ICLR

Vienna, Austria

Grounding Multimodal Large Language Models to the World

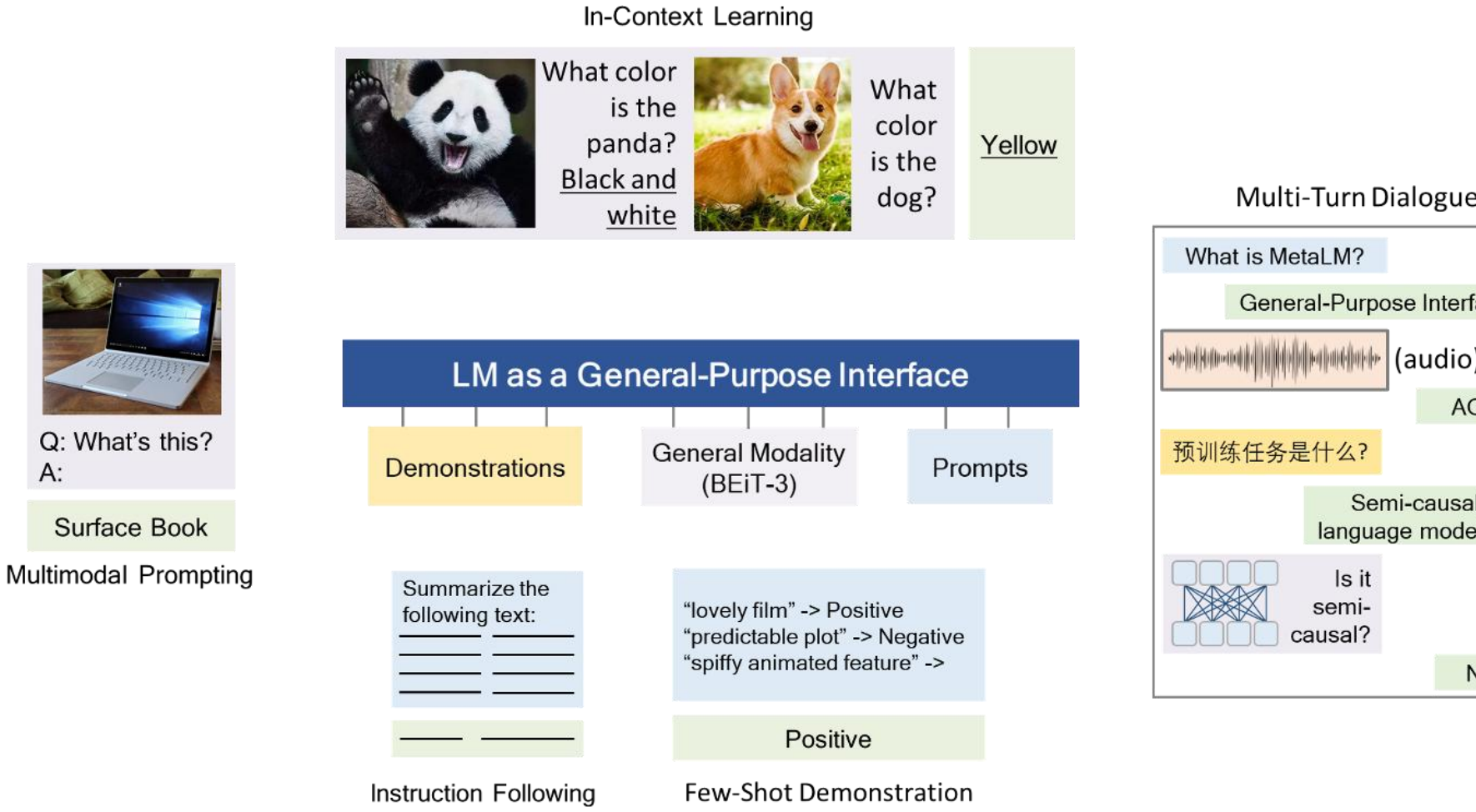
Zhiliang Peng*· Wenhui Wang*· Li Dong*· Yaru Hao · Shaohan Huang · Shuming Ma · Qixiang Ye · Furu Wei

