MoAI: *Mixture of All Intelligence for Large Language and Vision Model*

Milano, Italy

Byung-Kwan Lee, Beom Chan Park, Chae Won Kim, Yong Man Ro

Introduction

Scene Perception: Detecting and Judging Objects Undergoing Relational Violations

IRVING BIEDERMAN, ROBERT J. MEZZANOTTE, AND JAN C. RABINOWITZ
State University of New York at Buffalo

Scene Perception in the Human Brain

Russell A. Epstein[1] and Chris I. Baker[2]

[1]Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; email: epstein@psych.upenn.edu

[2]Section on Learning and Plasticity, Laboratory of Brain and Cognition, National Institute of Mental Health, Bethesda, Maryland 20892, USA; email: bakerchris@mail.nih.gov

Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models

Sivan Doveh[1,2] Assaf Arbelle[1] Sivan Harary[1] Roei Herzig[1,3] Donghyun Kim[4]
Paola Cascante-bonilla[4] Amit Alfassy[1,5] Rameswar Panda[4] Raja Giryes[3]
Rogerio Feris[4] Shimon Ullman[2] Leonid Karlinsky[4]

[1]IBM Research, [2]Weizmann Institute of Science, [3]Tel-Aviv University, [4]MIT-IBM Watson AI Lab, [5]Technion

From Cognitive Science to Machine Learning,

Real-world scene understanding

- *Recognizing object presences*
- *Determining their positions*
- *Identifying their states*
- *Understanding their relationships*
- *Extracting spatial scene layouts*
- *Grasping non-object notions*

Compositional Reasoning

KAIST EE
KAIST ELECTRICAL ENGINEERING

Introduction

Real-world scene understanding	From Rich CV Models
• *Recognizing object presences* →	*Segmentation/Detection*
• *Determining their positions* →	*Segmentation/Detection*
• *Identifying their states* →	*Scene Graph Generation*
• *Understanding their relationships* →	*Scene Graph Generation*
• *Extracting spatial scene layouts* →	*Segmentation/Detection/Scene Graph Generation*
• *Grasping non-object notions* →	*OCR*

Proposed Method

External CV Models

- *Panoptic Segmentation (Mask2Former, CVPR 2022, Meta)* — *Swin-B/4 106M*
- *Open-World Object Detection (OWLv2, NeurIPS 2023, Google)* — *CLIP-B/16 154M (We use ADE20K-847+ImageNet)*
- *Scene Graph Generation (Panoptic SGG, ECCV 2022, Nanyang Tech)* — *ResNet-50 44M*
- *OCR (PaddleOCRv2, performant open-source OCR)* — *14M (Chinese & Englinsh, Rotated Text O)*

What do we propose?

