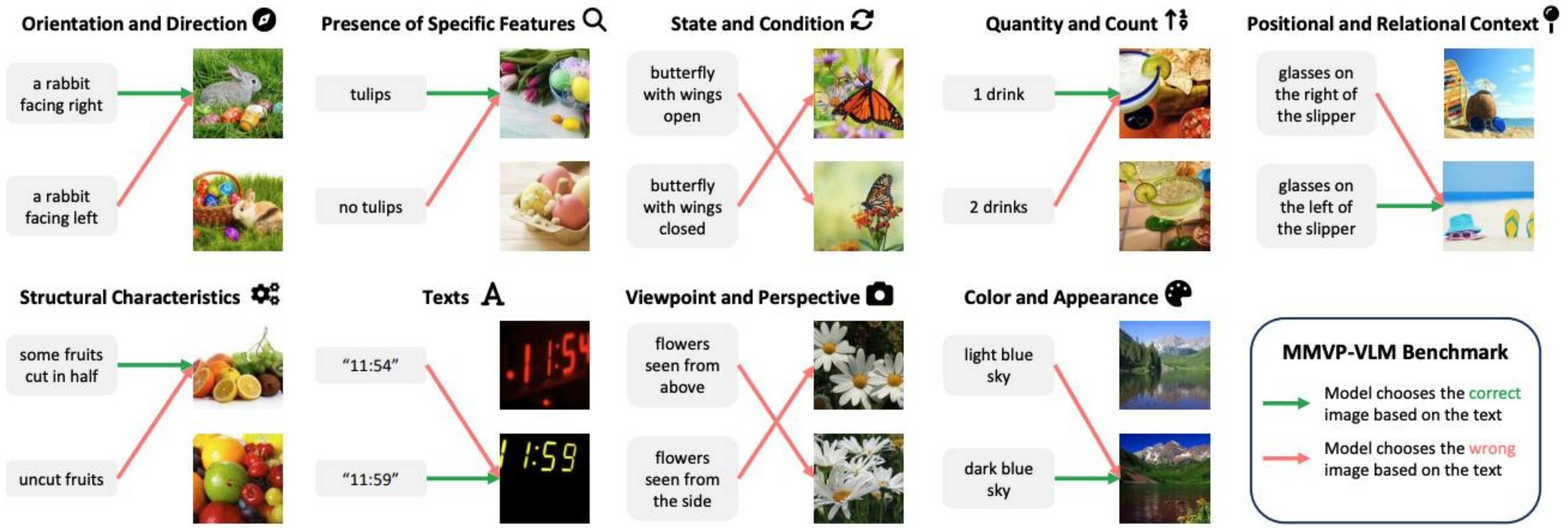
VITRIX-CLIPIN: Enhancing Fine-Grained Visual Understanding in CLIP via Instruction Editing Data and Long Captions

