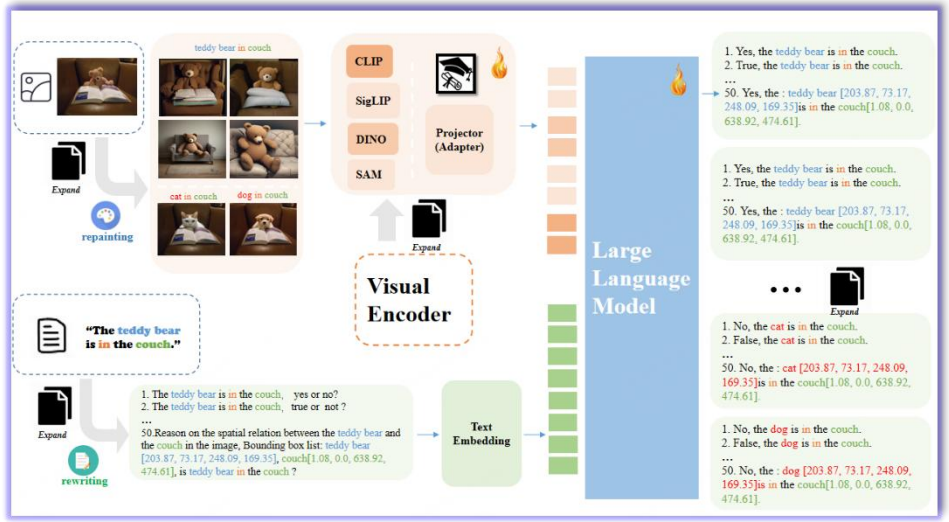
Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