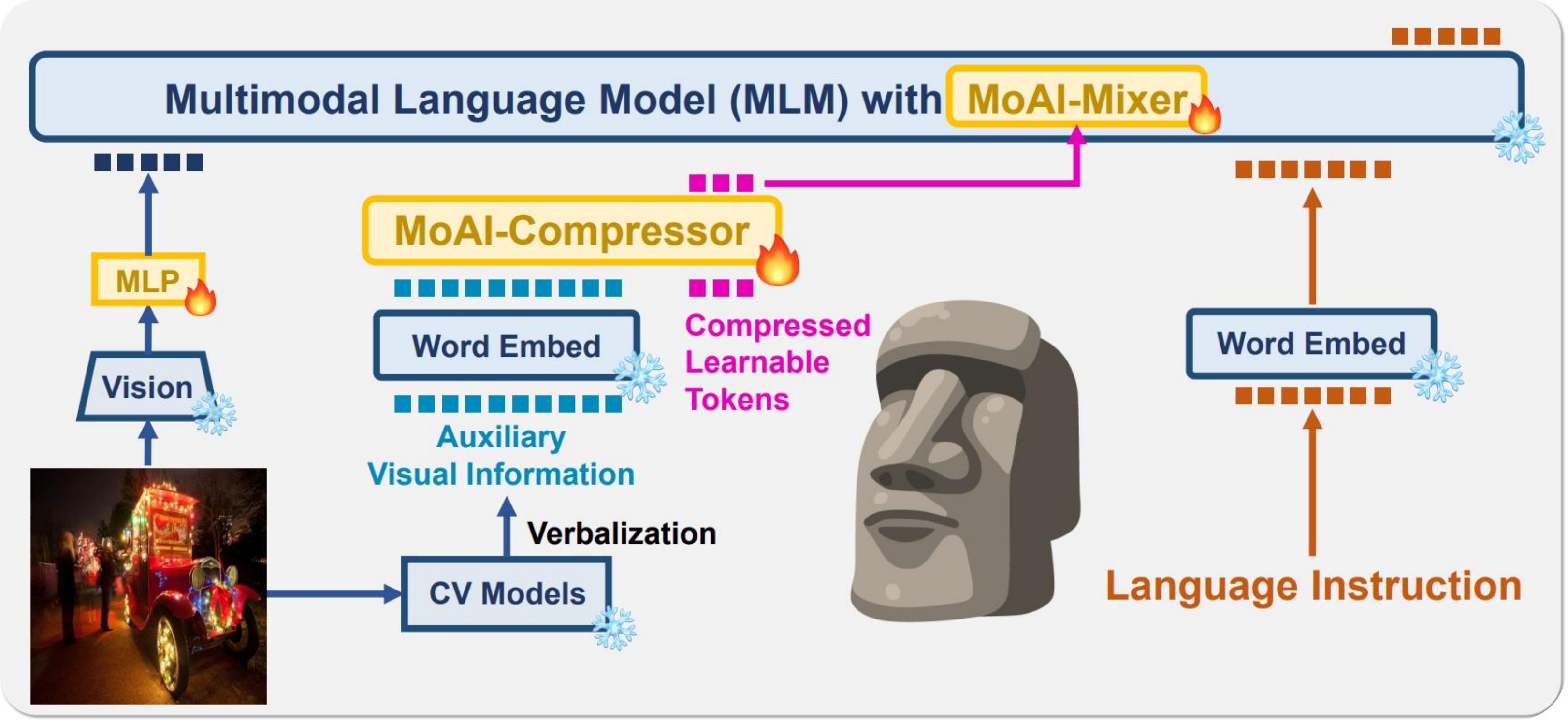
***Compressor**: Compressing Auxiliary Visual Information from external CV models*

***Mixer**: Blending Aux with Visual and Language Features in Multimoal LM (MLM)*

KAIST EE