NeurIPS 2025

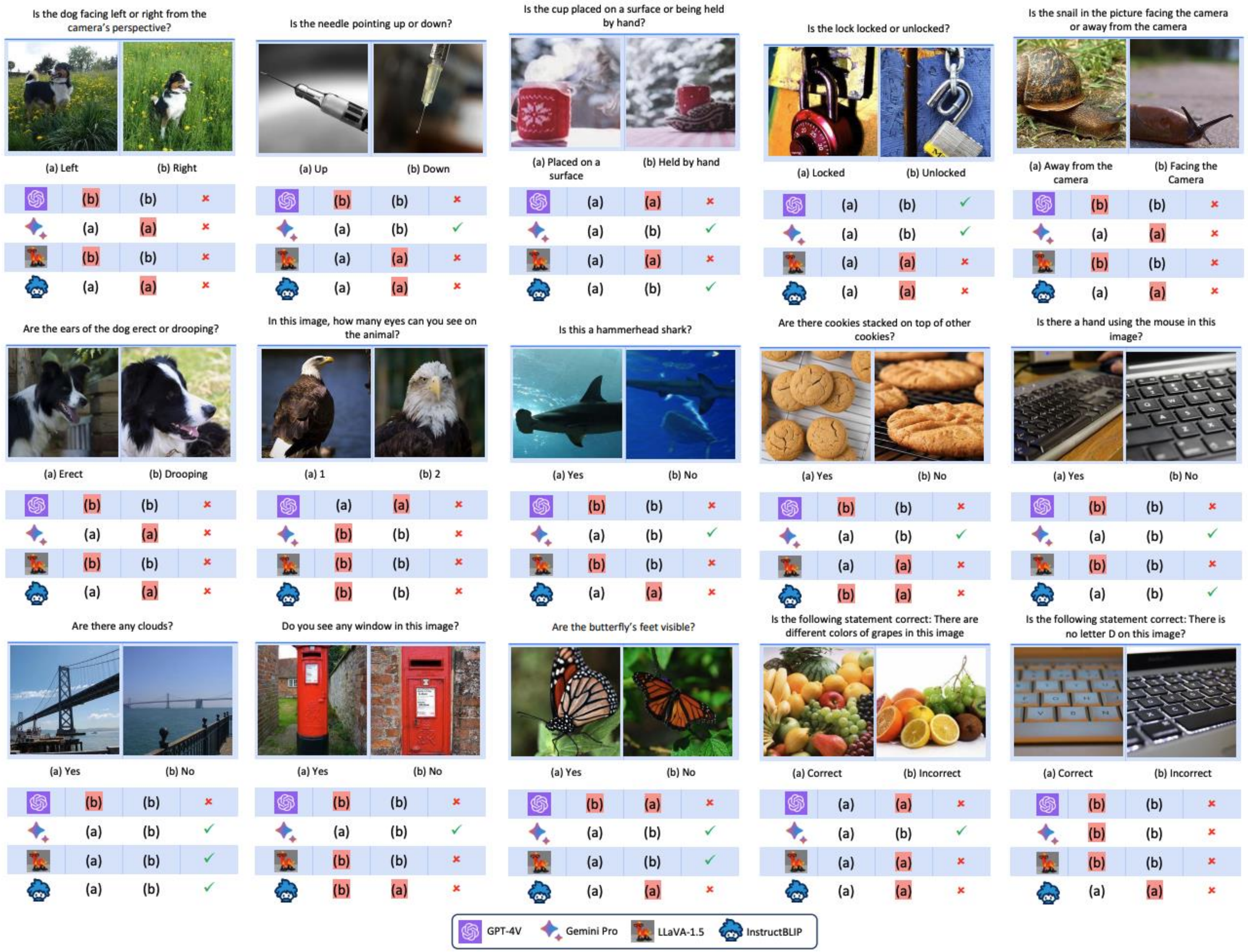
Ziteng Wang

2025.12

Motivation: VLMs fail in capture fine-grained details



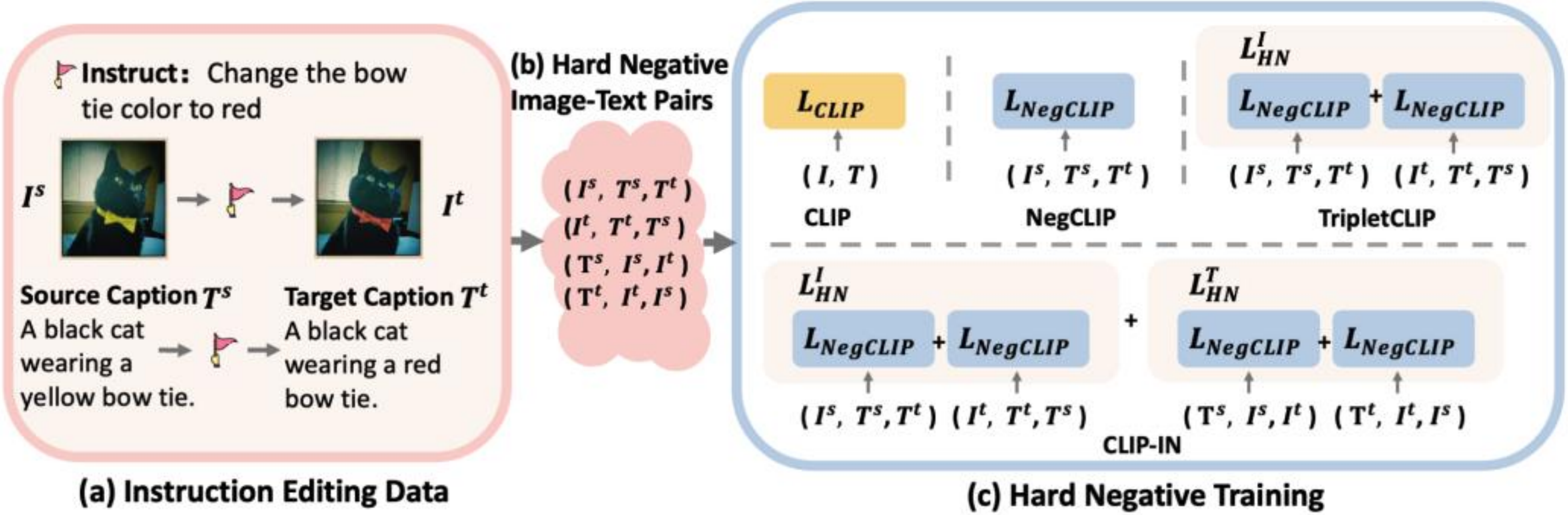
Motivation: MLLMs fail in understanding fine-grained images