中国科学院大学

University of Chinese Academy of Sciences

Microsoft

LLMs are General-Purpose Interfaces



Grounding Multimodal Large Language Models to the World

Kosmos-2 Capabilities

1. Language & Vision-Language Mastery
2. *Multimodal Grounding & Referring*
3. *Downstream Applications Potential*

Decode location tokens to coordinates

[**a campfire**](<loc$_4$> <loc$_{1007}$>)

Kosmos–2: Multimodal Large Language Model

[**It**](<loc$_{44}$> <loc$_{863}$>) sits next to

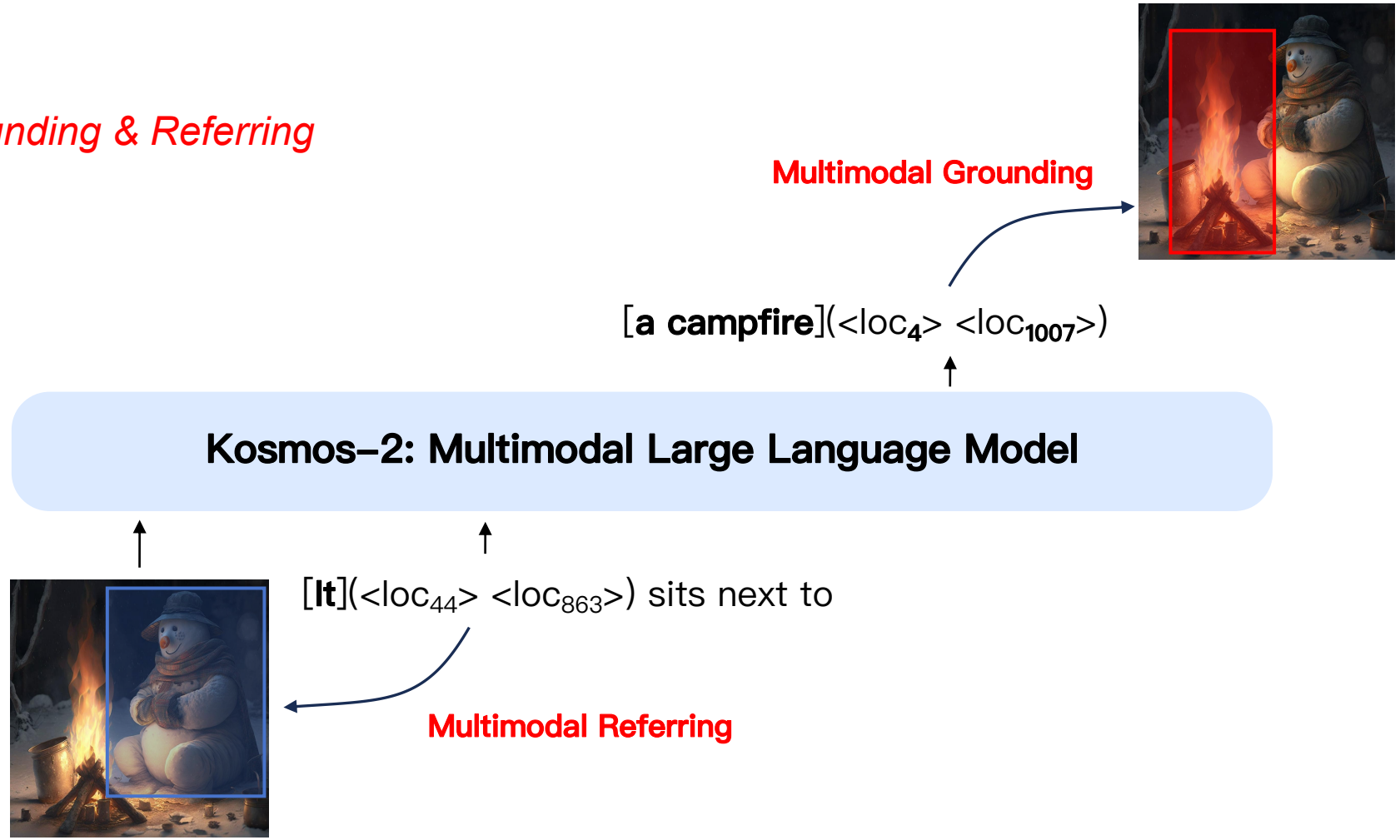
Encode coordinates of bounding box to location tokens

Top–left: (x1, y1)
Bottom–right: (x2, y2)

