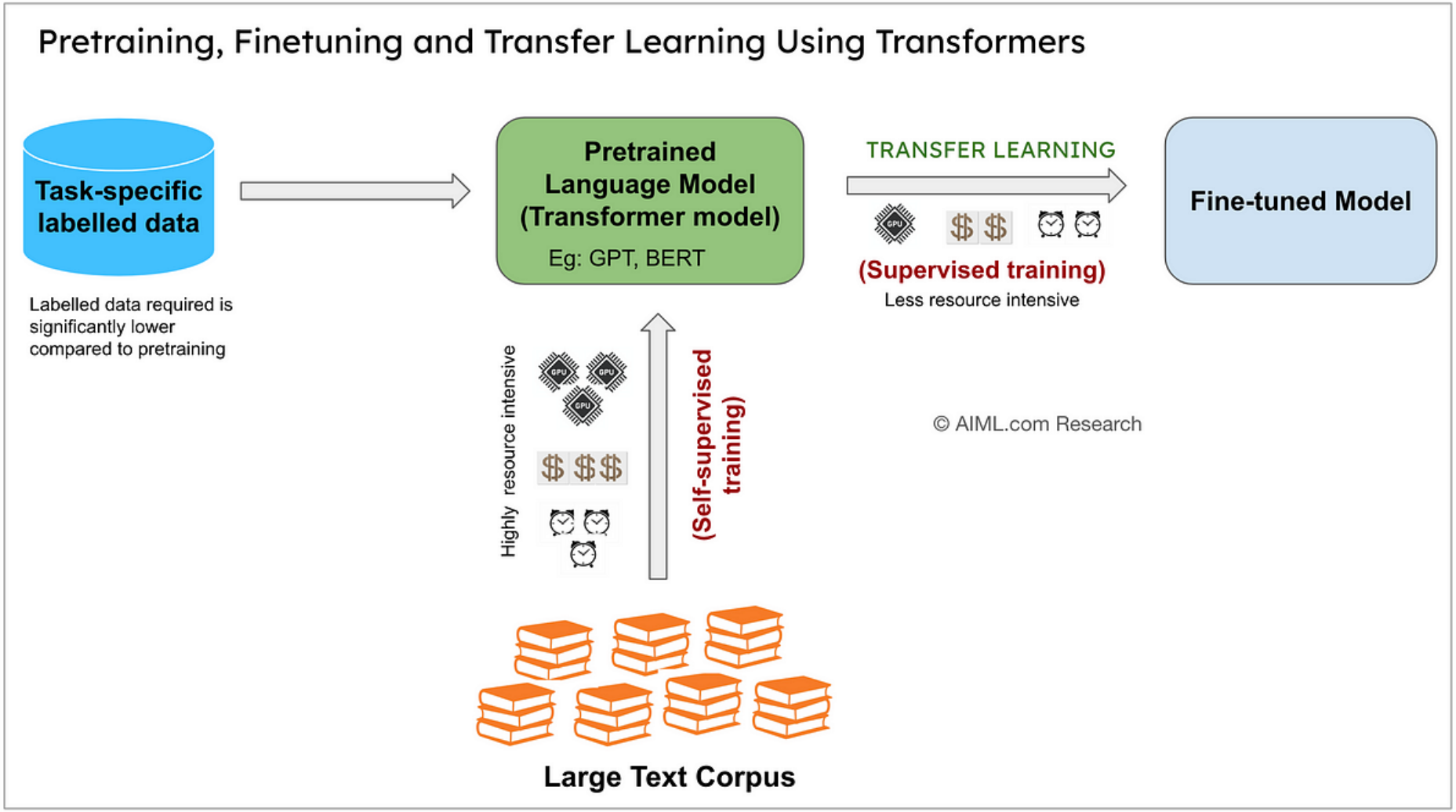
CoIN: A Benchmark of Continual Instruction Tuning for Multimodal Large Language Models

