

Native Multimodal Models: Architecture, Post-Training, and Evaluation

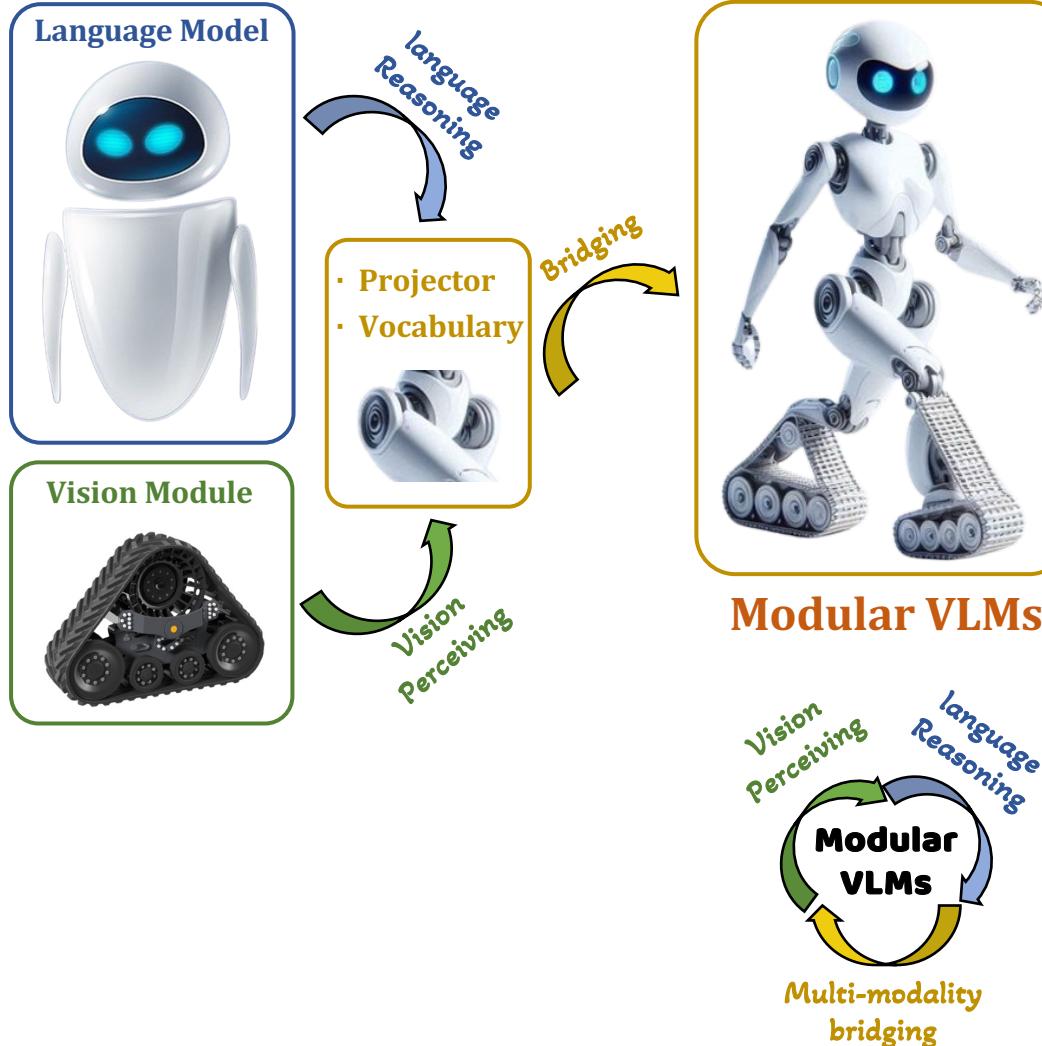
Ziwei Liu 刘子纬

Nanyang Technological University

<https://liuziwei7.github.io>



Background: Modular Vision-Language Models



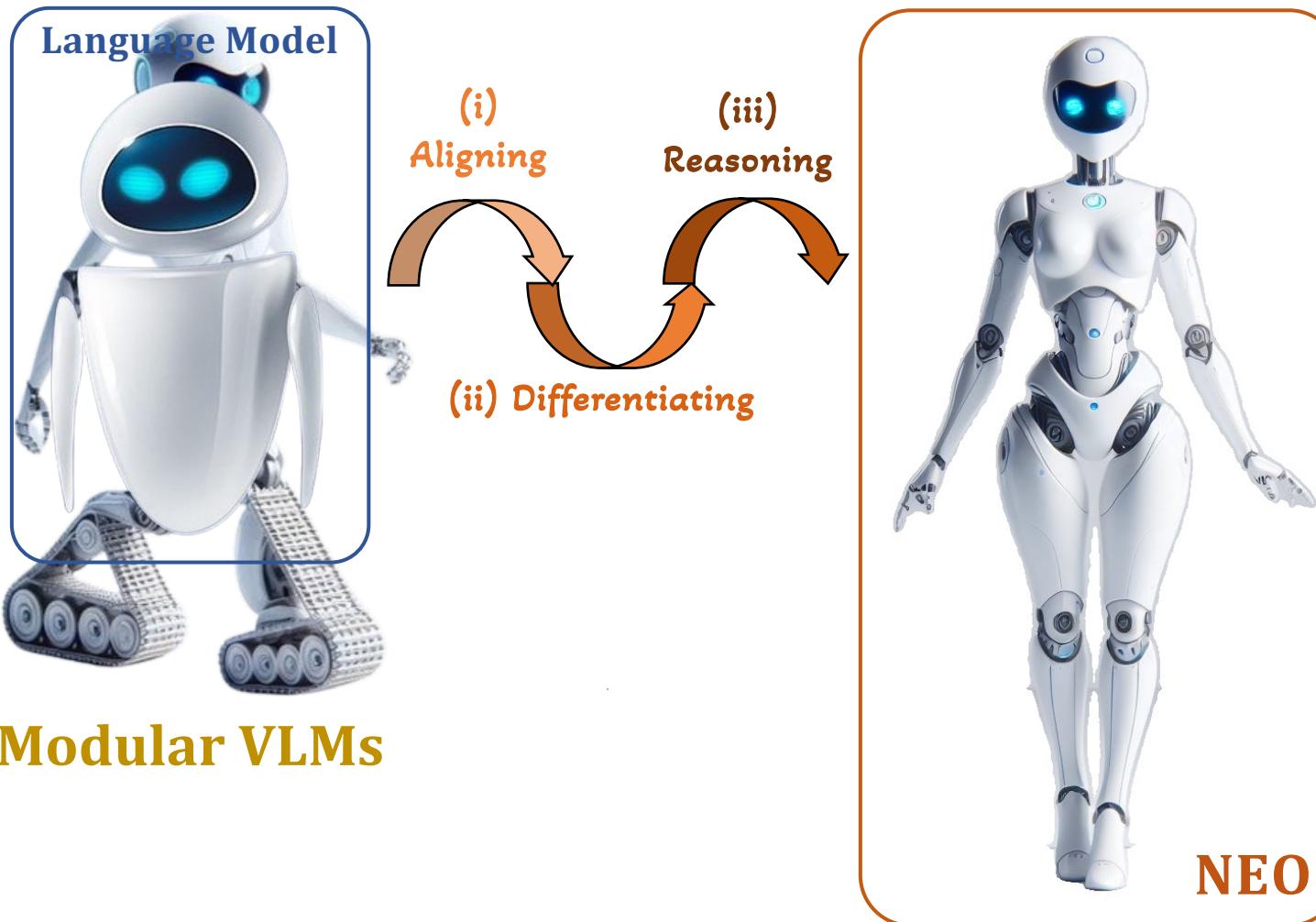
■ Dense Visual Encoder

- 🔥 Well pre-alignment across modules.
- 🔥 Minimal resource costs for adaptation.
- 🧊 Strong visual pre-training **inductive biases**
- 🧊 Complex infrastructure development and scaling analyses of separate components.

■ Discrete Visual Tokenizer

- 🔥 Efficiently model the **unified VLMs**.
- 🔥 Naturally **compatible with multiple modalities**.
- 🧊 Discretization results in **lossy visual features**.
- 🧊 Perform poorly in **fine-grained visual perception**.

Background: Native Vision-Language Models



Thinking:

- Can we remove **vision priors** from existing VLMs?
- How to **transfer** an LLM to a native VLM **efficiently**?
- How to **bridge the gap** between native and modular VLMs?
- How about **mutual synergy** on understanding and generation capabilities of existing VLMs ?

Outline: Native Vision-Language Models



Native Multimodal Evaluation

RealUnify: Do Unified Models Truly Benefit from Unification?
Dual-Evaluation Protocol, Understanding-Generation Synergy

Native Multimodal Post-Training

Visual Jigsaw Post-Training Improves MLLMs
Self-supervised Learning, Post-training, Reinforcement Learning

Native Multimodal Architecture

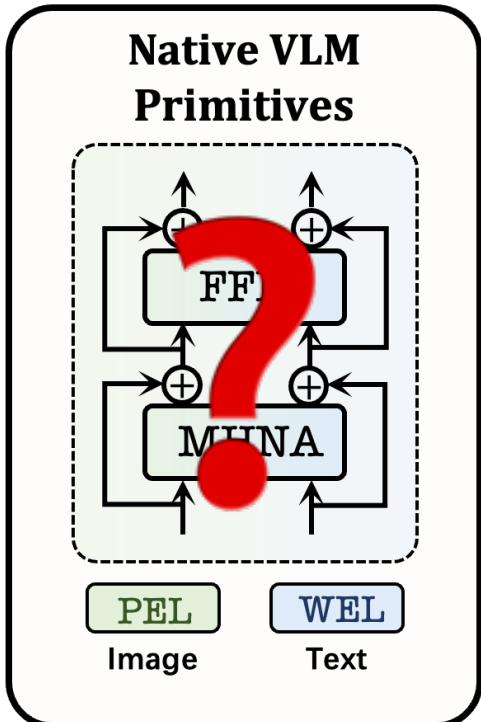
From Pixels to Words: Towards Native Vision-Language Primitives at Scale
Native Vision-Language Primitive, Holistic Vision-Language Buffer

Native Multimodal Architecture

From Pixels to Words: Towards Native Vision-Language Primitives at Scale

Haiwen Diao, Mingxuan Li, Silei Wu, Linjun Dai, Xiaohua Wang, Hanming Deng, Lewei Lu, Dahua Lin, Ziwei Liu

Motivation

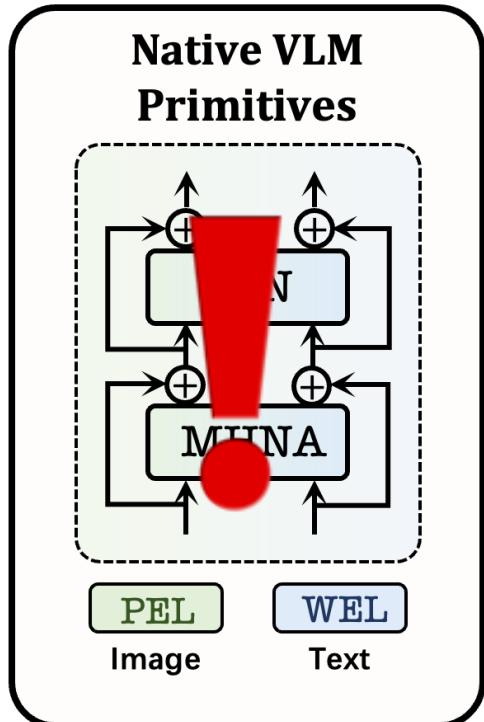


Question:

- What fundamental constraints set native VLMs apart from modular ones, and to what extent can these barriers be overcome?
- How to make research in native VLMs more accessible and democratized, thereby accelerating progress in the field.