KAIST ELECTRICAL ENGINEERING

Proposed Method

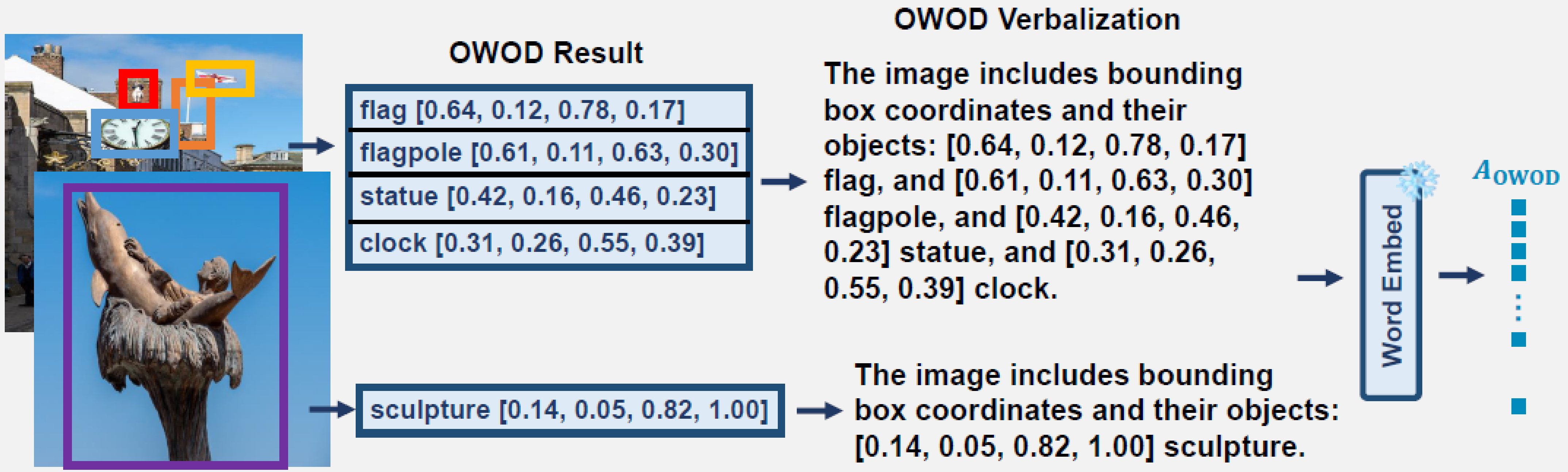
MoAI: *Mixture of All Intelligence*
for Large Language and Vision Model



