Multimodal Task Vectors Enable Many-Shot Multimodal In-Context Learning

Brandon Huang* Chancharik Mitra* Assafe Arbelle Leonid Karlinsky
Trevor Darrell Roi Herzig

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

LMMs Struggle with Context Length

- LMMs typically have limited context length
  - E.g 8K for Idefics2
- Images are token expensive! This makes ICL for MLLM even more expensive.

	%Text Tokens	%Image Token
Vizwiz	6.6	93.4
OKVQA	8.4	91.6
Flower	22.7	77.3
CUB	24	76

Task Vectors Enables Efficient ICL for LLMs

- Recent interpretability works demonstrates the existence of Task Vector in LLM and VIT
  - ICL synthesize task representation that can be extracted and reused at inference time
- We show the existence of Task Vector in LMMs, which we call **MTV**
- It enables time and memory efficient ICL for MLLM compared to vanilla ICL
- We observe a scaling law with the no. of examples using MTV

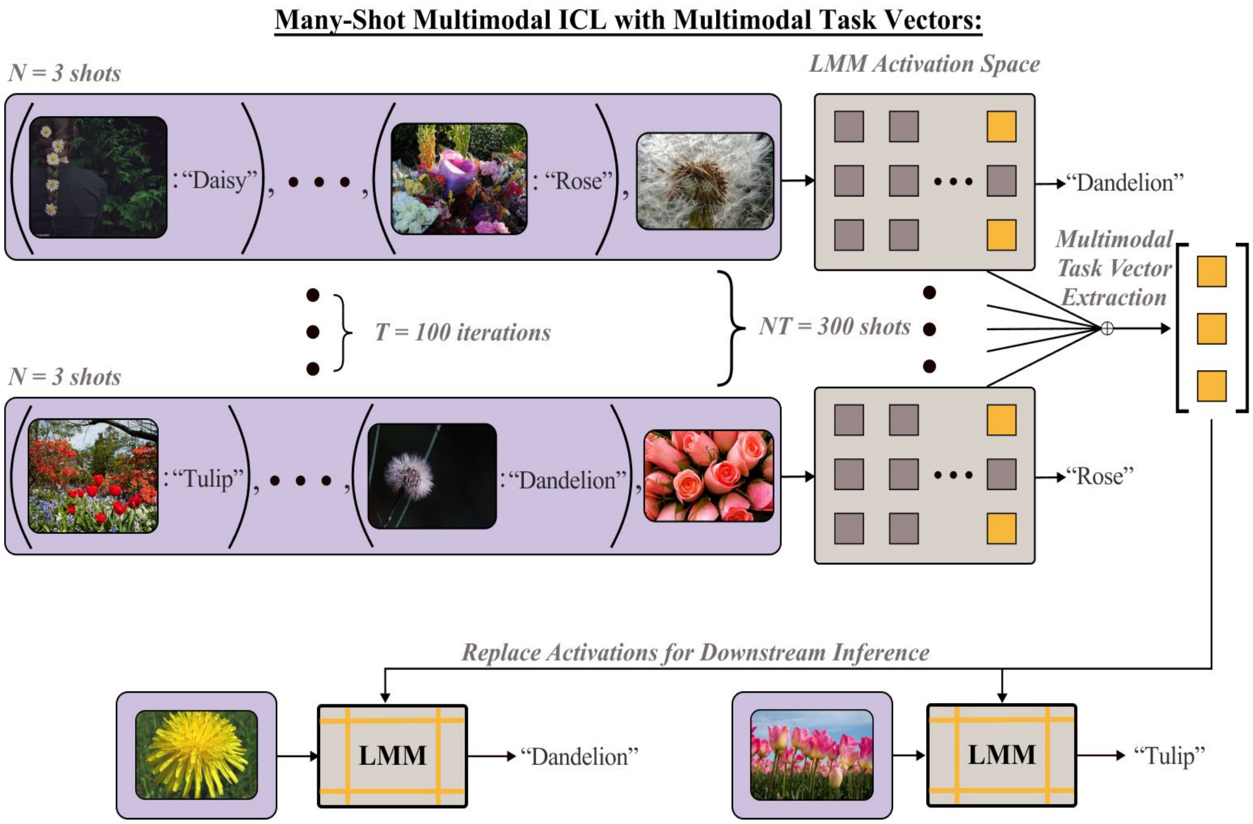
Calculate MTV

1. Get mean activations corresponding to the last token of the input.

2. Locate attention head locations that captures the task using REINFORCE. All model weights frozen.

MTV at Inference Time

- Replace the current activation with the mean activation at the selected heads.



Experiments

Models: QwenVL, Idefics2-8B, Vila-1.5-8B

Evaluation Datasets:

- Visual Question-Answering: VizWiz, OK-VQA
- Object Classification: Flowers, Caltech’s CUB Dataset on Birds