Cheng Chen, Junchen Zhu, Xu Luo, Heng Tao Shen, Jingkuan Song, Lianli Gao*

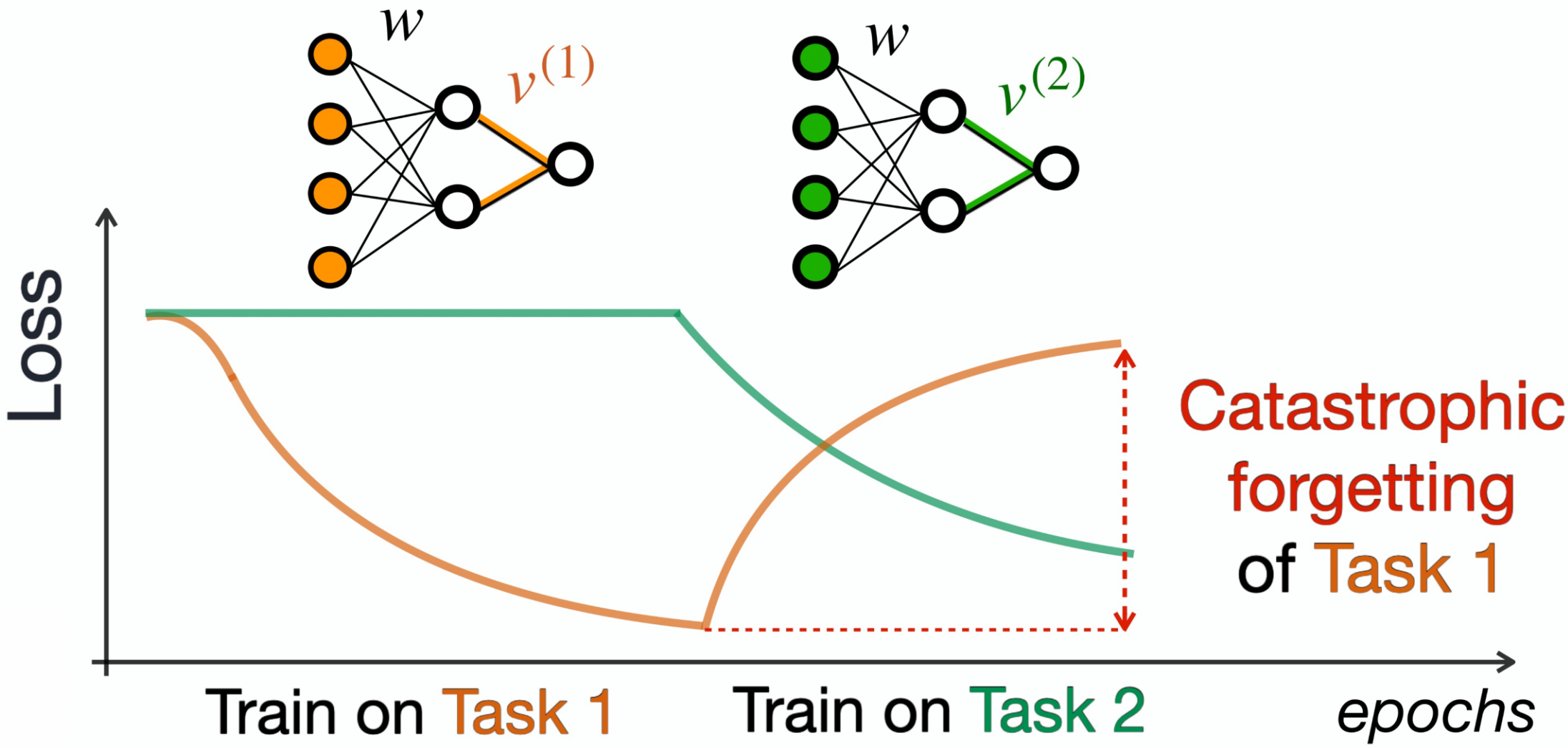
Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
University of Electronic Science and Technology of China, Tongji University

Background



Cheng Chen et al. "CoIN: A Benchmark of Continual Instruction Tuning for Multimodal Large Language Models." *NeurIPS 2024*

Problem


w
$v^{(1)}$
w
$v^{(2)}$
Loss
Catastrophic
forgetting
of Task 1
Train on Task 1
Train on Task 2
epochs

Motivation

- Investigate the catastrophic forgetting of MLLMs in instruction tuning phrase.