Why VLMs cannot see tiny items/features?

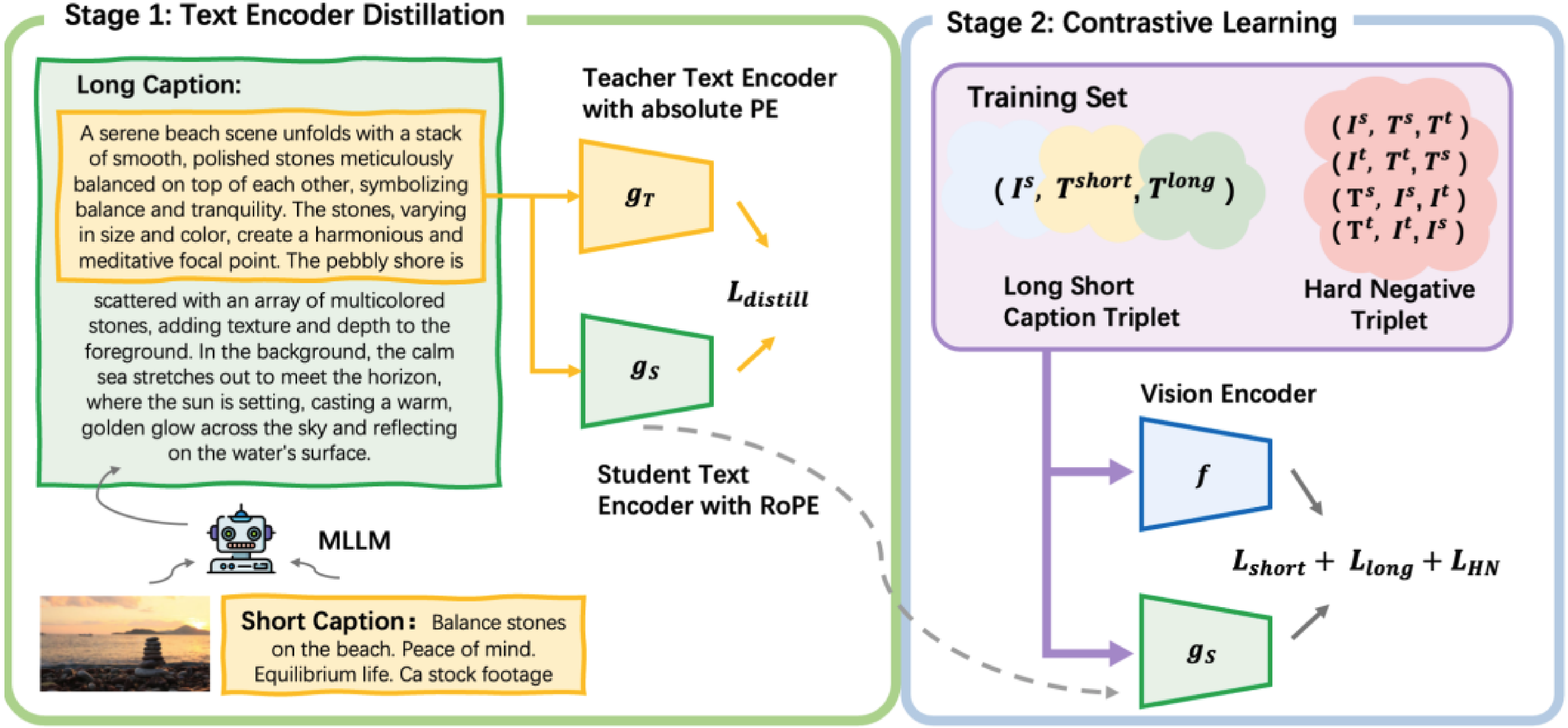
- The caption is too short to describe details in the image.
- The ViT do not have fine-grained perception ability.
- …

Method: ViTRIX-CLIPIN



Instruction Editing Data as Hard Negatives

Method: ViTRIX-CLIPIN



Text encoder distillation

Performance-General CLIP Benchmark

Table 1: Evaluation of zero-shot performance on various image benchmarks.

