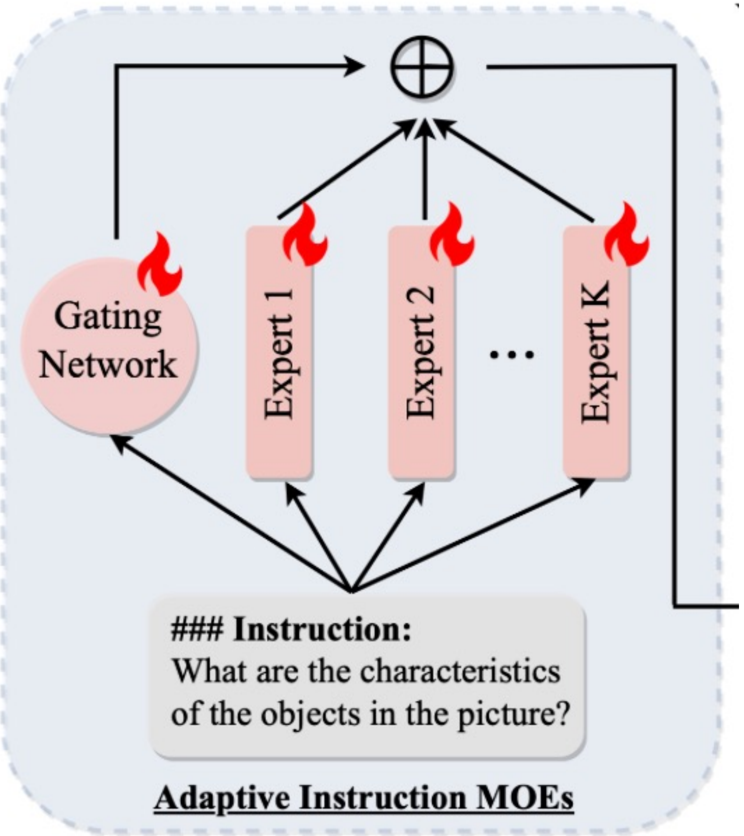
- Approximates the attention weights in Transformer architecture based on CP decomposition with a dynamic weighting scheme



$$\mathbf{H}^m = \Gamma \mathbf{W}_0 \mathbf{X}^m + \left(\sum_{r=1}^{R} \lambda_r^m (\mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{p}_r) \right) \mathbf{X}^m$$

Adaptive Instruction MOEs Module

- Extracts information from multiple perspectives by setting multiple instructions for each image with a different focus then stack them in a text paragraph
- Then features are extracted and merged then



Score-based Information bottleneck loss

- Mutual information(MI) loss that enhances semantics of instructions
- Calculate loss based on the product of normalized features and instructions
- Maximizes the mutual information between the representation and the target and minimizes between the representation and the input

$$\max \mathbf{MI} = \max I(Z;Y) - \eta I(Z;X)$$

$$\mathcal{L}_{IB} = \mathbf{MI}(\hat{\mathbf{z}}; y)$$

Outline

- Introduction
- PETAL Architecture
- Evaluation
- Summary

Tasks, Datasets, and Baselines

Tasks: Image Captioning and Question Answering

Datasets

- Captioning: Flickr30K, TextCaps
- Q/A: OKVQA, A-OKVQA, TextVQA

Baselines: InstructBLIP and LLaVA on 4 PEFT METHODS

- HEAD TUNING
- MAPLE
- LLAMA-ADAPTER
- LORA

Implementation?

- Apply PET exclusively to the Q-Former enhanced by approximation techniques
- LLMs Used: FlanT5 and Vicuna-7B
- GPU: 5 epochs 8x A100 (80GB) GPU

Results: Image Captioning

Table 1. We select two captioning datasets, Flickr30K and TextCaps, for performance comparison on the famous image captioning benchmark. We report CIDEr score, ROUGE-1 F1 score, ROUGE-1 recall, ROUGE-L F1 score, and ROUGE-1 recall for both of them, where ↑ and ↓ respectively indicates how much our method has improved or declined compared to the best parameter-efficient baseline.