ICLR

Grounding Multimodal Large Language Models to the World

Multimodal Grounding & Referring



ICLR

Grounding Multimodal Large Language Models to the World

Hyperparameters	
Number of layers	24
Hidden size	2,048
FFN inner hidden size	8,192
Attention heads	32
Dropout	0.1
Attention dropout	0.1
Activation function	GeLU [HG16]
Vocabulary size	64,007
Soft tokens V size	64
Max length	2,048
Relative position embedding	xPos [SDP+22]
Initialization	Magneto [WMH+22]

Training Data

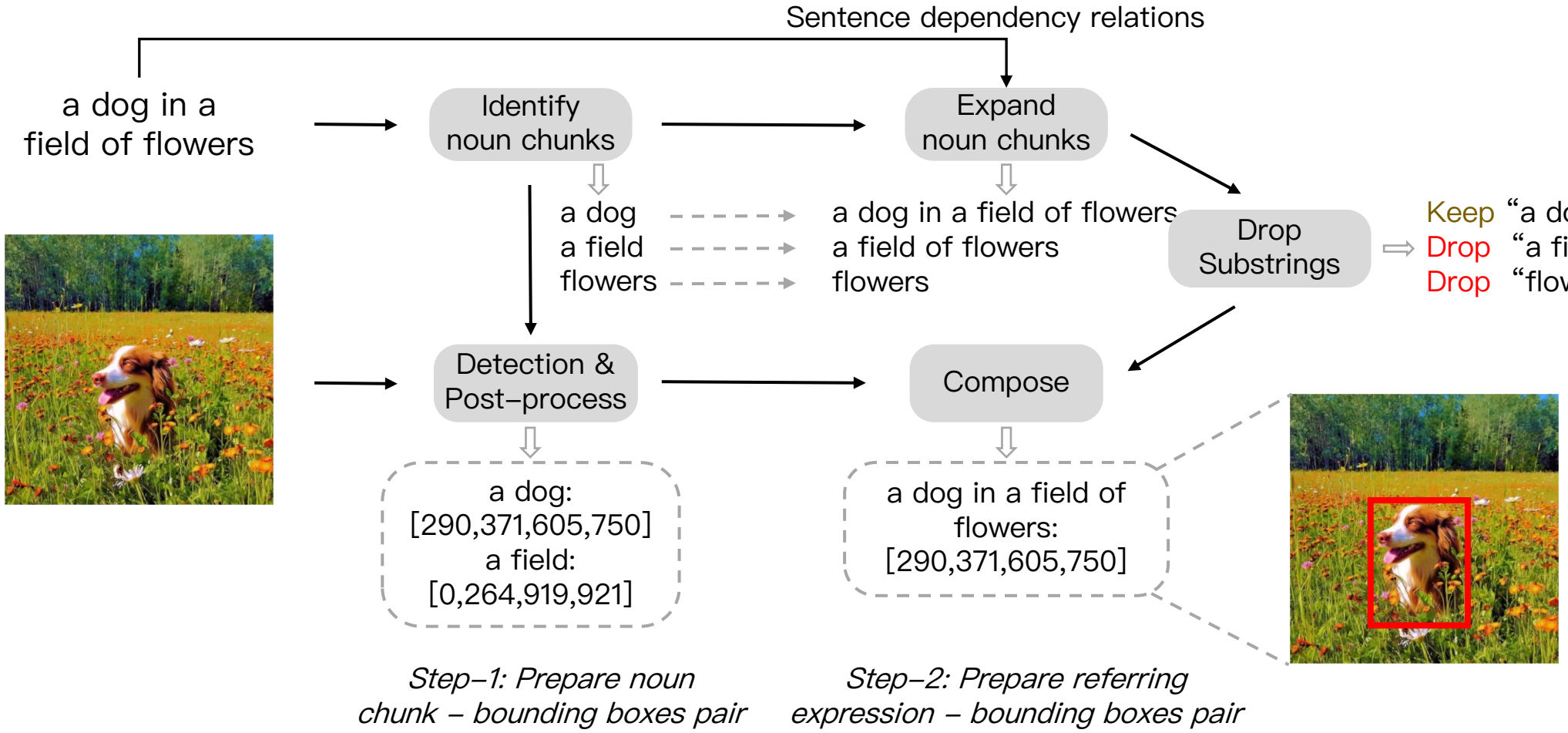
Text Corpora

Interleaved Image-Text Data

Grounded Image-Text Data

- **Academic**: NIH Exporter
- **Internet**: Pile-CC, OpenWebText2, Wikipedia (English), CC-2020-50, CC-2021-04, Realnews
- **Prose**: BookCorpus2, Books3, Gutenberg [RPJ+20], CC-Stories