These issues prompts us to think about what a native primitive should look like and what characteristics it should have?

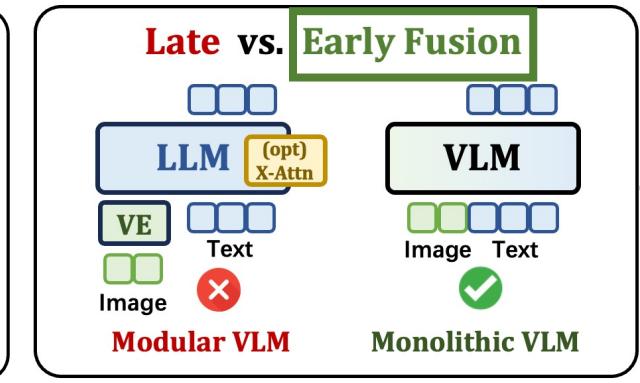
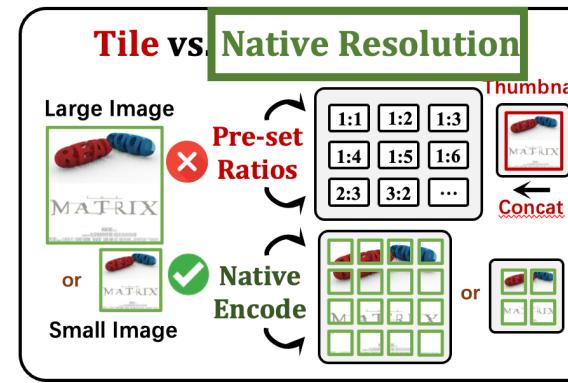
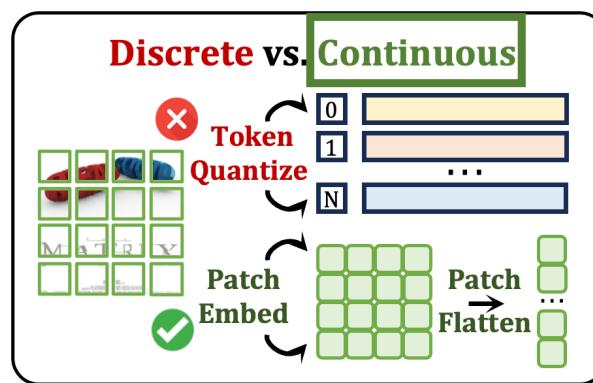
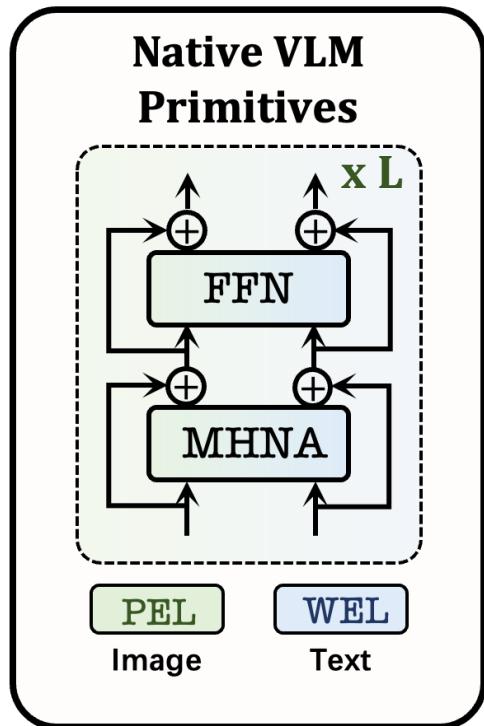
Motivation



From first principles, one native VLM primitive should :

- effectively align pixel and word representations within a shared semantic space;
- seamlessly integrate the strengths of formerly separate vision and language modules;
- inherently embody various cross-modal properties that support unified vision-language encoding, aligning, and reasoning

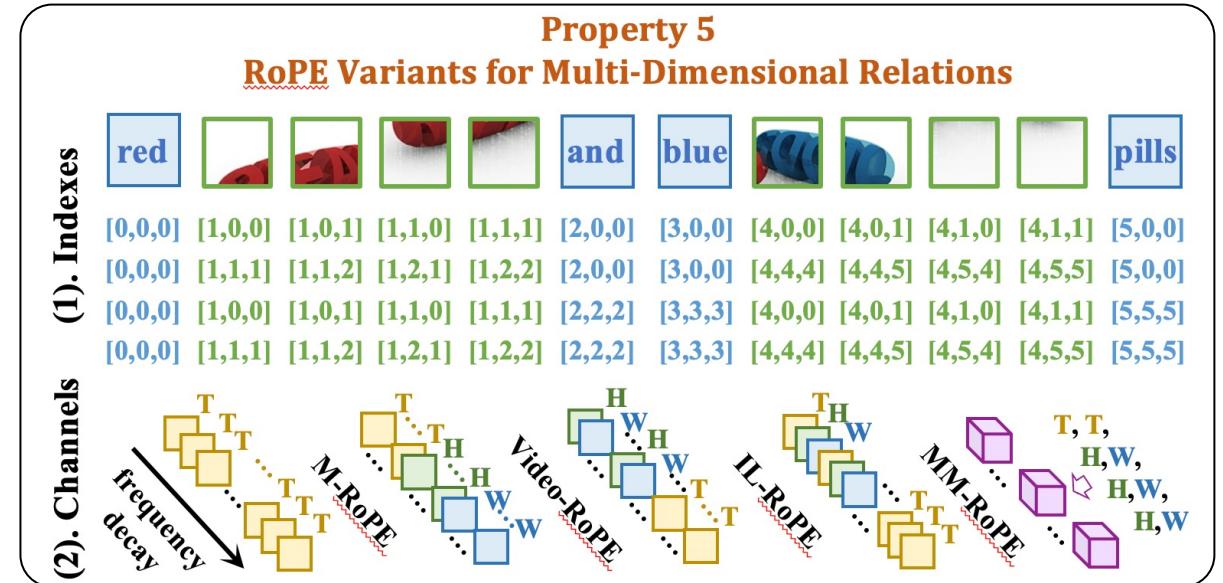
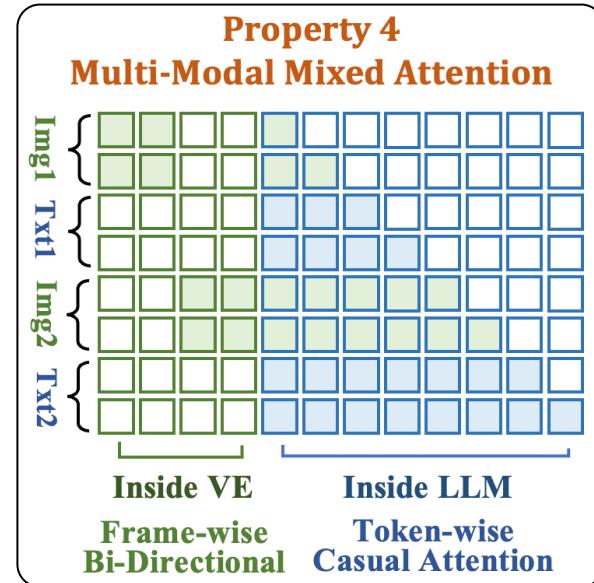
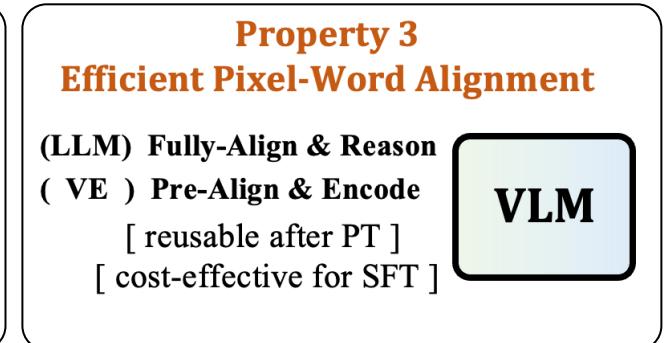
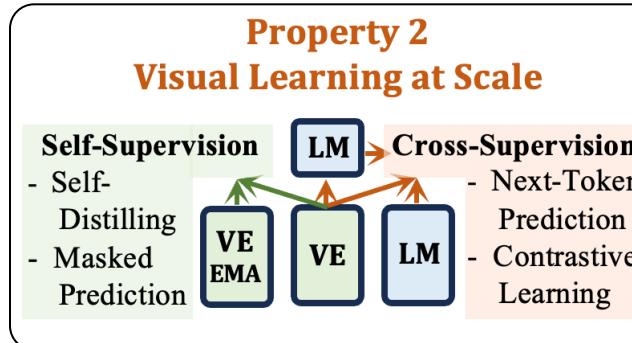
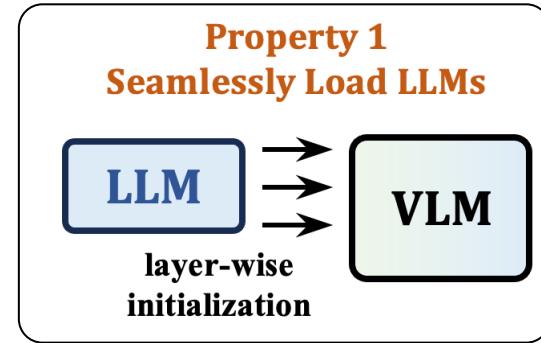
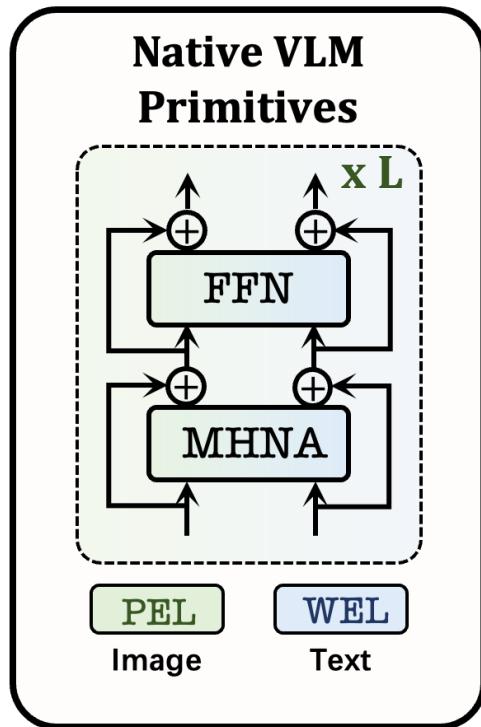
Motivation



Build native VLMs from first principles !!!