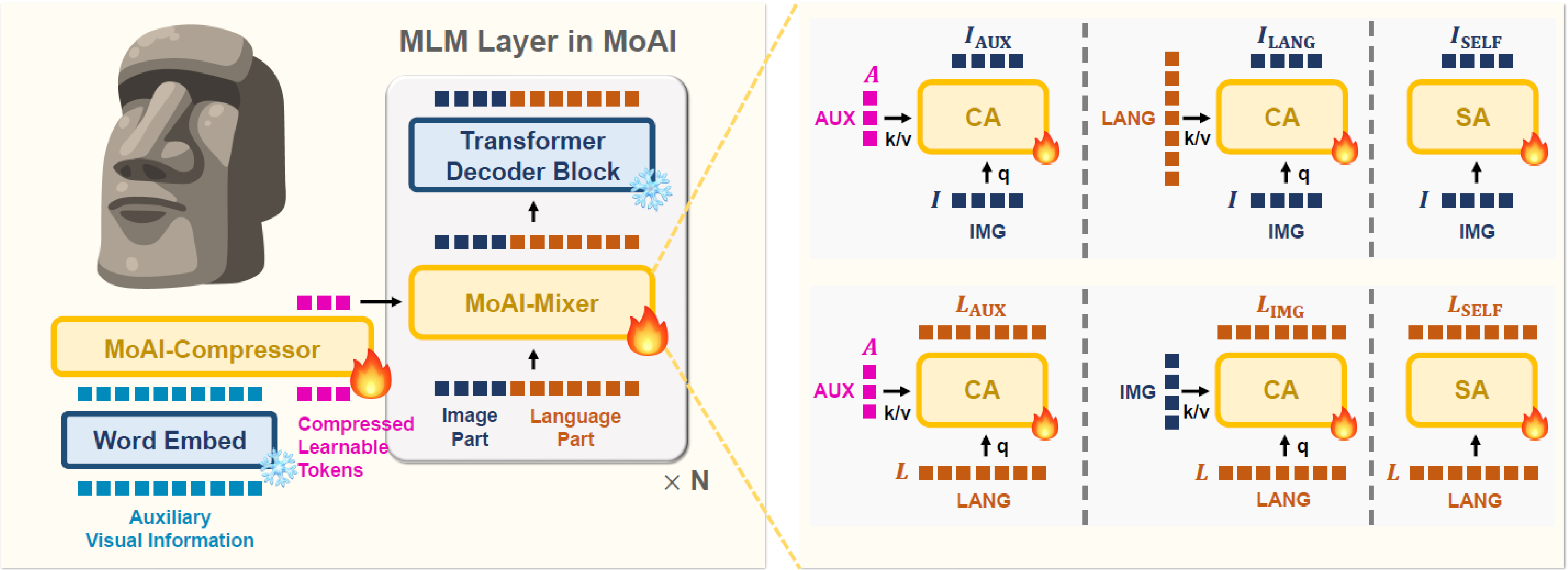
KAIST EE
KAIST ELECTRICAL ENGINEERING

Proposed Method

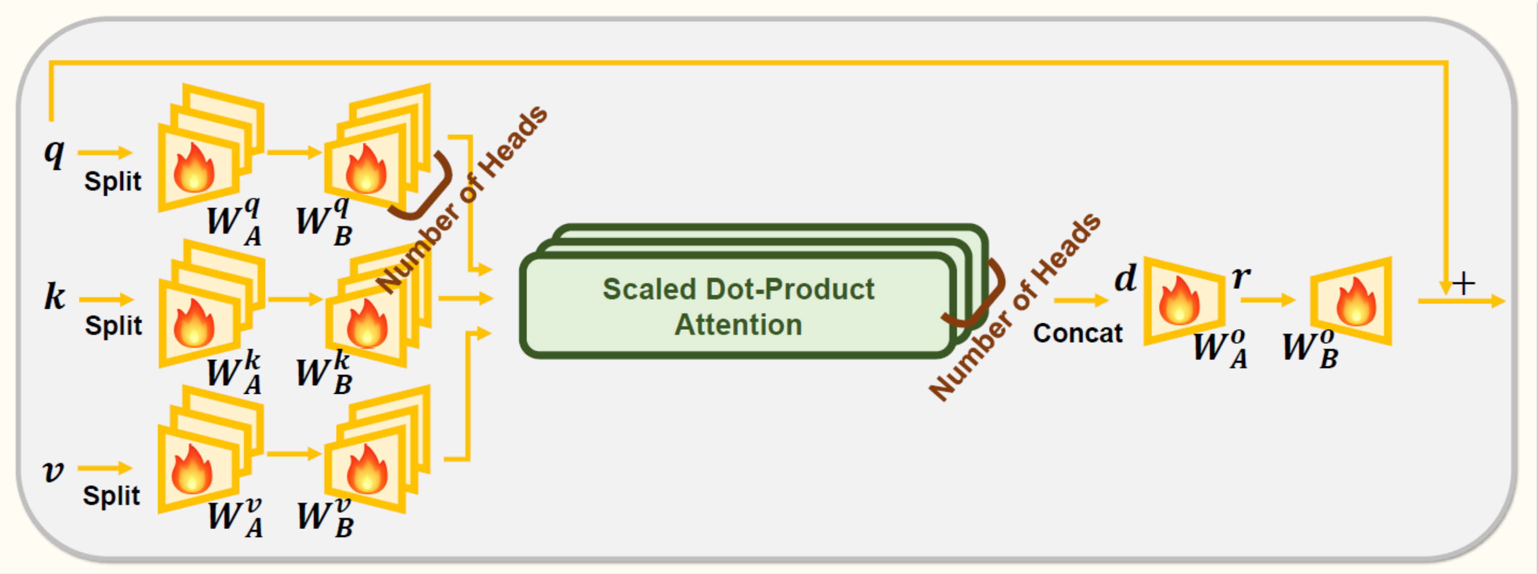
- **Open-World Object Detection (OWOD)**