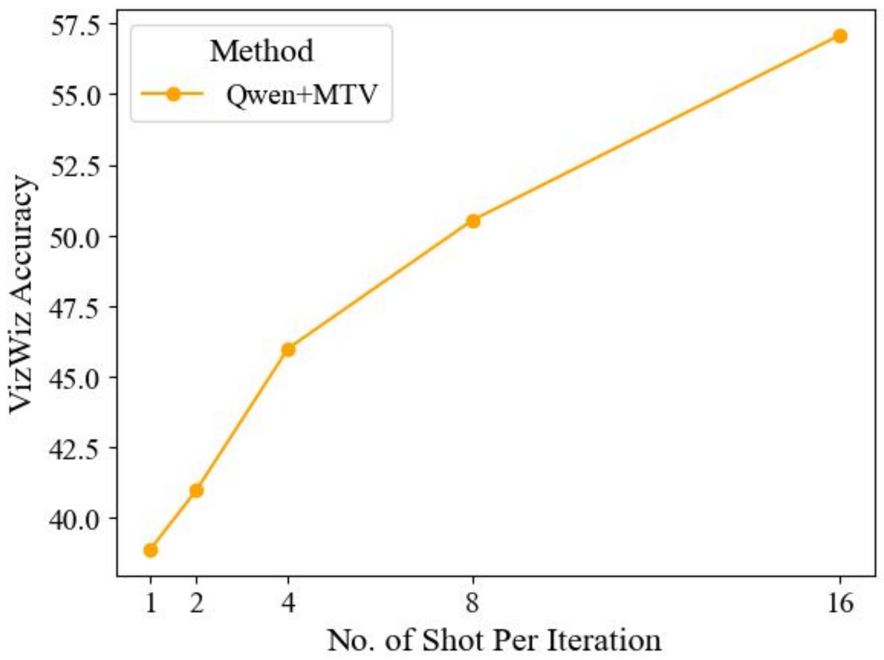
(a) **MTV on VQA Benchmarks**

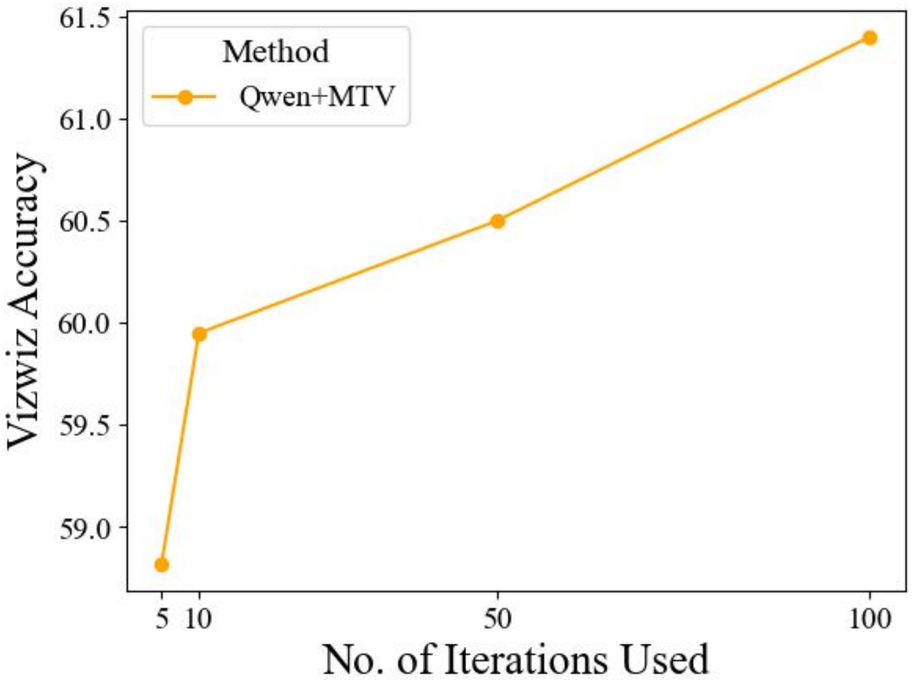
Model	VizWiz	OK-VQA
Flamingo 9B	28.8	44.7
+4-shot ICL	34.9	49.3
+8-shot ICL	39.4	50.0
Blip3	21.2	26.5
+4-shot ICL	38.4	49.2
+8-shot ICL	44.3	49.1
Qwen-VL-7B	35.2	58.6
+4-shot ICL	42.0	**62.0**
+8-shot ICL	44.3	61.5
+**MTV**	**45.6**	**62.0**
Idefics2	31.3	52.4
+4-shot ICL	40.8	51.5
+8-shot ICL	43.8	52.3
+**MTV**	**52.5**	**53.0**
Llama3-VILA-1.5-8B	28.0	32.8
+4-shot ICL	39.3	35.6
+8-shot ICL	44.2	36.5
+**MTV**	**55.2**	**40.6**

(b) **MTV on Object Classification**

Model	Flowers	CUB
LLaVA-1.5-13B		
+ 1-shot ICL	58.60	58.24
LLaVA-1.6-13B		
+ 1-shot ICL	65.58	67.90
Flamingo 9B		
+ 1-shot ICL 9B	48.78	51.2
IDEFICS-9B		
+ 1-shot ICL	55.29	62.0
Emu 37B		
+ 1-shot ICL	52.76	53.56
Qwen-VL-7B		
+ 1-shot ICL	55.0	56.5
+ **MTV**+1-shot ICL	**78.1**	**80.0**
Idefics2		
+ 1-shot ICL	82.8	88.7
+ **MTV**+1-shot ICL	**83.8**	**89.8**