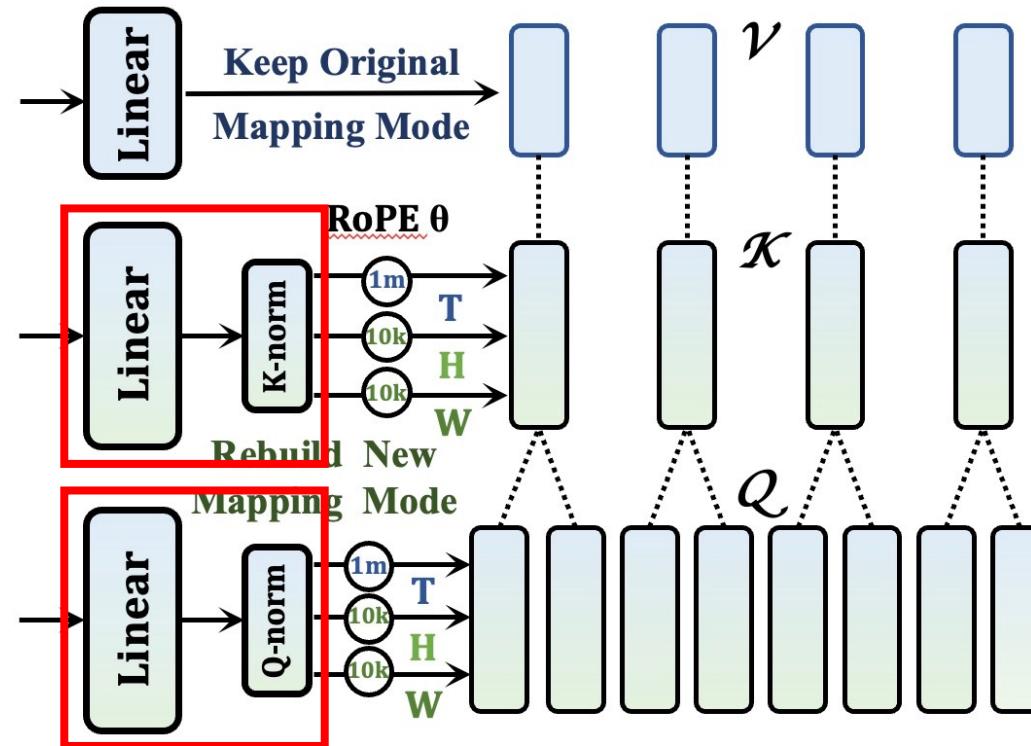
Motivation

Build native VLMs by leveraging the strengths of existing VLM designs !!!

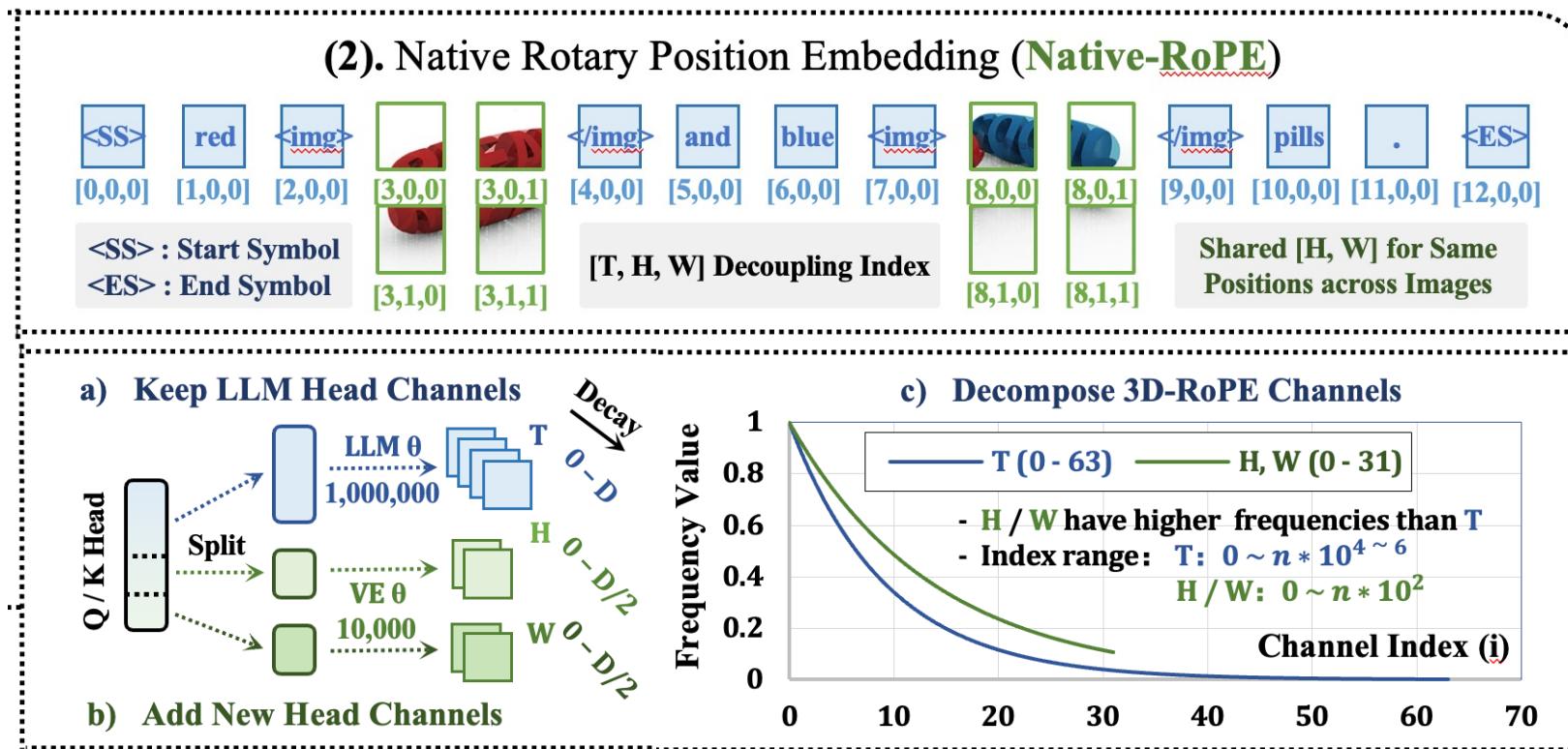


Methodology

(1). Introduce new FC/Norm into original Q, K for H, W



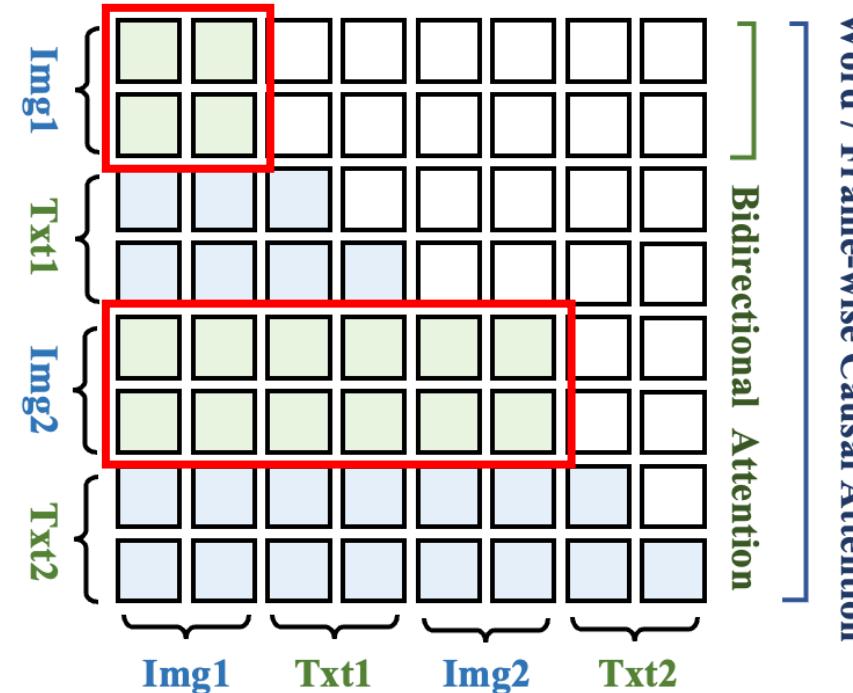
Methodology



- Native Rotary Position Embeddings (Native-RoPE) **eliminates index correlations** and **decouples channel allocation** between H / W and T;
- Native-RoPE with **modality-specific frequencies** captures **local dependencies** across H / W / T and **long-range relations** across T;

Methodology

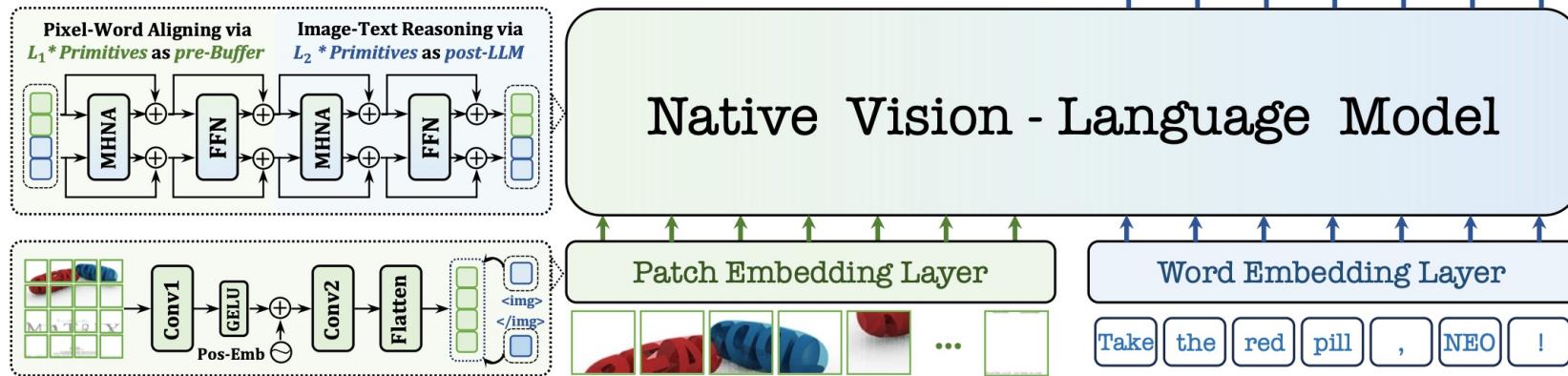
(3). Introduce **Frame-wise Native Multi-Modal Attention**



- Native Multi-Modal Attention captures **rich spatial correspondence within images** and **contextual vision-language dependencies**.

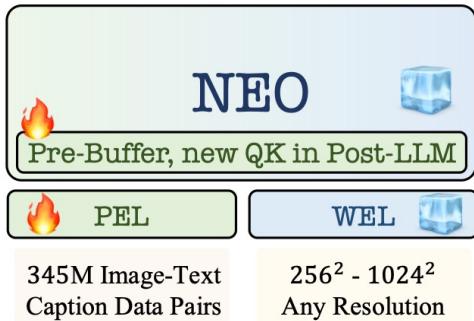
Methodology

(-) Model Architecture

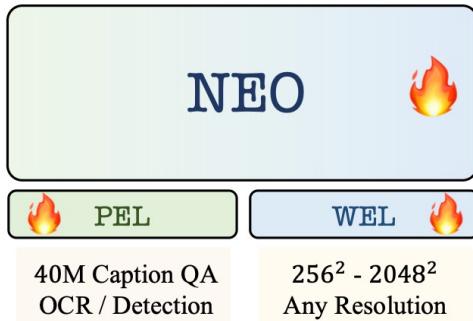


(-) Training Recipe

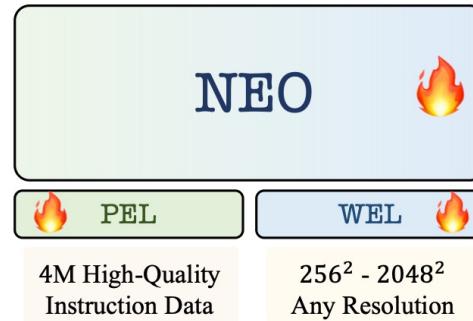
Stage 1: Pre-Training



Stage 2: Mid-Training



Stage 3: Supervised Fine-Tuning



- Modality-shared pre-Buffer maps vision and language into a unified representation space.