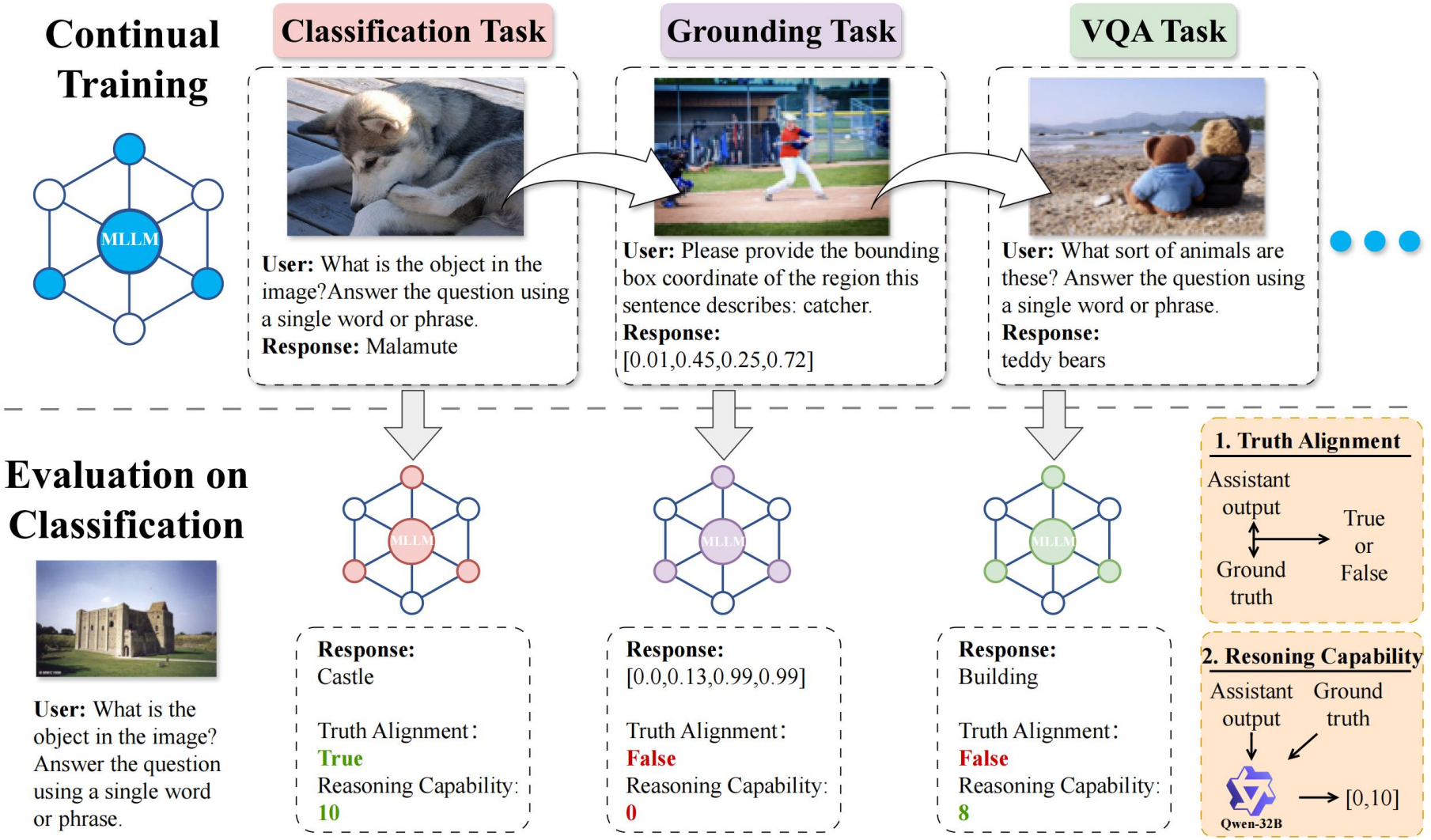
- Lacking of a suitable instruction tuning benchmark, we construct a continual instruction tuning benchmark by publicly available vision-language datasets.