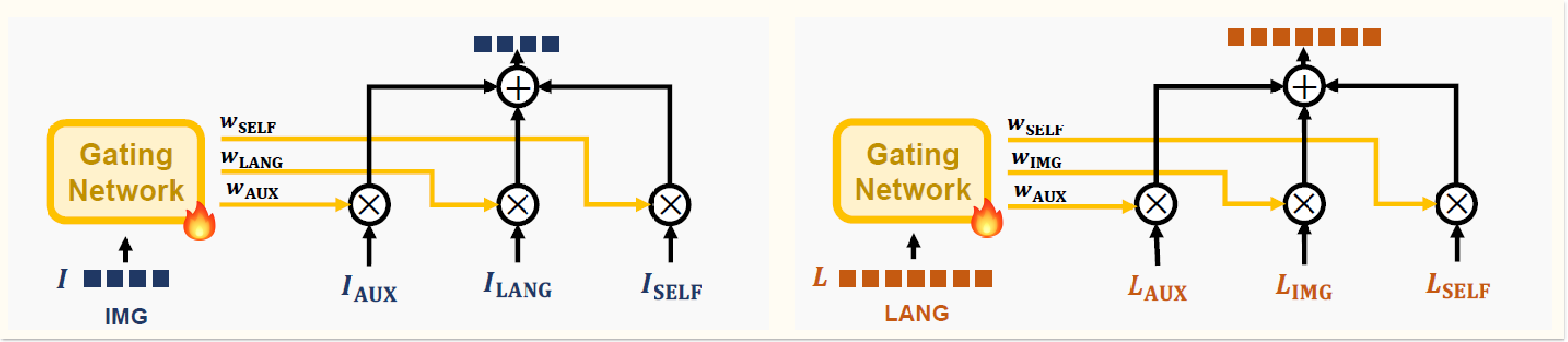
Proposed Method



Proposed Method

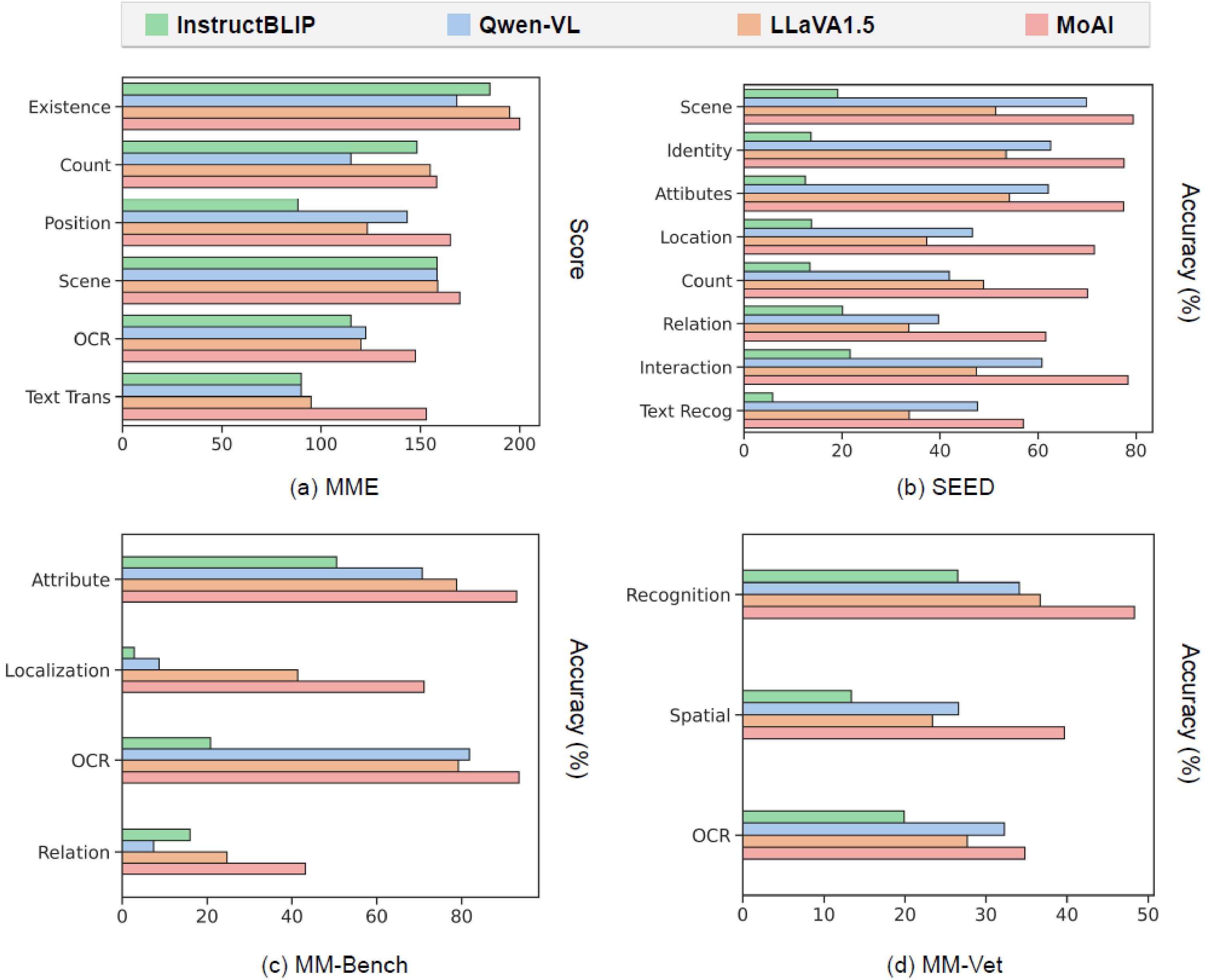


(a) CA/SA with Low Rank Adaptation (LoRA) for Expert Modules



(b) Gating Networks for MoAI-Mixer

Experiments



KAIST EE
KAIST ELECTRICAL ENGINEERING

Experiments

VLMs	Q-Bench	SQA-IMG	TextVQA	POPE	MME-P	MME-C	MM-Bench	MMB-CN	MM-Vet
BLIP2-13B [42]	-	61.0	42.5	85.3	1294	290	-	-	22.4
InstructBLIP-7B [16]	56.7	60.5	50.1	-	-	-	36.0	23.7	26.2
InstructBLIP-13B [16]	-	63.1	50.7	78.9	1213	-	-	-	25.6
Shikra-13B [9]	54.7	-	-	-	-	-	58.8	-	-
IDEFICS-9B [38]	-	-	25.9	-	-	-	48.2	25.2	-
IDEFICS-80B [38]	-	-	30.9	-	-	-	54.5	38.1	-
Qwen-VL-7B [4]	59.4	67.1	63.8	-	-	-	38.2	7.4	-
Qwen-VL-Chat-7B [4]	-	68.2	61.5	-	1488	361	60.6	56.7	-
MiniGPT-4-7B [83]	-	-	-	-	582	-	23.0	-	22.1
Otter-7B [40]	47.2	-	-	-	1292	-	48.3	-	24.6
LLaVA-7B [50]	-	38.5	-	-	807	248	34.1	14.1	26.7
MiniGPT-v2-7B [8]	-	-	-	-	-	-	-	-	-
MiniGPT-v2-Chat-7B [8]	-	-	-	-	-	-	-	-	-
LLaVA1.5-7B [48]	58.7	66.8	58.2	85.9	1511	294	64.3	58.3	30.5
LLaVA1.5-13B [48]	62.1	71.6	61.3	85.9	1531	295	67.7	63.6	35.4
mPLUG-Owl-7B [75]	58.9	-	-	-	967	-	46.6	-	-