Reusable for extensible ecosystem

- Post-LLM absorbs strong language proficiency and powerful reasoning capabilities of pre-trained LLMs.

- End-to-End Training Procedure

- Quite Efficient with Limited Data

With **390M** image-text samples, NEO efficiently **develops visual perception from scratch** while **mitigating vision-language conflicts** inside one model.

Main Results

(-) Evaluation Results on General Understanding

Model	LLM	# Data	MMMU	MMB	MMVet	MMStar	SEED-I	POPE	HallB
▼ Modular Vision-Language Models (2B)									
Qwen2-VL	Qwen2-1.5B	- / - / -	41.1	74.9	49.5	48.0	-	-	41.7
InternVL2.5	InternLM2.5-1.8B	>6B / 100M / 16M	43.6	74.7	60.8	53.7	-	90.6	42.6
Qwen2.5-VL [†]	Qwen2.5-1.5B	- / - / -	51.2	79.1	61.8	55.9	-	-	46.3
InternVL3 [†]	Qwen2.5-1.5B	>6B / 100M / 22M	48.6	81.1	62.2	60.7	-	89.6	42.5
Encoder-Base	Qwen3-1.7B	>6B / 40M / 4M	47.1	75.8	37.4	52.7	73.6	87.0	44.4
▼ Native Vision-Language Models (2B)									
Mono-InternVL	InternLM2-1.8B	1.2B / 143M / 7M	33.7	65.5	40.1	-	67.4	-	34.8
Mono-InternVL-1.5	InternLM2-1.8B	400M / 150M / 7M	39.1	64.0	54.0	-	66.9	-	32.5
HoVLE	InternLM2-1.8B	550M / 50M / 7M	32.2	73.3	43.8	-	70.9	87.4	38.4
OneCAT	Qwen2.5-1.5B	436M / 70M / 13M	39.0	72.4	42.4	-	70.9	-	-
NEO	Qwen3-1.7B	345M / 40M / 4M	48.6	76.0	49.6	54.2	74.2	87.5	43.1
▼ Modular Vision-Language Models (8B)									
Qwen2-VL	Qwen2-7B	- / - / -	54.1	83	62.0	60.7	-	88.1	50.6
InternVL2.5	InternLM2.5-7B	>6B / 50M / 4M	56.0	84.6	62.8	64.4	-	90.6	50.1
Qwen2.5-VL [†]	Qwen2.5-7B	- / - / -	55.0	83.5	67.1	63.9	-	86.4	52.9
InternVL3 [†]	Qwen2.5-7B	>6B / 100M / 22M	62.7	83.4	81.3	68.2	-	91.1	49.9
Encoder-Base	Qwen3-8B	>6B / 40M / 4M	54.1	84	60.0	63.5	76.2	87.8	51.4
▼ Native Vision-Language Models (8B)									
Fuyu	Persimmon-8B	- / - / -	27.9	10.7	21.4	-	59.3	84.0	-
Chameleon	from scratch	1.4B / 0M / 1.8M	25.4	31.1	8.3	-	30.6	19.4	17.1
EVE	Vicuna-7B	33M / 0M / 1.8M	32.6	52.3	25.7	-	64.6	85.0	26.4
SOLO	Mistral-7B	44M / 0M / 2M	-	67.7	30.4	-	64.4	78.6	-
Emu3	from scratch	- / - / -	31.6	58.5	37.2	-	68.2	85.2	-
EVEv2	Qwen2.5-7B	77M / 15M / 7M	39.3	66.3	45.0	-	71.4	87.6	-
BREEN	Qwen2.5-7B	13M / 0M / 4M	42.7	71.4	38.9	51.2	-	-	37.0
VoRA	Qwen2.5-7B	30M / 0M / 0.6M	32.0	61.3	33.7	-	68.9	85.5	-
SAIL	Mistral-7B	512M / 86M / 6M	-	70.1	46.3	53.1	72.9	85.8	54.2
NEO	Qwen3-8B	345M / 40M / 4M	54.6	82.1	53.6	62.4	76.3	88.4	46.4

Main Results

Model	Input	RoPE	Backbone	AI2D	DocVQA	ChartQA	InfoVQA	TextVQA	OCRBench
▼ Modular Vision-Language Models (2B)									
Qwen2-VL	Any Res.	M-RoPE	Dense	74.7	90.1	73.5	65.5	79.7	80.9
InternVL2.5	Tile-wise	1D-RoPE	Dense	74.9	88.7	79.2	60.9	74.3	80.4
Qwen2.5-VL [†]	Any Res.	M-RoPE	Dense	81.6	93.9	84.0	77.1	79.3	79.7
InternVL3 [†]	Tile-wise	1D-RoPE	Dense	78.7	88.3	80.2	66.1	77.0	83.5
Encoder-Base	Tile-wise	1D-RoPE	Dense	77.4	89.9	78.4	65.9	73.3	83.5
▼ Native Vision-Language Models (2B)									
Mono-InternVL	Tile-wise.	1D-RoPE	MoE	68.6	80.0	73.7	43.0	72.6	76.7
Mono-InternVL-1.5	Tile-wise.	1D-RoPE	DaC	67.4	81.7	72.2	47.9	73.7	80.1
HoVLE	Tile-wise.	1D-RoPE	Dense	73.0	86.1	78.6	55.7	70.9	74.0
OneCAT	Any Res.	M-RoPE	Dense	72.4	87.1	76.2	56.3	67.0	—
NEO	Any Res.	Native-RoPE	Dense	80.1	89.9	81.2	63.2	74.0	77.1
▼ Modular Vision-Language Models (8B)									
Qwen2-VL	Any Res.	M-RoPE	Dense	83.0	94.5	83	76.5	84.3	86.6
InternVL2.5	Tile-wise	1D-RoPE	Dense	84.5	93.0	84.8	77.6	79.1	82.2
Qwen2.5-VL [†]	Any Res.	M-RoPE	Dense	83.9	95.7	87.3	82.6	84.9	86.4
InternVL3 [†]	Tile-wise	1D-RoPE	Dense	85.2	92.7	86.6	76.8	80.2	88
Encoder-Base	Tile-wise	1D-RoPE	Dense	82.9	92.1	83.5	75	77.1	85.3
▼ Native Vision-Language Models (8B)									
Fuyu	Any Res.	1D-RoPE	Dense	64.5	—	—	—	—	36.6
Chameleon	Fix Res.	1D-RoPE	Dense	46.0	1.5	2.9	5.0	4.8	0.7
EVE	Any Rat.	1D-RoPE	Dense	61.0	53.0	59.1	25.0	56.8	39.8
SOLO	Any Res.	1D-RoPE	Dense	61.4	—	—	—	—	12.6
Emu3	Fix Res.	1D-RoPE	Dense	70	76.3	68.6	43.8	64.7	68.7
EVEv2	Any Rat.	1D-RoPE	DaC	74.8	—	73.9	—	71.1	70.2
BREEN	Any Res.	1D-RoPE	MoE	76.4	—	—	—	65.7	—
VoRA	Any Res.	1D-RoPE	Dense	61.1	—	—	—	58.7	—
SAIL	Any Res.	M-RoPE	Dense	76.7	—	—	—	77.1	78.3
NEO	Any Res.	Native-RoPE	Dense	83.1	88.6	82.1	60.9	75.0	77.7

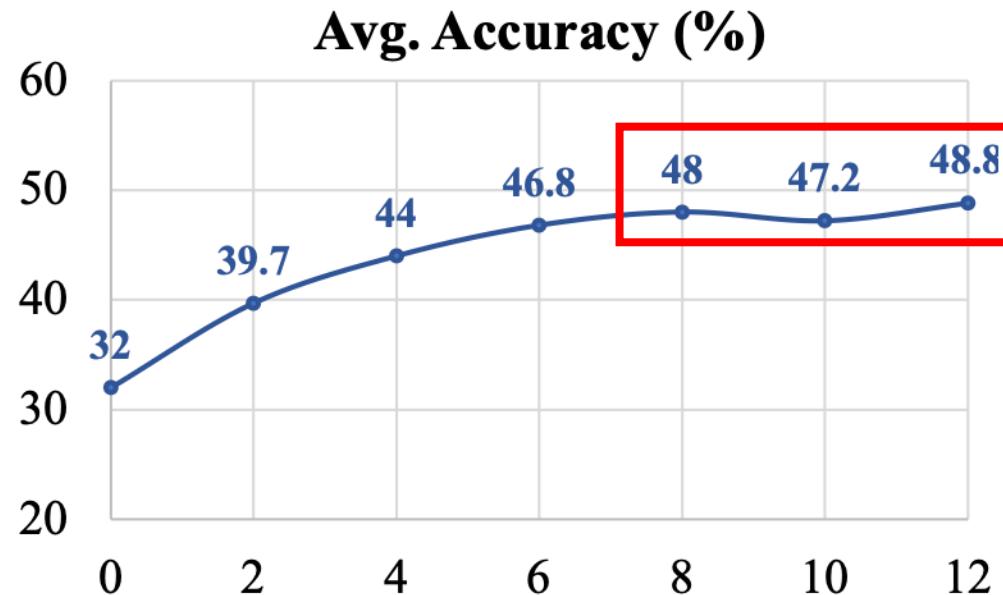
-- With quite **limited** pre-training and supervised fine-tuning data and

-- Without reinforcement learning (**RL**)

▪ **Approaches** the performance of top-tier modular VLMs, e.g., **Qwen2 / 2.5-VL, InternVL2.5 / 3.**

▪ **Delivers substantial gains** on diverse visual-centric benchmarks over the best competitors, from **EVE** series to **SAIL**.

Ablation Studies



Here **8-12** primitive layers for pre-Buffer is a good trade-off for pre-alignment.

Figure 5: Configurations of pre-Buffer.

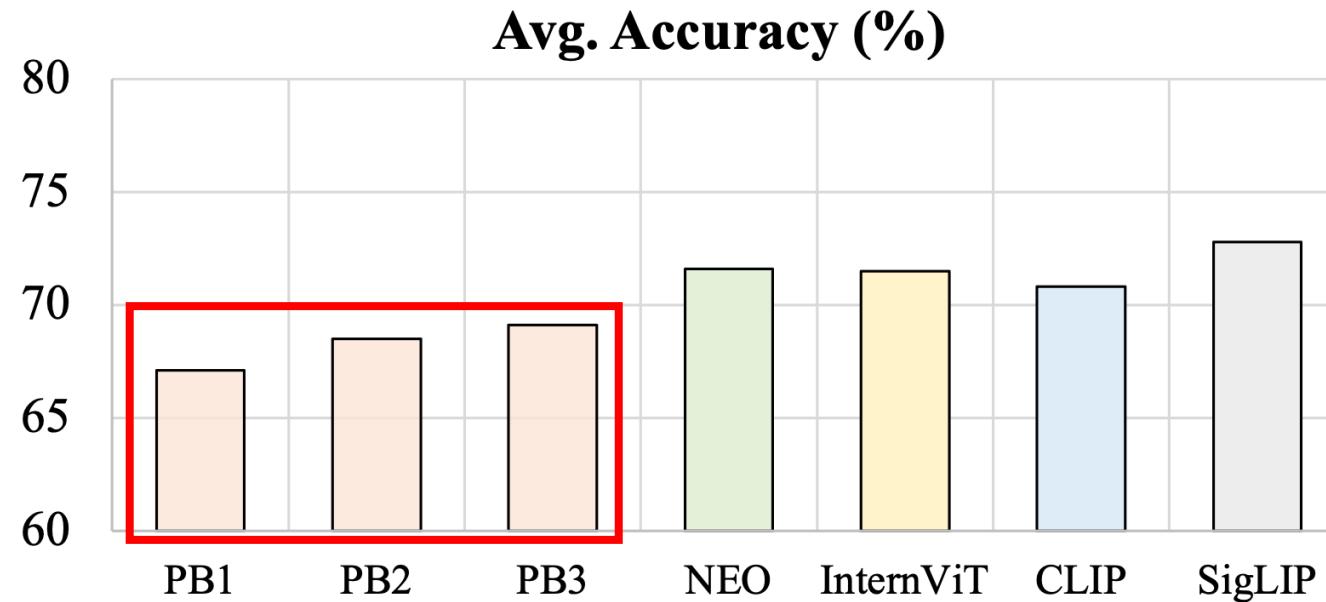
Ablation Studies

Table 3: Configurations of attention and RoPE. MMS, CQA, IVQA, and OCRB denote MMStar, ChartQA, InfoVQA, and OCRCbench. \star indicates that the base RoPE frequencies for height and width are set to 1M. To ensure fairness, we add new head dimensions of equal size across all models.