Method	Tunable	Flickr30K					TextCaps				
		CIDEr	ROGUE-1		ROGUE-L		CIDEr	ROGUE-1		ROGUE-L	
			F1	Recall	F1	Recall		F1	Recall	F1	Recall
Fine-tune InstructBLIP ($FlanT5_{xxl}$)	188 M	**63.5**	34.9	33.4	31.6	30.4	46.6	28.2	26.3	24.7	23.0
Head ($FlanT5_{xxl}$)	11.8 M	60.8	34.5	31.9	30.7	29.2	43.6	28.9	**28.8**	24.6	24.5
MAPLE ($FlanT5_{xxl}$)	2.9M	59.4	34.2	31.8	30.8	28.1	43.7	27.8	26.5	24.3	22.4
LLaMA-Adapter ($FlanT5_{xxl}$,R=128)	4.8 M	60.5	33.7	31.2	30.5	28.3	44.5	28.3	23.9	24.5	23.1
LoRA ($FlanT5_{xxl}$,R=64)	5.0 M	59.8	34.0	31.5	30.8	28.6	45.4	28.0	25.5	24.7	22.5
PETAL ($FlanT5_{xxl}$,R=64)	**1 M**	63.3	**35.2**	**32.9**	**35.2**	**32.9**	**46.7**	**29.1**	27.8	**29.1**	**28.5**
		2.5↑	0.7↑	1.0↑	4.2↑	3.7↑	1.3↑	0.2↑	1.0↓	4.4↑	4.0↑
Fine-tune InstructBLIP (Vicuna-7B)	188 M	**65.8**	35.5	34.3	**36.0**	34.3	**48.9**	30.8	**31.5**	30.9	**31.5**
LLaVA (Vicuna-7B)	7B	64.6	**35.9**	33.7	34.2	34.0	48.5	30.9	27.9	28.7	-
Head (Vicuna-7B)	11.8 M	61.9	35.1	32.9	33.8	33.7	48.3	30.3	30.6	30.3	30.6
MAPLE (Vicuna-7B)	2.9M	61.3	35.6	33.0	35.1	32.4	48.1	30.5	30.8	30.7	30.1
LLaMA-Adapter (Vicuna-7B,R=128)	4.8 M	62.4	36.6	33.2	33.4	32.9	48.3	30.2	30.0	30.4	30.2
LoRA (Vicuna-7B,R=64)	5.0 M	62.5	35.0	33.8	35.0	33.9	48.5	30.8	31.2	30.8	31.2
PETAL (Vicuna-7B,R=64)	**1 M**	63.6	35.7	**34.3**	36.2	**34.3**	**48.8**	**31.1**	30.5	**32.0**	31.3
		1.1↑	0.4↓	0.5↑	0.6↑	0.4↑	0.3↑	0.3↑	0.7↓	1.2↑	0.1↑

Results: Question Answering

Table 2. Performance comparison on VQA benchmarks. We conduct experiments on three datasets: A-OKVQA, OKVQA, and TextVQA, using accuracy as the evaluation metric. ↑ and ↓ respectively indicates our improvement compared to the best baseline.

Method	Tunable	A-OKVQA	TextVQA	OKVQA
Based on FlanT5$_{xxl}$				
Fine-tune	188 M	**56.7**	**24.1**	**55.2**
Head	11.8 M	54.3	21.0	52.1
MAPLE	2.9M	54.1	21.2	52.4
LLaMA-Adapter	4.8 M	53.2	20.9	52.8
LoRA	5.0 M	54.5	21.4	53.4
PETAL	1.0 M	55.8	21.5	53.6
		↑1.3	↑0.1	↑0.2
Based on Vicuna-7B				
Fine-tune	188 M	63.6	**62.4**	27.7
LLaVA	7B	52.2	52.7	21.3
Head	11.8 M	63.2	60.1	25.8
MAPLE	2.9M	62.8	60.3	24.5
LLaMA-Adapter	4.8 M	63.0	60.5	26.4
LoRA	5.0 M	63.5	60.8	26.3
PETAL	1.0 M	**63.8**	61.8	**27.7**
		↑ 0.3	↑ 1.0	↑ 1.3

PETAL captures greater object information

Instruction:

- What objects are in the picture?
- What is the relationship between the objects in the picture?
- What is the relationship between the objects in the picture ?

Ground Truth

A man wearing a red t-shirt sweeps the sidewalk in front of a brick building

Full finetune:

A man in a red shirt is sweeping the sidewalk

LoRA:

A man in a red shirt is sweeping the sidewalk

PETAL:

A man in a red shirt is sweeping the sidewalk in front of a brick building

PETAL captures the relationship between objects

Instruction:
- What objects are in the picture?
- What is the relationship between the objects in the picture?
- What is the relationship between the objects in the picture ?

Ground Truth:
On a sunny, dry day, wearing full football gear, a Texas A&M football player tries to reach an Iowa State football player, for the football during the game

Full finetune:
The football player is wearing a red jersey and a yellow helmet

LoRA:
A football player is running with the ball

PETAL:
A football player is being tackled by another player

PETAL boasts SOTA Few-shot results

Table 3. Results of few-shot instruction tuning. We have two configurations: we extract 50/150 data items from the training set for training. We conducted tests on AOKVQA, OKVQA, FLickr30K, and TextCaps. For AOKVQA and OKVQA, we calculate the accuracy, while for Flickr30K and TextCaps, we compute the CIDEr Score and ROGUE-1 F1 Score.