Constructing Grounded Image-Text Data



Resulting 90M Images, 137M objects

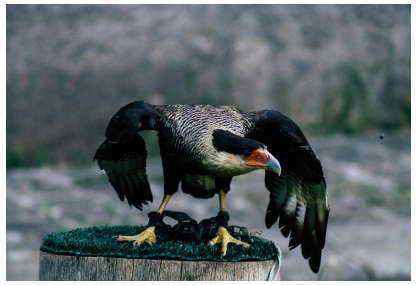
Examples

Input Prompt

[The left eye of emoji]

Question: How many cows are here? Answer:

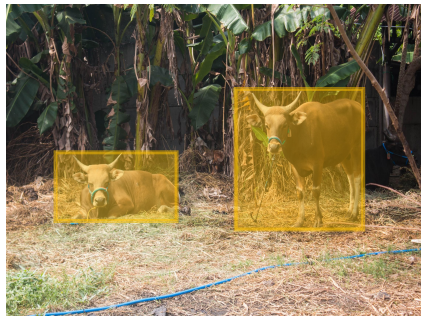
Question: What does it say? Answer:



Describe this bird in detail:

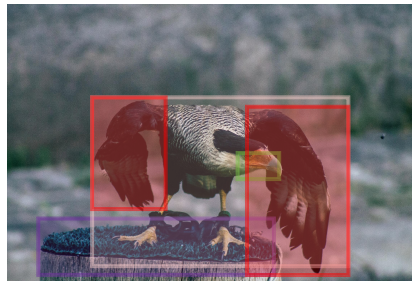
Completion

two cows present in the image.



"Je suis Charlie!" is written in French on a chalkboard.

The eagle is a large bird with a black and white body and a yellow beak. It has large, dark wings and a pointed tail. It stands on a stump and spreads its wings, displaying its magnificent plumage.



Referring Expression Comprehension

VQA Counting

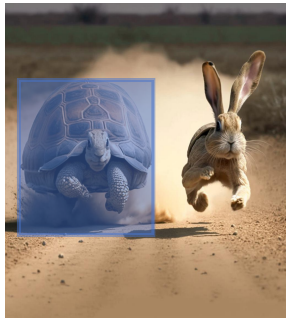
Text Recognition

Image Caption

ICLR

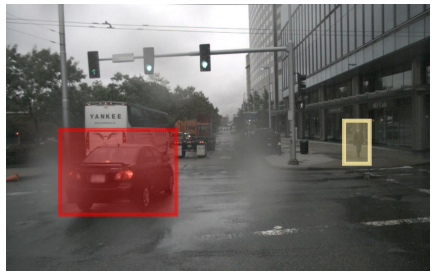
Examples

Input Prompt



Question: Why is this animal unusual? Answer:

Question: What is it ? Answer:



You are driving this car. To pick up that man, you need to

Question: What is the biggest difference betwen bottle-1 and bottle-2? Answer:

Completion

This animal is unique because it is a turtle. Turtles are known for their slow and steady pace, whereas rabbits are known to be agile and quick. It is not common for a turtle to be faster than a rabbit, especially in terms of speed.

a buoy attached to the boat.

go through the intersection and turn right.

The biggest difference between bottle-1 and bottle-2 is the label.

Kosmos-2 Performance

Model	Zero-shot	Val Split			Test Split		
		R@1	R@5	R@10	R@1	R@5	R@10
VisualBert (Li et al., 2019)	✗	70.4	84.5	86.3	71.3	85.0	86.5
MDETR (Kamath et al., 2021)	✗	83.6	93.4	95.1	84.3	93.9	95.8
GLIP (Li et al., 2022b)	✗	86.7	96.4	97.9	87.1	96.9	98.1
FIBER (Dou et al., 2022)	✗	87.1	96.1	97.4	87.4	96.4	97.6
GRILL (Jin et al., 2023)	✓	-	-	-	18.9	53.4	70.3
KOSMOS-2	✓	77.8	79.2	79.3	78.7	80.1	80.1

Table 1: Phrase grounding results on Flickr30k Entities. We report the R@1, R@5, and R@10 metrics, where R@1/5/10 means calculating the recall using the top 1/5/10 generated bounding boxes.