mPLUG-Owl2-7B [76]	62.9	68.7	58.2		1450	-	64.5	-	36.2
ShareGPT4V-7B [10]	63.4	68.4	-		1567	376	68.8	62.2	37.6
CogVLM-17B [71]	-	68.7	58.2		-	-	65.8	55.9	**54.5**
LLaVA-XTuner-20B [15]	-	-	-	-	-	-	75.1	73.7	37.2
Intern-XC-7B [78]	64.4	-	-		1528	391	74.4	72.4	35.2
MoAI-7B	**70.2**	**83.5**	**67.8**	**87.1**	**1714**	**561**	**79.3**	**76.5**	43.7

Experiments

Table 2: Illustrating the effectiveness of external computer vision (CV) models compared by the perception scores in MME [25] and MM-Bench [51]. 'TT' denotes text translation task that requires OCR as a priority.

			MME					MM-Bench			
PS+OWOD	SGG	OCR	Existence	Position	Scene	OCR	TT	Recognition	Localization	Spatial	OCR
✗	✓	✓	187	154	161	145	138	77.6	54.0	32.6	84.6
✓	✗	✓	198	145	164	147	150	89.7	65.3	35.8	90.9
✓	✓	✗	199	163	166	120	95	91.8	69.2	42.8	80.1
✓	✓	✓	**200**	**165**	**170**	**148**	**153**	**92.9**	**71.1**	**43.2**	**93.5**

KAIST EE

KAIST ELECTRICAL ENGINEERING

Experiments

Table 3: Ablation study for training step choice, selecting top-k expert modules in MoAI-Mixer, and the type of weights for gating network.