Method	Parameter	A-OKVQA	OKVQA		Flickr30K		TextCaps		
		Accuracy	Accuracy	Avg	CIDEr	ROGUE-1 F1	CIDEr	ROGUE-1 F1	Avg
50-shot									
Fine-tuning	188 M	53.2	52.0	52.6	**53.5**	33.1	42.8	27.1	39.1
LoRA	5 M	53.1	52.1	52.6	52.7	33.1	43.0	27.1	38.8
Ours	1 M	**53.3**	**52.6**	**53.0**	52.9	**33.2**	**43.1**	**27.4**	**39.2**
150-shot									
Fine-tuning	188 M	53.0	52.2	52.6	**53.5**	33.2	42.7	27.1	39.1
LoRA	5 M	53.1	52.1	52.6	52.7	33.1	42.9	27.0	38.9
Ours	1 M	**53.1**	**52.6**	**52.9**	53.1	**33.2**	**43.2**	**27.4**	**39.2**

Visualization Comparison of Instruction Enhancement

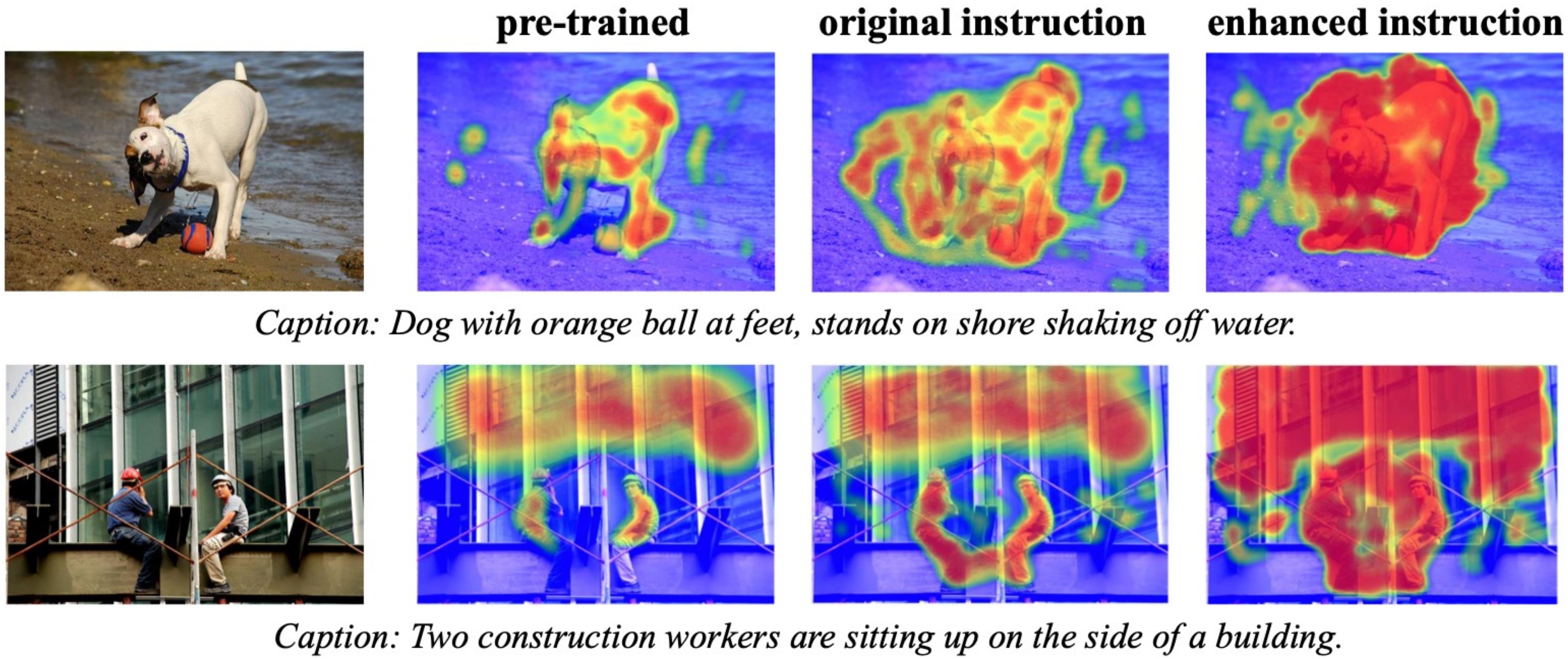


Figure 4. Cross-attention visualization results on Flickr30k dataset.

Training Time and Parameter Size Comparison

Table 5. Training time and parameter size comparison.