Model	Attention	RoPE	MMMU	MMB	MMS	SEED-I	AI2D	CQA	IVQA	TVQA	OCRB	POPE	Avg.
A	Causal	1D-RoPE	40.2	48.6	36.1	55.3	63.6	16.1	22.5	16.2	13.9	78.6	39.1
B	Mixed	1D-RoPE	40.8	48.8	36.4	57.3	63.7	16.0	21.9	17.4	16.0	79.2	39.8
C	Mixed	IL-RoPE	40.0	47.3	36.3	57.6	62.0	18.8	23.4	17.9	13.2	78.8	39.5
D	Mixed	M-RoPE	40.3	49.6	37.2	57.8	64.2	23.7	25.2	20.4	18.8	79.3	41.7
E	Mixed	MM-RoPE	40.5	50.8	37.6	58.2	65.8	25.7	26.3	22.1	18.2	78.8	42.4
F	Mixed	Video-RoPE	40.6	51.3	37.8	58.8	64.3	27.4	26.1	23.7	21.3	81.0	43.2
G	Causal	Native-RoPE	40.2	49.2	36.3	57.1	63.7	19.2	23.5	19.5	16.7	77.8	40.3
H	Mixed	Native-RoPE	40.7	51.9	38.2	58.9	65.8	30.6	26.9	24.1	23.2	80.0	44.0
I	Mixed	Native-RoPE \star	40.4	50.4	36.9	57.0	64.1	25.6	25.2	21.7	20.1	78.7	42.0

- Modality-specific RoPE frequency does count !
- RoPE indexes allocation for H, W, T does count !
- Mixed Multi-Modality Attention Mechanism does count !

Ablation Studies



PB 1–3 denotes the Pre-Buffer after stage 1–3.

PB3 shows only an average gap of 2.5 / 2.4 / 1.7 / 3.7% over NEO / InternViT / CLIP / SigLIP, reducing training costs of building native VLMs for subsequent research.

Native Multimodal Post-Training

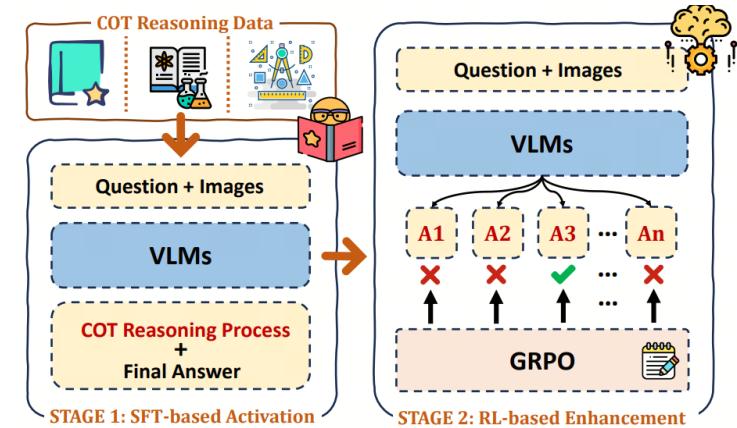
Visual Jigsaw Post-Training Improves MLLMs

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, Ziwei Liu

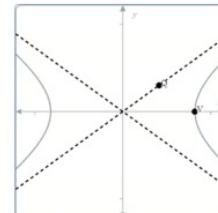
Motivation

RL-based Post-training for MLLMs

- **Text-centric reasoning on math/science/coding problems**
- Specific vision tasks (grounding, detection, segmentation, counting)
- Tool-using (thinking with images)



Most works focused on enhancing **text-centric reasoning** where **visual inputs work only as context**



Question: Given the hyperbola shown in the image, can we determine the equation of the hyperbola by analyzing the relationship between the slope of the asymptote and the distance from the center to the vertex?
Choices: A. Yes B. No

Reason-RFT:

<think>Let's analyze the image. The hyperbola is shown with two asymptotes intersecting at the center. To find the equation of the hyperbola, we need to consider the relationship between the slope of the asymptote and the distance from the center to the vertex. The equation of a hyperbola with its center at the origin and aligned with the coordinate axes is given by $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$, where a and b are the distances from the center to the vertices and asymptotes, respectively. The center to the vertices is given, which suggests that $a^2 + b^2$ is given. The slope of the asymptotes is given by $\pm \frac{b}{a}$. In this case, the slope is 3, which implies that $\frac{b}{a} = 3$. Therefore, the equation of the hyperbola can be determined by these relationships.</think>

<answer>A</answer>

[1] Tan, Huajie, et al. "Reason-rft: Reinforcement fine-tuning for visual reasoning." *arXiv preprint arXiv:2503.20752* (2025).

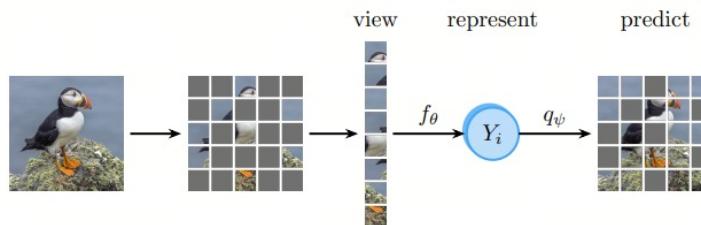
Motivation

How to improve intrinsic vision-centric capabilities of MLLMs?

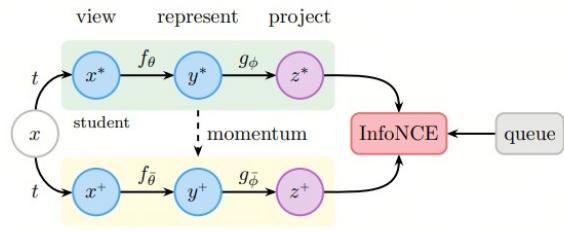
- Methods like ROSS [1] shows dense image reconstruction helps understanding, but requiring additional vision generation modules and designs. Do we need dense pixel-level reconstruction?
- Unified Multimodal Models (UMMs) only shows **understanding benefits visual generation**

How do we learn good vision representation?

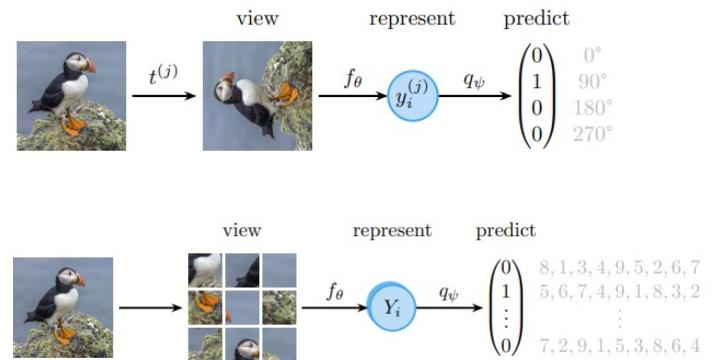
Self-supervised learning!



Reconstruction-based methods



Discriminative/contrastive methods



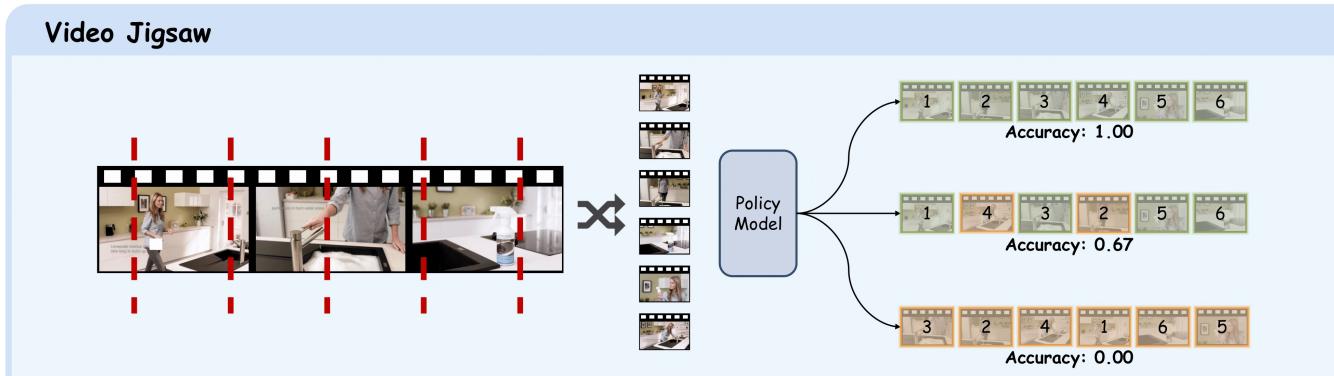
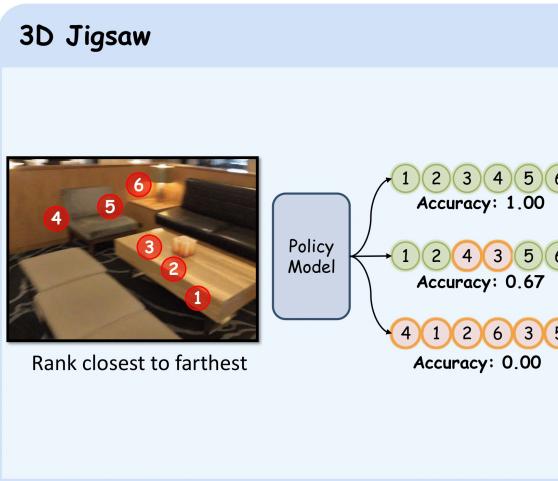
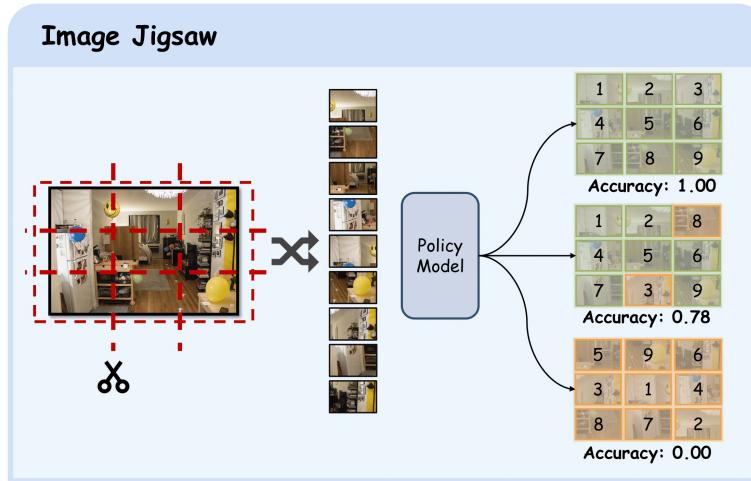
Other pretext tasks like rotation prediction and **jigsaw-style** tasks

Easier version of reconstruction
Suitable for MLLMs
(compatible with text-output MLLM)

[1] Wang, Haochen, et al. "Reconstructive visual instruction tuning." *arXiv preprint arXiv:2410.09575* (2024).