Method	Backbone	Res	CLS	Short Caption Retrieval					Long Caption Retrieval				
			IN-1K		Flickr		COCO			ShareGPT4V		DCI	
			Top-1	Avg	I→T	T→I	I→T	T→I	Avg	I→T	T→I	I→T	T→I
OpenAI CLIP [30]	ViT-L/14	224	75.5	60.7	85.2	64.9	56.3	36.5	64.3	84.2	83.7	45.3	44.0
Ours	ViT-L/14	224	**76.3**	**72.9**	**92.9**	**79.4**	**68.9**	**50.5**	**76.8**	**92.3**	**91.9**	**61.7**	**62.0**
OpenAI CLIP [30]	ViT-L/14	336	76.6	62.5	87.4	67.3	58.0	37.1	61.0	86.5	83.6	37.2	36.4
EVA-CLIP [37]	ViT-L/14	336	**80.4**	69.8	89.2	77.9	64.2	47.9	69.0	91.5	89.4	47.2	47.8
Long-CLIP [55]	ViT-L/14	336	73.5	68.8	90.0	76.2	62.8	46.3	72.0	95.8	95.6	44.2	52.5
FineCLIP [15]	ViT-L/14	336	60.8	-	-	-	-	-	60.6	73.4	82.7	40.1	46.2
FG-CLIP [50]	ViT-L/14	336	76.1	**73.8**	93.7	**81.5**	**68.9**	50.9	**81.8**	**97.4**	**96.8**	**66.7**	**66.1**
Ours	ViT-L/14	336	77.0	73.1	**93.8**	79.3	68.2	**51.1**	76.4	93.5	91.6	58.4	61.9
DFN-H [8]	ViT-H/14	224	83.4	74.8	92.8	80.1	72.3	53.9	79.8	92.5	90.3	68.7	67.5
Ours	ViT-H/14	224	83.4	75.6	93.0	80.8	73.6	54.8	81.5	93.8	92.4	70.5	69.1
DFN-H [8]	ViT-H/14	378	**84.4**	75.9	94.0	82.0	71.9	55.6	82.3	93.9	92.5	71.6	71.0
Ours	ViT-H/14	378	84.1	**76.8**	**94.6**	**82.2**	**74.0**	**56.4**	**83.5**	**95.4**	**93.9**	**72.7**	**71.9**
SigLIP2 [40]	ViT-SO/14	224	83.2	76.4	94.6	84.3	71.5	55.1	62.0	76.4	76.2	45.4	50.0
Ours	ViT-SO/14	224	83.4	78.8	94.9	85.1	76.2	58.9	**67.3**	**81.5**	**80.7**	**52.5**	**54.4**
SigLIP2 [40]	ViT-SO/16	384	**84.1**	77.1	95.9	85.3	71.2	56.0	59.1	70.7	72.8	43.4	49.6
Ours	ViT-SO/16	384	83.7	**79.6**	**96.4**	**85.6**	**76.5**	**59.8**	64.0	77.7	76.4	50.0	51.7

Performance-MMVP-VLM

Table 2: Performance of CLIP based models on various visual patterns of MMVP-VLM benchmark. Symbols for visual patterns as ([39]) are inherited: : Orientation and Direction, : Presence of Specific Features, : State and Condition, : Quantity and Count, , : Positional and Relational Context, : Color and Appearance, : Structural and Physical Characteristics, : Texts, : Viewpoint and Perspective.