(a) Training step choice

Step	MME-P	MME-C
First	1542	369
Second	1654	511
Combined	**1714**	**561**

(b) Selecting Top-k Experts

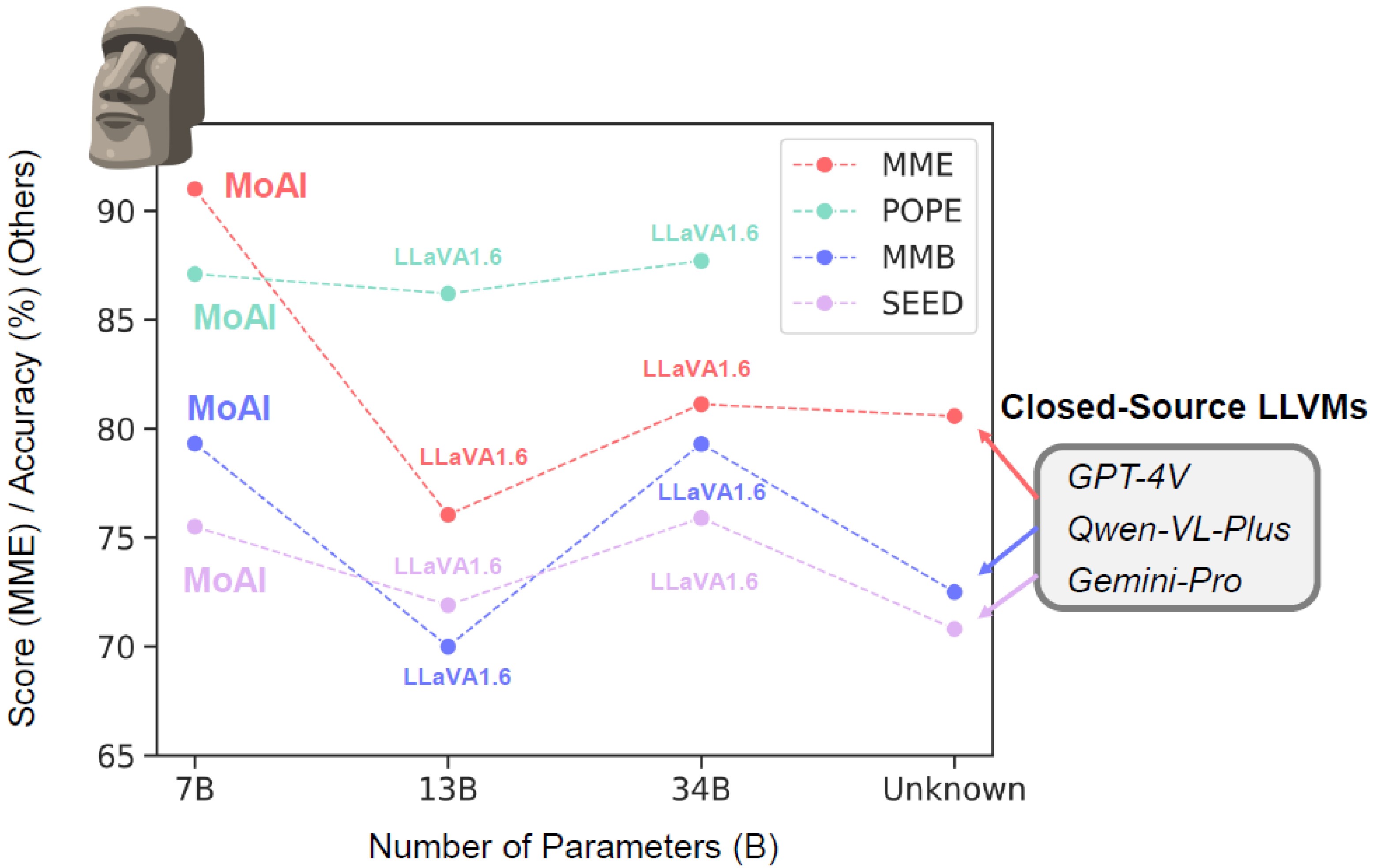
k	MME-P	MME-C
1	1588	387
2	1638	451
3	**1714**	**561**