[2] Uelwer, Tobias, et al. "A survey on self-supervised representation learning." *arXiv preprint arXiv:2308.11455* (2023).

Visual Jigsaw



Visual Data → Partitioning → Shuffling

Model reconstruct the data by predicting the indices in correct order

Optimize using the GRPO algorithm

Image Jigsaw

Image → $3 * 3$ image patches

Mentally reconstruct the image and output the patch indices in the correct raster scan order.



1 2 3 4 5 6 7 8 9



1 2 3 4 5 6 7 8 9



1 2 3 4 5 6 7 8 9

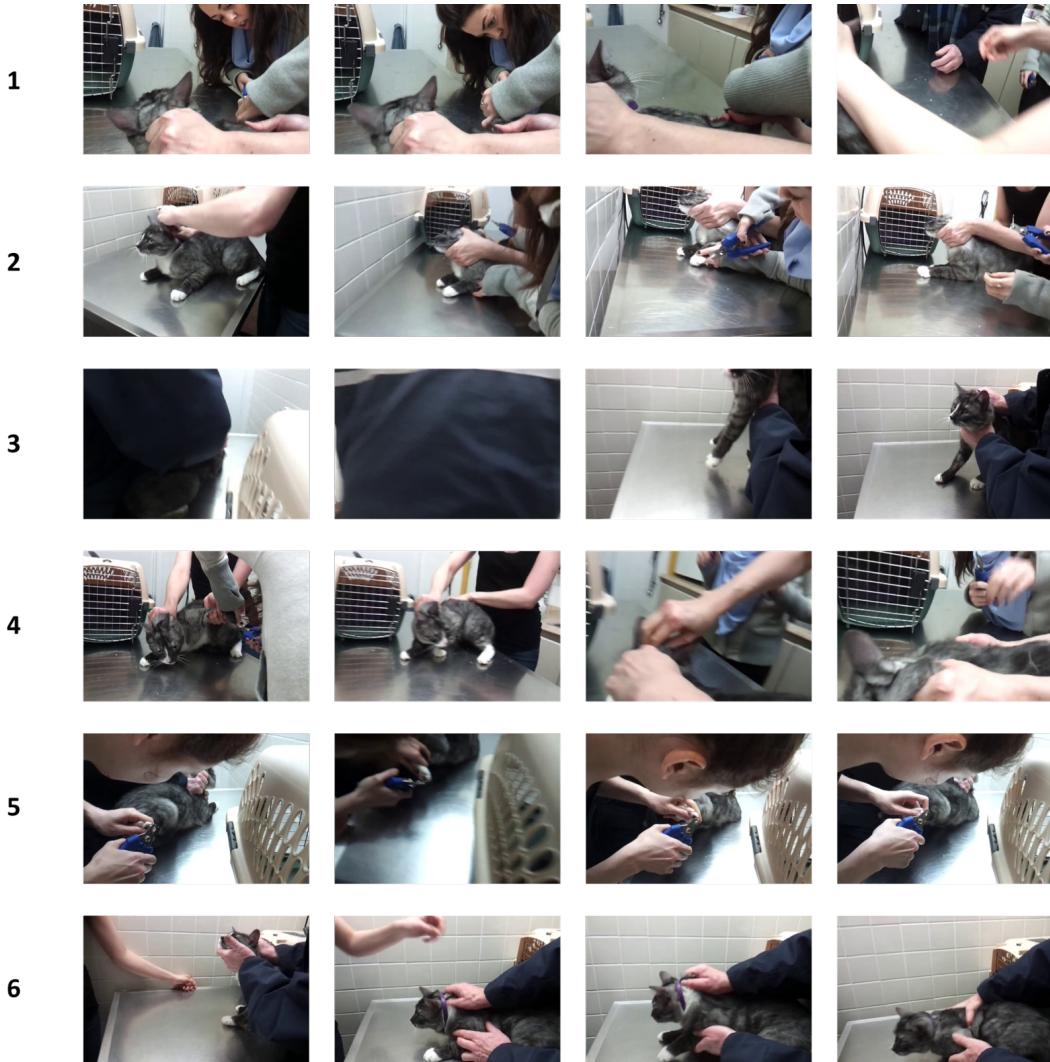
Image Jigsaw

Model	Fine-grained Perception & Understanding										Spatial Und (Mono)		Compositional Und	
	MMVP	MMStar (fine-grained)	MMBench	HR-Bench-8K	V*	MME-RealWorld	LISA-Grounding	OVD-Eval	VSR	OmniSpatial	DA-2K	Winoground	SugarCrepe++	
	test	fine	en_dev	test	test	lite	test	test	test	test	val	g-acc	test	
ThinkLite-VL	55.33	59.95	84.19	68.12	76.96	46.17	73.70	35.78	78.09	42.60	58.46	35.25	61.49	
VL-Cogito	55.33	56.64	82.98	69.62	79.58	47.63	72.26	35.78	79.82	44.29	56.43	38.25	63.59	
LLaVA-Critic-R1	53.33	57.80	83.16	67.50	78.01	45.18	68.52	35.28	78.50	42.73	53.82	34.75	61.93	
Qwen2.5-VL-7B	54.66	59.75	83.33	67.38	76.96	43.41	71.89	35.07	77.68	42.66	54.45	37.00	61.59	
Image Jigsaw (SFT)	56.00	60.94	83.67	69.75	80.10	43.88	66.59	34.35	80.68	43.55	61.46	38.75	62.03	
Image Jigsaw	60.66	65.81	84.45	71.13	80.63	45.96	74.54	36.49	80.36	44.49	60.35	39.00	63.02	
(Gain)	+6.00	+6.06	+1.12	+3.75	+3.66	+2.55	+2.65	+1.42	+2.68	+1.83	+5.90	+2.00	+1.43	

Enhance vision-centric capabilities:

- Fine-grained perception & understanding
- Monocular spatial understanding
- Compositional visual understanding

Video Jigsaw



Video → 6 video clips

Mentally reconstruct the video and output the clip indices in the correct chronological order.

Video Jigsaw

Model	Frames	AoTBench		Vinoground		TOMATO		FAVOR-Bench		TUNA-Bench		VideoMME		TempCompass		TVBench		MotionBench		LVBench		VSI-Bench		Video-TT		CVBench	
		vqa	group	test	test	test	wo subs	mc	test	val	test	test	mcq	test	test	val	test	val	test	test	mcq	test	test	mcq	test		
Video-R1	16	45.06	9.40	27.29	49.47	53.00	56.62	70.19	51.80	55.82	34.53	34.34	42.95	47.50													
Video-R1	32	47.53	10.20	27.29	49.90	54.26	59.88	71.77	53.54	56.12	38.61	35.11	42.63	48.10													
Video-R1	64	48.68	10.60	27.36	50.51	54.33	60.85	72.59	53.43	56.09	38.80	36.61	42.74	48.69													
Qwen2.5-VL-7B	16	45.52	12.60	25.87	48.54	53.14	57.44	71.77	49.94	55.56	33.51	32.79	38.39	47.70													
Qwen2.5-VL-7B	32	49.48	18.20	26.34	49.34	54.88	60.70	72.59	51.96	56.47	39.19	35.34	41.57	49.60													
Qwen2.5-VL-7B	64	52.41	21.80	26.35	50.86	55.79	63.44	72.84	53.74	56.29	40.35	37.74	42.25	51.50													
Video Jigsaw	16	51.67	15.20	27.56	49.69	55.10	58.07	73.10	51.33	56.87	36.41	35.39	40.19	49.80													
(Gain)		+6.15	+2.60	+1.69	+1.15	+1.96	+0.63	+1.33	+1.39	+1.31	+2.90	+2.60	+1.80	+2.10													
Video Jigsaw	32	55.00	21.40	28.03	50.56	56.49	62.37	73.60	53.31	57.99	39.70	38.47	43.27	51.60													
(Gain)		+5.52	+3.20	+1.69	+1.22	+1.61	+1.67	+1.01	+1.35	+1.52	+0.51	+3.13	+1.70	+2.00													
Video Jigsaw	64	57.64	25.20	28.30	52.27	56.63	64.74	73.60	54.18	57.91	41.83	40.40	44.11	54.50													
(Gain)		+5.23	+3.40	+1.95	+1.41	+0.84	+1.30	+0.76	+0.44	+1.62	+1.48	+2.66	+1.86	+3.00													

- Enhances general video perception and comprehension
- Large gain on temporal-centric understanding and reasoning about temporal directionality (e.g. AoTBench)
- Improved cross-video understanding and reasoning (CVBench)

Video Jigsaw

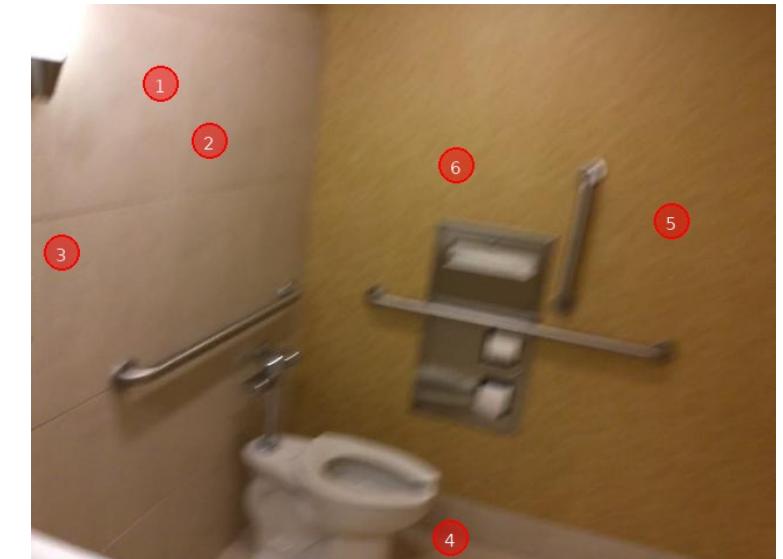
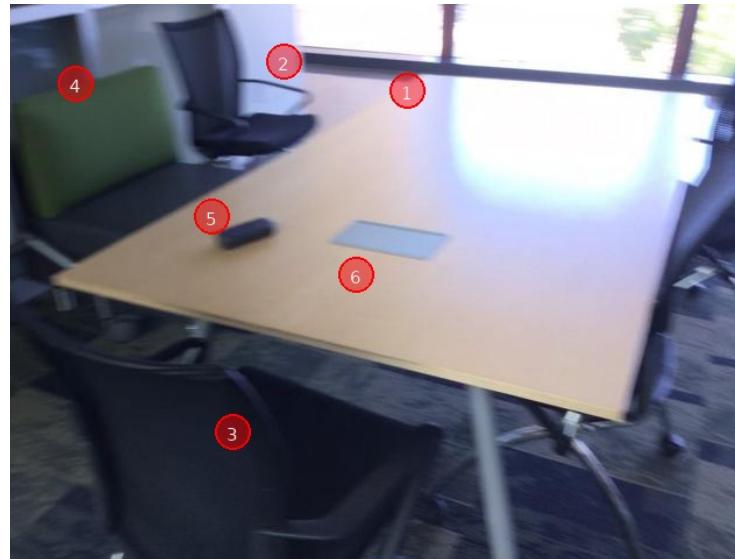
Model	AoTBench	Vinoground	TOMATO	FAVOR-Bench	TUNA-Bench	VideoMME	TempCompass	TVBench	MotionBench	LVBench	VSI-Bench	Video-TT	CVBench
	vqa	group	test	test	test	wo subs	mc	test	val	test	test	mcq	test
MiMo-VL-7B-SFT-2508	65.00	15.60	34.16	53.33	58.80	68.07	76.13	56.48	57.86	40.86	41.59	46.00	63.00
Video Jigsaw	69.77	21.60	37.33	54.31	62.29	68.55	77.21	61.50	59.73	42.93	44.27	48.50	65.20
(Gain)	+4.77	+6.00	+3.17	+0.98	+3.49	+0.48	+1.08	+5.02	+1.87	+2.07	+2.68	+2.50	+2.20

- Consistent improvement on stronger base model: MiMo-VL-7B-SFT-2508

3D Jigsaw

RGB-D → 6 points

Order the points from closest to farthest relative to the camera.