Llama3-VILA-1.5-8B		
+ 1-shot ICL	87.4	88.4
+ **MTV**+1-shot ICL	**89.3**	**89.7**

Scaling Law for MTV





Multimodal Task Vectors

(a) Attention Head Generalization

Model	VizWiz	OK-VQA
ViLA-1.5-8B	28.0	32.8
+ 4-shot-ICL	39.3	35.6
+ 8-shot-ICL	44.2	36.5
+ **MTV-Vizwiz**	**55.2**	**38.3**

Metric	**0-shot**	**4-shot**	**8-shot**	**16-shot**	**MTV (400-shot)**
Max GPU Memory (GB)	17.4	18.3	19.0	20.6	19.8
Runtime per 100 iterations (min)	1.1	2.7	3.1	3.3	1.9

Text-Only Tasks

- We compare MTV with many-shot ICL using LLaMA 3.1

	English-Spanish	Antonym Generation
10-shot	65.2	56.0
400-shot	68.5	57.6
MTV	76.7	61.7

English-Spanish e.g.: hello:hola, dog:perra, apple:?

Antonym Generation e.g.:good:bad, tall:short, big:?

Conclusion

- MTVs can effectively learn tasks from many-shot multimodal ICL examples without finetuning
- They scale with additional examples
- They are generally applicable to almost any vision-language task

Thank You!