(c) Gating network weights

Gating	MME-P	MME-C
Random	1520	348
Uniform	1617	485
Trained	**1714**	**561**

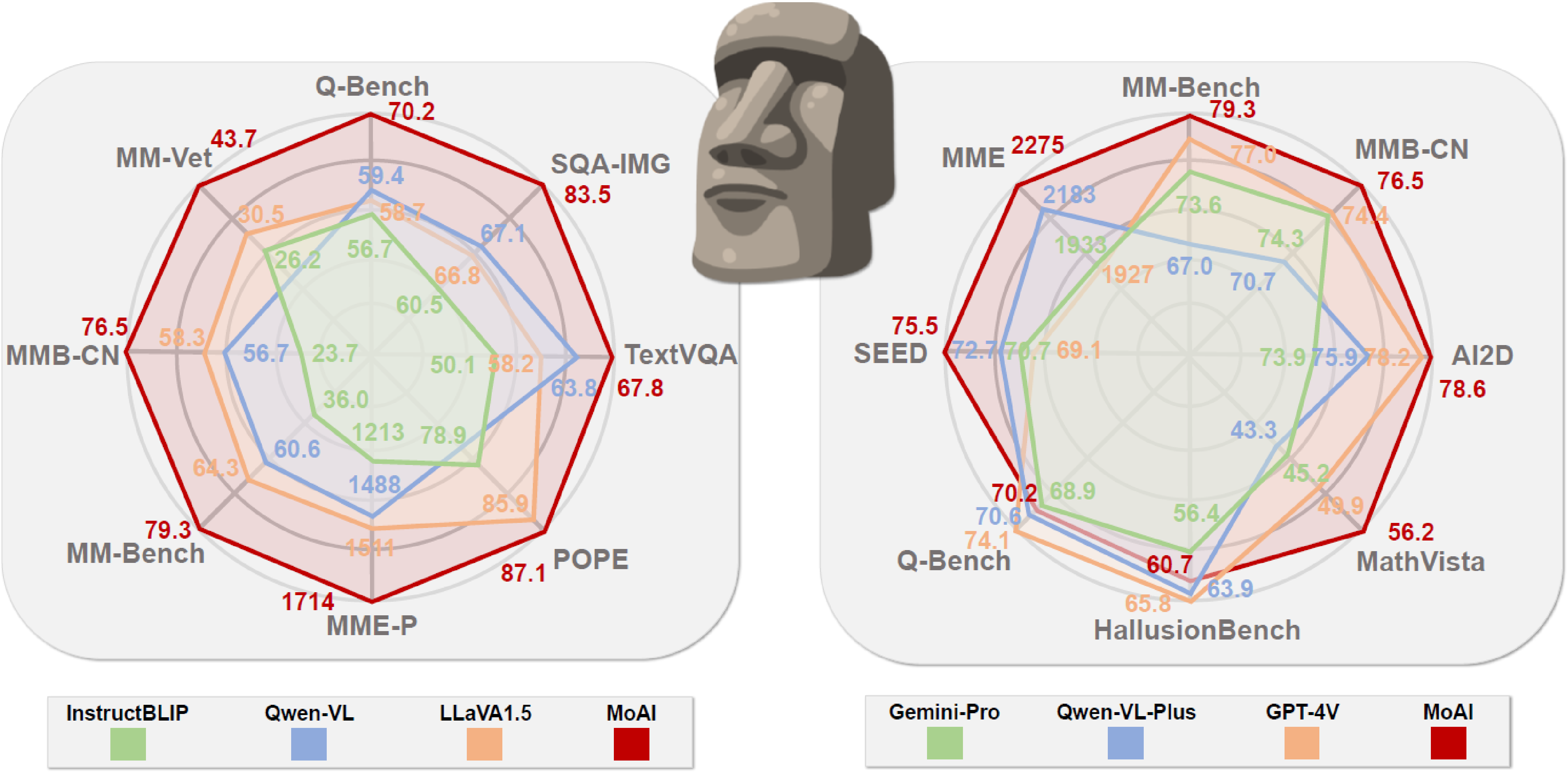
KAIST EE
KAIST ELECTRICAL ENGINEERING

Experiments



(a) MME/POPE/MMB/SEED by Scale

Experiments



(a) Open-source LLVMs

(b) Closed-source LLVMs

KAIST ELECTRICAL ENGINEERING

Discussion and Limitation

Discussion and Limitation. From the results, we can gain insight that prioritizing perception capabilities is more crucial than relying on extra curation of visual instruction datasets or scaling up model size. As illustrated in Fig. 7(a), MoAI-7B surpasses the zero-shot performances despite being relatively small compared to the considerably larger open-source and closed-source models. Notably, Fig. 7(b) also indicates that MoAI performs well even in hallucination zero-shot datasets: POPE [44] and HallusionBench [47]. This suggests that recognizing objects and their relationships accurately can help prevent LLVMs from doing mistakes. Looking ahead, as MoAI is tailored for real-world scene understanding, we will incorporate more external CV models to provide LLVMs with diverse capabilities for low-level vision understanding, common-sense knowledge, non-object notions beyond text descriptions such as charts, diagrams, signs, and symbols, and solving advanced math problems.

KAIST EE
KAIST ELECTRICAL ENGINEERING

Thank you