Method	#Tunable	Flickr30K	TextCaps	AOKVQA	OKVQA	TextVQA
Fine-tune	188M	1.00	1.00	1.00	1.00	1.00
MAPLE	4.8M	0.95	0.93	0.94	0.90	0.91
LoRA	5.0M	0.91	0.92	0.86	0.87	0.93
PETAL	1M	0.85	0.88	0.86	0.85	0.91

Outline

- Introduction
- PETAL Architecture
- Evaluation
- Summary

Summary of Key Contributions

- PETAL: novel approach for parameter efficient tuning in vision-language models
- Dynamic mode approximation increases efficiency
- Enhanced Instructions through **adaptive instruction MOEs** and **mutual information loss**

Discussion and Future Contributions

Strengths

- Multimodal paper with specific instruction tuning optimizations and dynamic mode approximation

Future Work

- Quantization
- Inference level optimizations for Multimodal inputs
- Switch Transformer optimization w/ MOEs

Appendix 1: Ablation Study PETAL Architecture

Table 4. Results of ablation studies that remove the important components. ↑ and ↓ respectively indicates how much the variant has improved or declined compared to our PETAL. DMA stands for Dynamic Mode Approximation, AIM represents Adaptive Instruction MOEs, and SIB is Score-based Information Bottleneck loss.

Method	DMA	AIM	SIB	A-OKVQA	Flickr30K			Avg.
				Accuracy	CIDEr	ROGUE-L		
						F1	Recall	
PETAL V1	✓	×	✓	55.2	61.1	34.5	32.3	45.8 (↓1.1)
PETAL V2	✓	✓	×	55.6	62.6	34.8	33.1	46.5 (↓0.4)
PETAL V3	✓	×	×	54.8	60.7	31.0	28.9	43.9 (↓3.0)
PETAL V4	×	✓	✓	55.4	61.2	33.8	31.9	45.6 (↓1.3)
PETAL w. random instruction				55.1	60.8	32.0	30.9	44.8 (↓2.2)
PETAL	✓	✓	✓	**55.8**	**63.4**	**35.1**	**33.4**	**46.9**

Appendix 1: Ablation Study PETAL Architecture

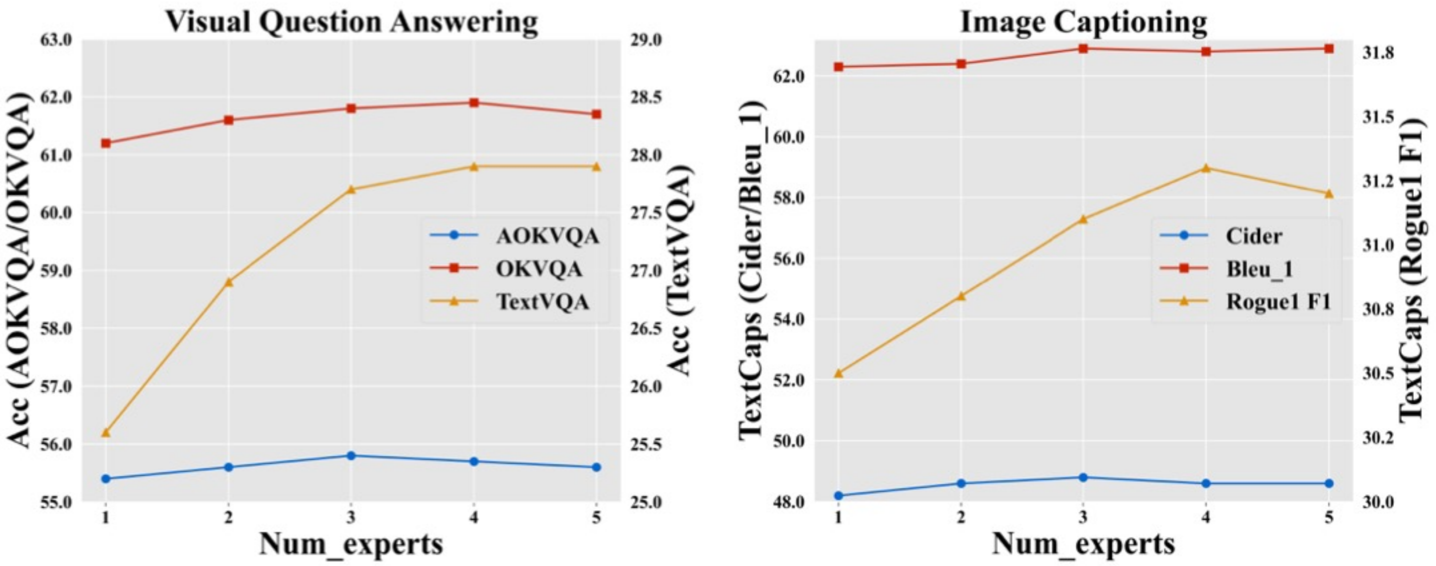


Figure 3. Results of different number of experts.