CoIN

Task	Dataset	Instruction	Train Number	Test Number
Grounding	RefCOCO RefCOCO+ RefCOCOg	Please provide the bounding box coordinate of the region this sentence describes: <description>	55k	31k
Classification	ImageNet	What is the object in the image? Answer the question using a single word or phrase	129k	5k
Image Question Answering (IQA)	VQAv2	Answer the question using a single word or phrase	82k	107k
Knowledge Grounded IQA	ScienceQA	Answer with the option's letter from the given choices directly	12k	4k
Reading Comprehension IQA	TextVQA	Answer the question using a single word or phrase	34k	5k
Visual Reasoning IQA	GQA	Answer the question using a single word or phrase	72k	1k
Blind People IQA	VizWiz	Answer the question using a single word or phrase	20k	8k
OCR IQA	OCR-VQA	Answer the question using a single word or phrase	165k	100k

Cheng Chen et al. "CoIN: A Benchmark of Continual Instruction Tuning for Multimodal Large Language Models." ***NeurIPS 2024***

CoIN



Cheng Chen et al. "CoIN: A Benchmark of Continual Instruction Tuning for Multimodal Large Language Models." *NeurIPS 2024*

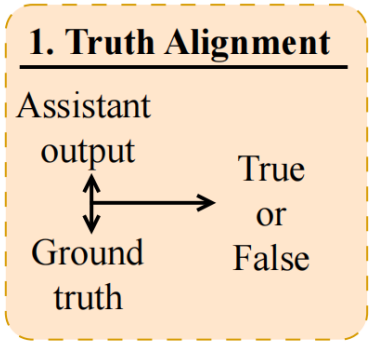
CoIN

Truth Alignment = Reasoning Capability + Instruction Following

Truth Alignment:

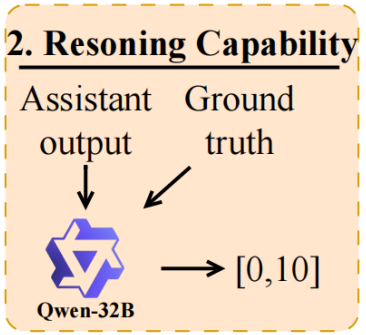
Directly compare the outputs of MLLMs with ground truths.

Two apples ⟺ **Two**



Reasoning Capability:

MLLMs may correctly answer the question logically as "Two apples" while the ground truth is "Two"



Metrics: Mean Average Accuracy (MAA):

$$MAA = \frac{1}{T}\sum_{j=1}^{T}\left(\frac{1}{j}\sum_{i=1}^{j} A_{j,i}\right)$$

Backward Transfer (BWT):

$$BWT = \frac{1}{T-1}\sum_{i=1}^{T-1} A_{T,i} - A_{i,i}$$

Cheng Chen et al. "CoIN: A Benchmark of Continual Instruction Tuning for Multimodal Large Language Models." *NeurIPS 2024*

Experiments

MLLM	Method	Accuracy on Each Task								Overall Results	
		ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
LLaVA	Multi-task	56.77	49.35	95.55	56.65	53.90	30.09	59.50	55.65	57.18	-
	Zero-shot	49.91	2.88	0.33	2.08	0.90	0.00	0.68	0.17	7.12	-
	Sequential Finetune	82.45	49.99	96.05	56.40	55.45	31.27	62.20	57.08	32.97	-32.62
		21.26	28.74	10.25	36.78	32.45	0.83	42.50	57.08		
Qwen-VL	Multi-task	25.70	60.88	17.05	56.77	35.58	6.78	68.67	63.50	41.87	-
	Zero-shot	64.56	48.15	11.82	44.50	9.57	0.00	64.10	27.50	33.78	-
	Sequential Finetune	67.69	66.36	53.70	59.30	36.38	63.10	71.00	47.80	43.35	-16.94
		31.05	42.45	29.57	55.57	15.30	40.33	67.75	47.80		
MiniGPT-v2	Multi-task	43.55	19.24	10.57	28.43	41.62	0.00	27.12	1.45	21.50	-
	Zero-shot	32.16	6.83	0.07	11.58	35.20	0.00	12.20	0.03	12.26	-
	Sequential Finetune	28.81	10.40	7.25	31.55	41.35	0.00	36.10	6.15	25.45	6.04
		44.35	29.89	11.90	36.95	42.58	0.00	38.10	6.15		

The results evaluating the *Truth Alignment* ability are presented. The first line of **Sequential Finetune** are the results for each task evaluated when just tuned on the corresponding task, and the second line displays the final results of each task after fine-tuning on the last task.