3D Jigsaw

Model	SAT-Real	3DSRBench	ViewSpatial	All-Angles	Omnispatial	VSI-Bench	SPARBench	DA-2K
	test	test	test	test	test	test	tiny	test
Qwen2.5-VL-7B	48.66	57.42	36.52	47.56	42.66	37.74	35.75	54.45
3D Jigsaw <i>(Gain)</i>	64.00 +15.34	58.13 +0.71	38.62 +2.10	49.06 +1.50	45.99 +3.33	40.64 +2.90	38.31 +2.56	71.56 +17.11

- Largest gains on directly related task – DA-2K
- Consistent improvements on a wide range of other tasks (single-view, multi-views, egocentric video)

Ablation Studies

- RL outperforms SFT
- The difficulty of the jigsaw tasks matters
- Apply jigsaw task training before text-centric/long CoT reasoning training

Future Works

- Different 3D jigsaw designs on base models with stronger 3D capabilities
- Different jigsaw configurations and combinations
- Other **vision-centric** self- and weakly-supervised tasks

Native Multimodal Evaluation

RealUnify: Do Unified Models Truly Benefit from Unification?

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, Zuyan Liu, Bohan Zeng, Ruizhe Chen, Qixun Wang, Zhuoran Zhang, Xinlong Chen, Chengzhuo Tong, Bozhou Li, Chaoyou Fu, Qiang Liu, Haotian Wang, Wenjing Yang, Yuanxing Zhang, Pengfei Wan, Yi-Fan Zhang, Ziwei Liu

Motivation

Specific

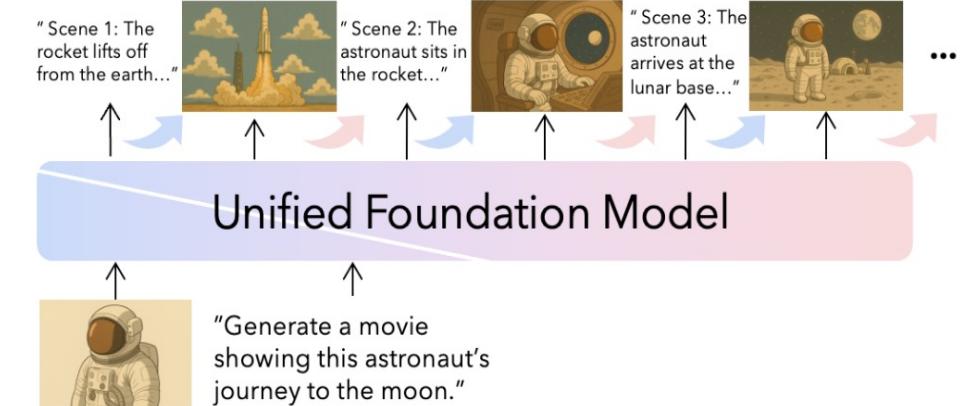
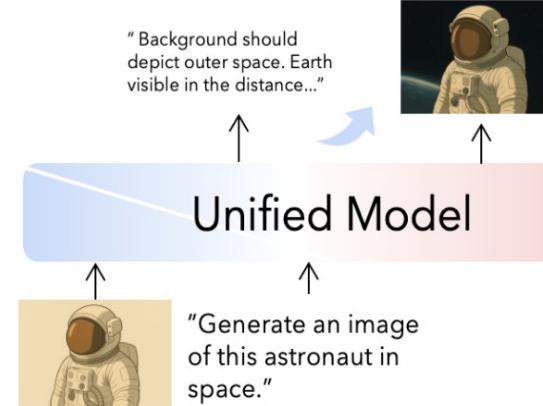
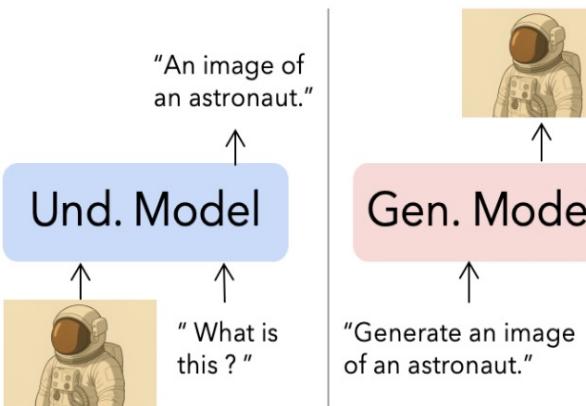
Combine

Emergent

Stage1. Specific

Stage2. Combine

Stage3. Emergent

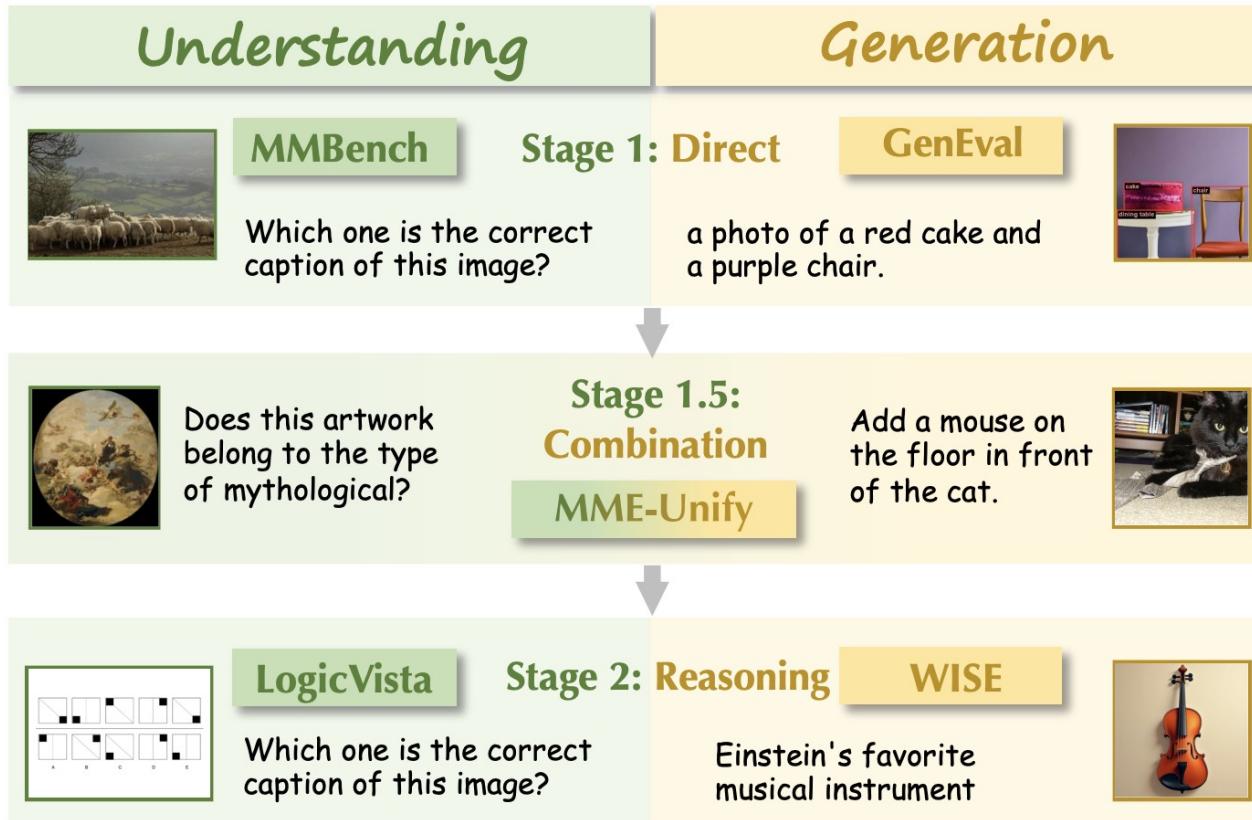


Easy to evaluate with current benchmarks

Lack of customized benchmarks

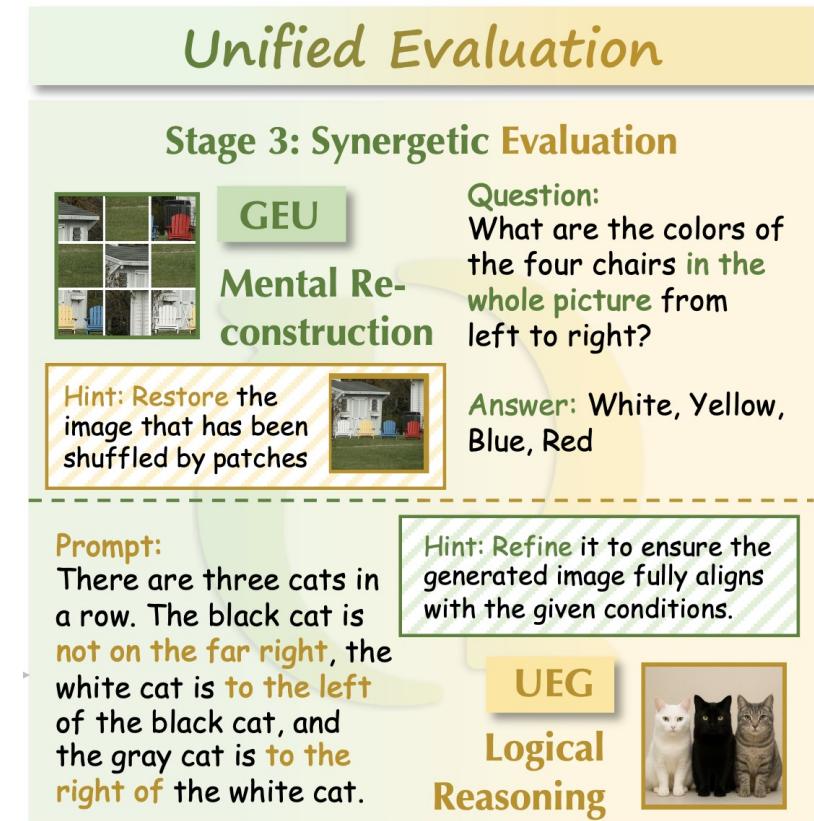
Motivation

Previous Benchmarks



Direct -> Combination -> Reasoning

Customized Unified Benchmark



True Unification

Task Taxonomy

Understanding -> Generation (UEG)

UNDERSTANDING ENHANCES GENERATION (UEG)