Model	Setting	RefCOCOg	
		Meteor	CIDEr
SLR (Yu et al., 2017)	Finetuning	15.4	59.2
SLR+Rerank (Yu et al., 2017)	Finetuning	15.9	66.2
KOSMOS-2	Zero-shot	12.2	60.3
	Few-shot ($k = 2$)	13.8	62.2
	Few-shot ($k = 4$)	14.1	62.3

Table 3: Results of referring expression generation on RefCOCOg.

Model	Zero-shot	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
UNITER (Chen et al., 2019)	✗	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
MDETR (Kamath et al., 2021)	✗	87.51	90.40	82.67	81.13	85.52	72.96	83.35	83.31
OFA (Wang et al., 2022c)	✗	90.05	92.93	85.26	84.49	90.10	77.77	84.54	85.20
FIBER (Dou et al., 2022)	✗	90.68	92.59	87.26	85.74	90.13	79.38	87.11	87.32
VisionLLM (Wang et al., 2023)	✗	86.70	-	-	-	-	-	-	-
GRILL (Jin et al., 2023)	✓	-	-	-	-	-	-	-	47.50
KOSMOS-2	✓	52.32	57.42	47.26	45.48	50.73	42.24	60.57	61.65

Table 2: Accuracy of referring expression comprehension.

Kosmos-2 Performance

Rank	Model	Accuracy(%)
1	KOSMOS-2	63.36
2	InstructBLIP	60.29
3	InstructBLIP Vicuna	60.20
4	BLIP2	59.12
5	MiniGPT-4	56.27
6	VPGTrans	51.87
7	mPLUG-Owl	49.68
8	VideoChat	47.12
9	LLaMA-Adapter V2	45.22
10	Otter	44.90

(1) Scene Understanding

Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	58.93
2	InstructBLIP	58.49
3	KOSMOS-2	57.07
4	BLIP2	53.90
5	MiniGPT-4	49.15
6	mPLUG-Owl	45.33
7	VPGTrans	44.13
8	VideoChat	43.80
9	Otter	38.56
10	LLaMA-Adapter V2	38.50

(2) Instance Identity

Rank	Model	Accuracy(%)
1	KOSMOS-2	43.96
2	InstructBLIP Vicuna	43.56
3	BLIP2	42.33
4	InstructBLIP	40.59
5	VideoChat	39.98
6	MiniGPT-4	37.93
7	mPLUG-Owl	36.71
8	VPGTrans	36.09
9	LLaMA-Adapter V2	33.03
10	Flan-T5	31.75

(3) Instance Location

Rank	Model	Accuracy(%)
1	InstructBLIP Vicuna	40.33
2	InstructBLIP	38.66
3	KOSMOS-2	37.90
4	BLIP2	36.68
5	VPGTrans	36.38
6	LLaMA-Adapter V2	35.46
7	VideoChat	34.55
8	mPLUG-Owl	32.72
9	MiniGPT-4	32.57
10	GVT	31.96

(4) Spatial Relations

Rank	Model	Accuracy(%)
1	KOSMOS-2	55.67
2	BLIP2	55.67
3	InstructBLIP Vicuna	52.58
4	InstructBLIP	51.55
5	MiniGPT-4	47.42
6	mPLUG-Owl	44.33
7	VideoChat	42.27
8	LLaMA-Adapter V2	39.18
9	Flan-T5	32.98
10	VPGTrans	31.96

(5) Instance Interaction

Rank	Model	Accuracy(%)
1	KOSMOS-2	60.72
2	MiniGPT-4	57.10
3	mPLUG-Owl	54.68
4	VPGTrans	53.17
5	LLaMA-Adapter V2	51.96
6	Otter	51.36
7	MultiModal-GPT	51.36
8	GVT	51.06
9	VideoChat	50.45
10	OpenFlamingo	50.15

(6) Visual Reasoning

ICLR

ICLR

University of Chinese Academy of Sciences

Microsoft

Thanks for your attention!

Paper

Official Code

Nvidia Host Demo

HF Space Demo