Experiments

MLLM	Method	Accuracy on Each Task								Overall Results	
		ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
LLaVA	Multi-task	80	75	97	72	42	86	73	79	75.50	-
	Zero-shot	93	83	69	64	48	35	64	66	65.25	-
	Sequential Finetune	92 82	75 74	97 55	72 56	42 47	58 52	75 58	78 78	71.28	-10.88
Qwen-VL	Multi-task	98	82	68	77	50	51	82	88	74.50	-
	Zero-shot	97	81	78	74	54	58	81	74	74.63	-
	Sequential Finetune	96 95	83 78	86 77	78 77	51 47	82 76	82 82	75 75	80.97	-3.25
MiniGPT-v2	Multi-task	96	76	58	62	44	89	63	59	68.38	-
	Zero-shot	98	72	48	63	48	80	64	61	66.75	-
	Sequential Finetune	97 89	71 73	55 59	61 60	44 44	91 94	63 63	52 52	75.05	0.00

The evaluation results of Reasoning Capability are presented

Experiments

Whether is Qwen a good evaluator?

Type	Accuracy on Each Task								Overall Results	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
Qwen-32B	92	75	97	72	42	58	75	78	71.28	-10.88
	82	74	55	56	47	52	58	78		
GPT-4	94	83	96	83	79	71	81	69	73.62	-11.50
	80	83	65	67	62	70	68	69		
User Study	96	82	98	85	80	65	86	70	74.35	-8.13
	85	80	85	71	76	57	73	70		

The comparison of Qwen with GPT-4 and user study as a evaluator are presented

Experiments

What factors affect the performance?

Order	Accuracy on Each Task								Overall Results	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
Random	82.45	49.99	96.05	56.40	55.45	31.27	62.20	57.08	32.97	-32.62
	21.26	28.74	10.25	36.78	32.45	0.83	42.50	57.08		
	GQA	Grounding	ImageNet	OCR-VQA	ScienceQA	TextVQA	VizWiz	VQAV2	MAA	BWT
Alphabet	62.68	37.73	97.30	62.00	59.98	50.98	60.10	67.28	31.08	-25.90
	53.92	0.00	8.57	37.75	44.37	53.37	25.27	67.28		

The results of LLaVA about different **task orders** are presented

Experiments

What factors affect the performance?

Type	Accuracy on Each Task								Overall Results	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
Original	82.45	49.99	96.05	56.40	55.45	31.27	62.20	57.08	32.97	-32.62
	21.26	28.74	10.25	36.78	32.45	0.83	42.50	57.08		
Diverse	82.45	50.14	96.03	55.65	51.42	34.00	59.17	52.92	32.92	-33.67
	26.00	25.38	8.40	33.07	26.52	0.10	40.00	52.92		
10Type	81.65	51.99	97.00	61.30	54.10	39.20	68.15	64.65	38.37	-31.75
	54.84	35.46	9.80	38.70	12.95	0.82	46.80	64.65		

The results of LLaVA about different **instruction templates** are presented

Experiments

What factors affect the performance?

Volume	Accuracy on Each Task								Overall Results	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
0.1	70.00	42.88	93.45	36.93	43.7	3.73	40.48	45.62	30.27	-16.17
	53.71	32.62	5.38	33.50	36.98	2.85	36.77	45.62		
0.2	69.86	46.86	94.38	44.98	44.15	4.81	32.55	52.10	30.33	-19.89
	41.12	33.25	5.53	33.80	25.85	1.77	37.10	45.62		
0.4	75.33	47.06	94.95	52.95	50.77	10.25	56.73	55.33	33.18	-24.85
	49.96	23.60	7.22	36.12	33.05	0.09	39.20	55.33		
0.6	78.09	47.65	95.85	55.93	53.08	10.00	59.17	46.33	31.47	-32.57
	27.42	19.54	7.03	33.52	13.15	0.05	38.48	46.33		
0.8	80.02	48.13	95.45	54.00	49.85	28.33	58.35	56.67	30.00	-33.60
	11.74	16.94	8.85	32.62	35.50	0.00	39.67	56.67		
1.0	82.45	49.99	96.05	56.40	55.45	31.27	62.20	57.08	32.97	-32.62
	21.26	28.74	10.25	36.78	32.45	0.83	42.50	57.08		