World Knowledge	
User	Prompt: The largest feline animal in terms of body size.
Evaluation	Question: 1. Does this image show a Tiger? ✓ 2. Can you identify a Tiger in this image? ✓
Mathematical Reasoning	
User	Prompt: A table with some books, which could be arranged into four stacks with 2 books each. Draw all the books on the table.
Evaluation	Question: 1. Are there a total of exactly 8 books on the table? ✓
Scientific Reasoning	
User	Prompt: A litmus solution is exposed to a carbon dioxide (CO_2) environment.
Evaluation	Question: 1. Is the solution red? ✓ 2. Is the solution blue? ✗ 3. Is the solution purple? ✗
Commonsense Reasoning	
User	Prompt: A slice of butter melting unevenly on hot toast.
Evaluation	Question: 1. Is there a slice of butter present? ✓ 2. Is the butter placed on a toast? ✓ 3. Is the butter shown melting unevenly rather than in a uniform manner? ✓
Logical Reasoning	
User	Prompt: Three birds, one blue and one gray, are lined up on a telephone pole. The blue bird is not in the middle, and the adjacent birds are different colors.
Evaluation	Question: 1. Is the blue bird not in the middle? ✓ 2. Are the adjacent different colors? ✓
Code To Image	
User	Prompt: Code: num = int(input()) if num > 0: print("A pair of shoes") elif num < 0: print("A pink pig rolling in the mud.") else: print("A fluffy sheep with a bell around its neck.")
Evaluation	Given the input: 0, generate the image based on the output of the code execution. Question: 1. Does the image show a fluffy sheep with a bell around its neck? ✓

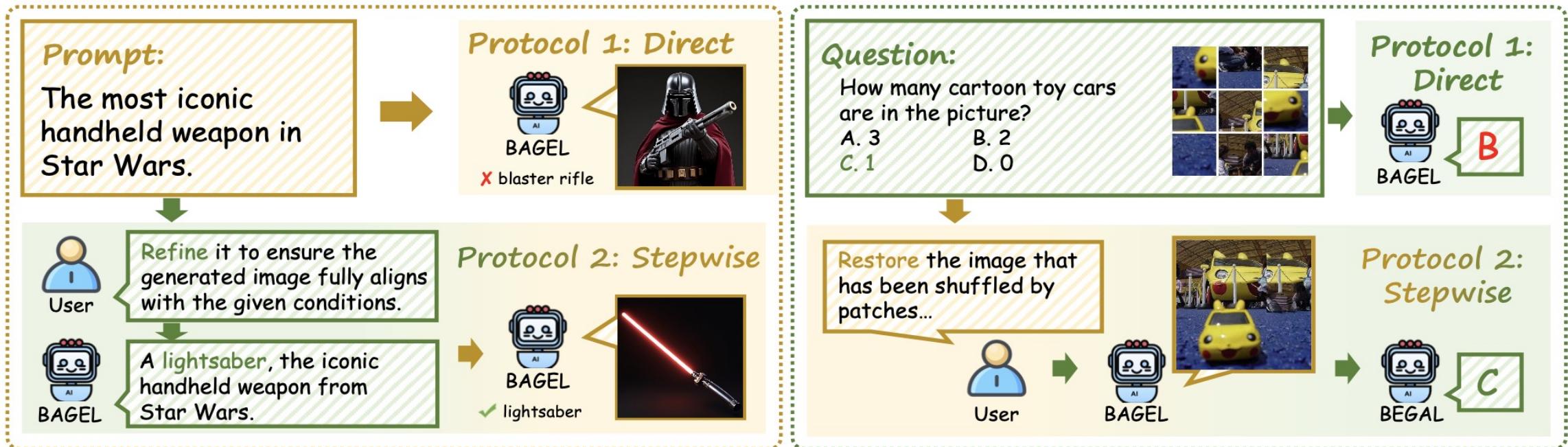
Hint for UEG: Refine it to ensure the generated image fully aligns with the given conditions.

Generation -> Understanding (GEU)

Mental Reconstruction	
User	Question: Including the photographer, how many cars are there in the photo? A. 3 B. 1 C. 4 D. 2
Evaluation	Hint: Restore the image that has been shuffled by patches...
Mental Tracking	
User	Question: Turn all black segments into orange, then turn all yellow into orange, then turn all green into red. Which digits are formed by the orange segments? A. "1,7" B. "4,7". C. "4,6". D. "7,6"
Evaluation	Hint: Apply the transformations to the contents of the image.
Attentional Focusing	
User	Question: What is the text written on the blue golf ball holder? A. MUTUAL INSURANC B. NEW YORK MUTUAL C. NEW MEXICO MUTUAL D. NEW MEXICO INSURANCE
Evaluation	Hint: Highlight the regions of the image that are relevant to the question.
Cognitive Navigation	
User	Question: On the shortest path from Penguin to Polar, which of the following animals can we see?? A. Rabbit B. Monkey C. Aviary D. Lion
Evaluation	Hint: Mark the path(s) in the image that are relevant to the question.

How to Evaluate

Direct -> Step-wise



Direct: Whether the model can leverage generation & understanding synergistically

Step-Wise: Decouple generation & understanding for better assignment

Evaluate with RealUnify

Model	Understanding Enhances Generation							Generation Enhances Understanding				Total
	WK	CR	MR-I	LR	SR	C2I	Avg	MR-II	MT	AF	CN	
<i>Proprietary Models</i>												
Nano Banana	89 / -	86 / -	34 / -	65 / -	48 / -	56 / -	63.0 / -	34 / -	27 / -	36 / -	30 / -	31.8 / -
<i>Open-Source Unified Models</i>												
MIO	24 / 35	26 / 33	18 / 13	9 / 10	10 / 11	0 / 8	14.5 / 18.3	26 / 23	19 / 18	35 / 19	23 / 21	25.8 / 20.3
Janus-Pro	25 / 26	77 / 71	16 / 7	13 / 17	16 / 20	3 / 10	25.0 / 25.2	21 / -	23 / -	28 / -	29 / -	25.3 / -
ILLUME+	44 / 52	62 / 62	22 / 22	23 / 25	26 / 26	1 / 7	29.7 / 32.3	27 / 27	19 / 20	35 / 38	30 / 25	27.8 / 27.5
Show-o2	30 / 42	56 / 50	25 / 25	21 / 21	18 / 20	18 / 19	28.0 / 29.5	36 / -	28 / -	36 / -	21 / -	30.3 / -
OmniGen2	36 / 55	61 / 60	21 / 26	29 / 28	16 / 20	19 / 6	30.3 / 32.5	30 / 42	21 / 24	51 / 38	28 / 19	32.5 / 30.8
UniPic2	61 / 62	73 / 72	31 / 30	28 / 38	25 / 26	7 / 15	37.5 / 40.5	26 / 28	20 / 24	27 / 27	23 / 16	24.0 / 23.8
UniWorld-V1	51 / 56	64 / 59	26 / 26	33 / 37	21 / 24	15 / 9	35.0 / 35.2	29 / 33	19 / 25	57 / 36	24 / 20	32.3 / 28.5
Ovis-U1	37 / 59	72 / 71	28 / 30	23 / 34	15 / 17	12 / 25	31.2 / 39.3	32 / 38	28 / 25	60 / 31	36 / 24	39.0 / 29.5
BLIP3-o	57 / 62	71 / 74	21 / 24	19 / 25	28 / 22	2 / 9	33.0 / 36.0	36 / -	25 / -	57 / -	32 / -	37.5 / -
OneCAT	61 / 64	70 / 65	32 / 20	29 / 27	24 / 31	9 / 27	37.5 / 39.0	26 / 29	25 / 26	43 / 26	31 / 36	31.3 / 29.3
BAGEL	46 / 74	70 / 80	23 / 26	29 / 37	21 / 29	7 / 40	32.7 / 47.7	37 / 38	31 / 25	50 / 52	39 / 28	39.3 / 35.8
												35.3 / 42.9

(a) Understanding Enhances Generation (UEG)

Model	WK	CR	MR-I	LR	SR	C2I	Total
<i>Specialized Models</i>							
GPT-Image-1	90	87	31	69	48	48	62.2
Qwen-Image	66	83	28	44	25	67	52.2
FLUX.1 Kontext	53	73	25	27	25	37	40.0
<i>Unified Models</i>							
Nano Banana	89	86	34	65	48	56	63.0
UniPic2	61	73	31	28	25	7	37.5
OneCAT	61	70	32	29	24	9	37.5

(b) Generation Enhances Understanding (GEU)

Model	MR-II	MT	AF	CN	Total
<i>Specialized Models</i>					
Gemini 2.5 Pro	30	73	73	43	54.8
GPT-4.1	38	23	56	37	38.5
Qwen2.5-VL	35	23	44	36	34.5
<i>Unified Models</i>					
BAGEL	37	31	50	39	39.3
Ovis-U1	32	28	60	36	39.0
BLIP3-o	36	25	57	32	37.5

12 SOTA models evaluated on RealUnify:

- UEG & GEU remain challenging
- Step-wise is better than direct answer
- All models lack true unification

Comparison with SOTA specialist:

- Unify models benefit from understanding
- Generation may not help understanding currently

How Far Can We?

Table 5: Comparisons with Gen-Und SOTA.

Model	WK	CR	MR-I	LR	SR	C2T	Total
Nano Banana	89	86	34	65	48	56	63
Und→Gen (SOTA)	93	86	43	70	53	91	72.7
Model	MR-II		MT	AF		CN	Total
BAGEL	37		31	50	39	39.3	
Gen→Und (SOTA)	29		27	21	50	31.8	

Comparison with Oracle Setting:

- Current unified models can still learn from oracle cases -> **Strong understanding leads to improved generation**
- Both unified models and oracle settings fall short on GEU tasks -> **Current generation models fall short in aiding real-world problem-solving.**

Error Analysis Unified models fall short in real-world image generation



Attribute Entanglement (rabbits and chickens)



Quantity Accuracy (8 flowers)



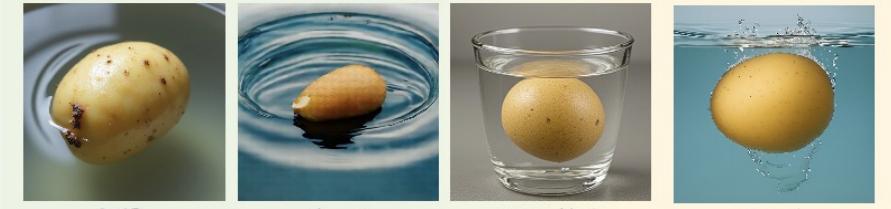
Attribute Fidelity (Olympic Rings)



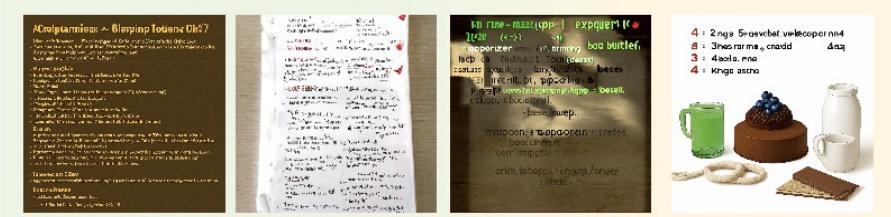
Positional Alignment (the green book cannot be on the far right)



Fine-grained Detail (hands and fingers)



Physical Law (the potato should sink to the bottom)



Text Distortion (font distortion, warping, and meaningless content)



Object Misclassification (the right side should be a lioness)



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

S-LAB
FOR ADVANCED
INTELLIGENCE

Thank You

Ziwei Liu 刘子纬

Nanyang Technological University