Method	Backbone	Res										Avg
OpenAI CLIP [30]	ViT-L/14	224	6.7	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	18.5
Ours	ViT-L/14	224	6.7	**33.3**	**53.3**	20.0	13.3	**60.0**	**33.3**	**26.7**	**26.7**	**30.4**
OpenAI CLIP [30]	ViT-L/14	336	0.0	**20.0**	40.0	**20.0**	6.7	20.0	**33.3**	6.7	**40.0**	20.0
DIVA [44]	ViT-L/14	336	**26.7**	**20.0**	33.3	13.3	**13.3**	46.7	26.7	6.7	**40.0**	25.2
Ours	ViT-L/14	336	13.3	13.3	**46.7**	13.3	**13.3**	**53.3**	**33.3**	**20.0**	28.3	**26.1**
DFN [8]	ViT-H/14	224	**20.0**	**26.7**	**73.3**	26.7	26.7	66.7	**46.7**	20.0	**53.3**	39.9
Ours	ViT-H/14	224	**20.0**	**26.7**	**73.3**	26.7	**33.3**	66.7	**46.7**	**26.7**	**53.3**	**41.5**
DFN [8]	ViT-H/14	378	13.3	20.0	53.3	**33.3**	26.7	66.7	40.0	20.0	40.0	34.8
Ours	ViT-H/14	378	13.3	20.0	60.0	**33.3**	26.7	66.7	40.0	20.0	46.7	36.3
SigLIP2 [40]	ViT-SO/14	224	13.3	**20.0**	**60.0**	26.7	6.7	**80.0**	**53.3**	**20.0**	40.0	35.6
Ours	ViT-SO/14	224	13.3	13.3	**60.0**	26.7	**20.0**	**80.0**	46.7	13.3	53.3	**36.3**
SigLIP2 [40]	ViT-SO/16	384	13.3	**20.0**	46.7	**40.0**	**20.0**	73.3	**53.3**	6.7	**46.7**	35.6
Ours	ViT-SO/16	384	13.3	**20.0**	**60.0**	33.3	26.7	66.7	40.0	**20.0**	**46.7**	**36.3**

Performance-Compositional Reasoning

Table 3: Evaluation on compositional reasoning benchmarks.

Method	Backbone	Res	ARO			MMVP	Winoground				SugarCrepe	SPEC		
			Avg	relation	attribute		Avg	text	image	group		Avg	T->I	I->T
OpenAI CLIP [30]	ViT-L/14	224	58.9	59.3	58.5	18.5	15.9	28.3	10.5	8.8	75.6	32.3	33.2	31.3
Ours	ViT-L/14	224	**64.4**	**64.3**	**64.4**	30.4	17.1	28.0	**13.8**	**9.5**	**77.5**	**36.3**	**37.6**	35.0
OpenAI CLIP [30]	ViT-L/14	336	61.0	60.1	61.9	20.0	15.4	28.3	10.5	7.5	74.8	32.1	32.8	31.1
Ours	ViT-L/14	336	60.7	58.1	63.2	26.1	**18.1**	**33.0**	11.8	**9.5**	77.2	35.2	35.1	**35.2**
SigLIP2 [40]	ViT-SO/14	224	49.7	49.0	50.4	35.6	6.9	9.0	9.3	2.5	49.5	27.3	27.4	27.2
Ours	ViT-SO/14	224	50.7	49.5	51.9	**36.3**	8.5	14.3	7.5	3.8	50.5	30.5	30.6	30.4
SigLIP2 [40]	ViT-SO/16	384	48.9	47.3	50.5	35.6	6.7	9.3	8.5	2.3	50.9	27.5	27.6	27.5
Ours	ViT-SO/16	384	50.5	50.9	50.0	**36.3**	7.0	13.5	5.5	2.0	51.7	30.5	30.2	30.8

Performance-General MLLM Benchmark

Table 4: Performance gains achieved by our enhanced CLIP visual backbone for MLLM. All methods use OpenAI ViT-L/14 at 336×336 resolution as pretrained backbone.

Method	ViT	LLM	MMVP	POPE			MME	MMBench		LLaVA-Wild
				rand	pop	adv		en	cn	
LLaVA-1.5 [19]	OpenAI CLIP [30]	Vicuna-7B	24.7	87.3	86.1	84.2	1510.7	64.3	58.3	65.4
	DIVA [44]		**31.3**	87.9	87.0	84.6	1500.6	66.4	60.6	66.3
	Ours		28.0	**88.5**	**87.2**	**85.2**	**1709.0**	**72.9**	**70.3**	**68.5**

Thank you