The results of LLaVA about different **data volumes** are presented

MoELoRA

Size	Accuracy on Each Task								Overall Results	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
7B	82.45	49.99	96.05	56.40	55.45	31.27	62.20	57.08	32.97	-32.62
	21.26	28.74	10.25	36.78	32.45	0.83	42.50	57.08		
13B	82.95	54.25	97.28	52.45	59.40	40.35	68.10	61.00	39.43	-28.79
	60.03	41.19	10.62	31.03	32.67	2.60	46.33	61.00		

Size	Accuracy on Each Task								Overall Results	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
7B	92	75	97	72	42	58	75	78	71.28	-10.88
		82	74	55	56	47	52	58	78	
13B	94	77	98	77	46	76	80	79	75.98	-11.00
	89	77	58	59	53	62	62	79		

The results evaluating the **Truth Alignment** and **Reasoning Capability** of LLaVA about different model size are presented

Cheng Chen et al. "CoIN: A Benchmark of Continual Instruction Tuning for Multimodal Large Language Models." ***NeurIPS 2024***

Example Analysis

1. TextVQA

User: What is written on the front of that building? Reference OCR token:M,AVAL,LOOCKYARD,BOCK, ESTA,1735,.2,LION,GATE,djakoeb. Answer the question using a single word or phrase.
Assistant: Sunscreen
Ground_truth: lion gate

Truth Alignment: False
Reasoning Capability: 0

2. ScienceQA

User: What is the name of the colony shown? A. Maryland B. New Hampshire C. Rhode Island D. Vermont. Answer with the option's letter from the given choices directly.
Assistant: New Hampshire
Ground_truth: B

Truth Alignment : False
Reasoning Capability : 10

3. GQA

User: What animals are on the dirt? Answer the question using a single word or phrase.
Assistant: Elephants
Ground_truth: Elephants

Truth Alignment : True
Reasoning Capability : 10

4. ImageNet

User: What is the object in the image? Answer the question using a single word or phrase.
Assistant: Car.
Ground_truth: Pickup

Truth Alignment : False
Reasoning Capability : 8

MoELoRA

Number	Accuracy on Each Task								Overall Results	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
Multi-task(1)	56.77	49.35	95.55	56.65	53.90	30.09	59.50	55.65	57.18	-
1	82.45	49.99	96.05	56.40	55.45	31.27	62.20	57.08	32.97	-32.62
	21.26	28.74	10.25	36.78	32.45	0.83	42.50	57.08		
2	79.93	51.37	95.92	59.60	55.33	32.29	63.15	54.15	35.75	-28.03
	47.77	31.67	10.75	37.10	40.98	1.44	43.65	54.15		
4	80.35	52.21	96.25	59.62	58.05	34.47	64.40	62.73	40.24	-26.57
	65.36	40.28	11.10	37.20	34.77	0.49	43.60	62.73		
8	75.78	51.73	96.70	59.42	58.88	37.50	64.22	60.08	42.76	-25.91
	63.09	38.63	10.50	37.38	43.62	0.59	43.15	60.08		

The results of LLaVA about different **numbers of experts** are presented

MoELoRA

Method	Accuracy on Each Task								Overall Results	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
Baseline	75.33	47.06	94.95	52.95	50.77	10.25	56.73	55.33	33.18	-24.85
	49.96	23.60	7.22	36.12	33.05	0.09	39.20	55.33		
LwF	75.33	48.18	96.90	48.58	44.12	6.60	38.58	62.35	35.89	-19.27
	63.14	39.60	8.90	34.83	14.53	2.48	40.67	62.35		
EWC	75.28	48.37	96.83	42.77	44.25	8.65	60.27	61.02	40.36	-17.94
	67.41	40.41	8.18	35.05	37.88	2.67	41.27	61.02		
MoELoRA	75.85	49.05	93.95	56.53	48.70	25.57	61.9	55.35	41.05	-22.50
	58.92	38.59	8.85	37.10	44.25	2.45	41.40	55.35		

The comparison with **other continual learning methods** based on LLaVA is presented

Cheng Chen et al. "CoIN: A Benchmark of Continual Instruction Tuning for Multimodal Large Language Models." ***NeurIPS 2024***

Thanks

GitHub: https://github.com/zackschen/CoIN

Contacting us if you have any questions:

Email: cczacks@gmail.com