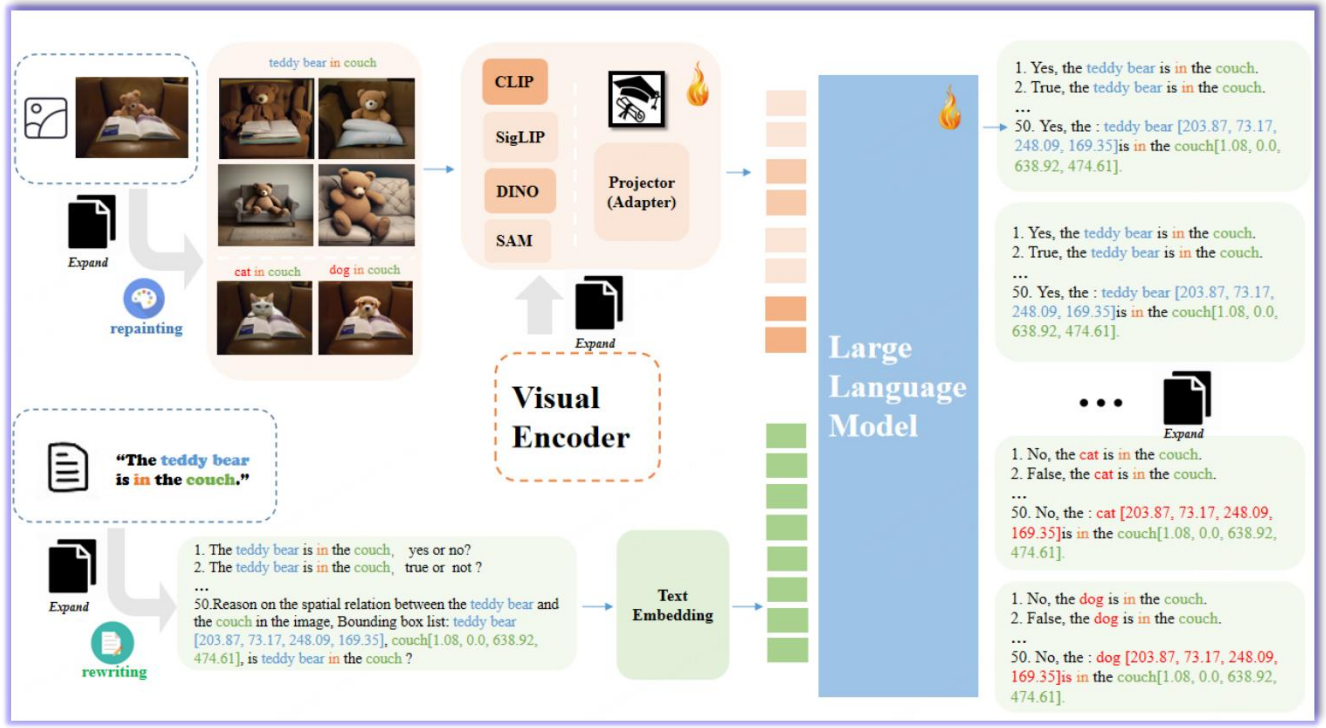
Peijin Xie 2025/1/10

Harbin Institute of Technology



Motivation
Expansion
Experiment
Result

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules



Motivation

Expansion

Experiment

Result

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

Re-evaluation on VSR

1. Over-sensitivity to language instructions
2. Under-sensitivity to visual positional information
3. Answer bias

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

- Re-evaluation on VSR
  - Variance and In-consistency on VSR

Models	Version	LLM	Vision Encoder	ACC(%)
Prismer (Liu et al., 2024)		RoBERTa	CLIP	68.4
	+ Normal			68.3
	+ Edge			67.8
	+ Seg.			68.4
	+ OCR Det.			67.2
	+ Obj. Det.			68.3
	No Experts			65.6
	+ 6 Experts			68.7
Prism* (Karamcheti et al., 2024)	Prism-CLIP 7B (C)	LLaMA 2 7B	CLIP	66.6
	Prism-CLIP 7B			57.7
	Prism-SigLIP 7B (C)		SigLIP	65.1
	Prism-SigLIP 7B			56.7
	Prism-DINOSigLIP 7B (C)		DINO+SigLIP	66.2
	Prism-DINOSigLIP 7B			59.5
	Prism-CLIP 13B (C)	LLaMA 2 13 B	CLIP	65.9
	Prism-CLIP 13B			71.8
	Prism-SigLIP 13B (C)		SigLIP	62.8
	Prism-SigLIP 13B			64.5
	Prism-DINOSigLIP 13B (C)		DINO+SigLIP	71.8
	Prism-DINOSigLIP 13B			72.1
BLIVA (Hu et al., 2023)		vicuna13B	CLIP + QF	62.2
		FlanT5XXL		68.8
VisLingInstruct (Zhu et al., 2024)		FlanT5XL	EVA	64.1
		FlanT5XXL		66.9
		vicuna-7B		60.1
		vicuna-13B		56.2
Text-only LMs (Azkune et al., 2024)		BERT-base (110M)		73.6
		BERT-large (336M)		74.4
		T5-base (220M)		73.1
		T5-large (770M)		74.4
		T5-3B (3B)		74.5
VisualBERT (Li et al., 2019)				51.0
LXMERT (Tan and Bansal, 2019)				61.2
VILT (Kim et al., 2021)				63.0

Table 2: Part 2 Summary of popular Models performance on VSR dataset including VLLMs, Text-only LMs and traditional pre trained VLMs. Prism (Karamcheti et al., 2024) investigated the design space of visually-conditioned language models and provide more 55 results in its last page of the appendix.

VLMM	Version	LLM	Vision Encoder	ACC(%)
GPT-4o	our template			84.6
Bard (Shao et al., 2023a)	*	*	*	82.0
GPT-4V (Lei et al., 2024)	naive	-	-	67.8
	CoT	-	-	70.4
	Scaffolding	-	-	74.4
	Coordinates			
mPLUG-OWL (Ye et al., 2024)		LLaMA-7B	CLIP	46.0
OpenFlamingo-V2 (Awadalla et al., 2023)	*	*	*	58.0
Otter	*	*	*	24.0
Otter-I (Li et al., 2023a)	*	*	*	56.0
VPGTrans (Zhang et al., 2023a)	*	*	*	40.0
PandaGPT (Su et al., 2023)		vicuna-13B	Imagebind	46.7
LLaMA-Adapter (Zhang et al., 2023b)	LA	LLaMA 7B	CLIP (Multi-scale)	50.6
	LA v2	LLaMA 7B	CLIP (Multi-scale) + caption expert	52.0
Cobra (Zhao et al., 2024)		Mamba-2.8B	DINOv2+SigLIP	63.6
Mini-GPT4 (Zhu et al., 2023)	13B	LLaMA2-chat (7B)	EVA	41.6
	7B			60.6
	7B chat			62.9
LLaVA (Liu et al., 2023)	v1.5	vicuna-7B	CLIP	51.4
		vicuna-13B	CLIP	51.2
InstructBLIP (Dai et al., 2023)		FlanT5XL	CLIP + QF	64.8
		FlanT5XXL		65.6
		vicuna-7B		54.3
		vicuna-13B		52.1
BLIP2 (Li et al., 2023b)		FlanT5XL	CLIP + QF	60.5
		FlanT5XXL		68.2
		vicuna-7B		50.0
		vicuna-13B		50.9
ImageBind-LLM (Han et al., 2023)	(D)	LLaMA	imagebind	49.3
				49.7
Qwen-VL (Bai et al., 2023)	Qwen7B	Qwen7B	ViT-bigG	63.8
	Qwen7B-chat			67.5
SPHINX (Lin et al., 2023)	1k	LLaMA 2	mixed*	62.6
	2k			62.9
				63.1

Table 1: Part 1 Summary of popular VLMMs' performance on VSR dataset. In the first row, we evaluated the GPT-4o with it's API using our own prompt mentioned later in our *Test-S* set. Label "*" are result gathered from LVLM-eHub(Shao et al., 2023b) and "-" are from *SCAFFOLD*(Lei et al., 2024) release.

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

- Re-evaluation on VSR
  - Variance and In-consistency on VSR
  - Hyper-senstivity on Language Prompt

Index	Templates of phase 1(the simplest questioning style)	answered	acc_t	acc_a
1	[caption], Yes or no.	1.00	0.61	0.61
2	[caption], True or not.	0.96	0.50	0.53
3	[caption], True or false.	1.00	0.72	0.72
4	Whether the [subject] is [relation] [object]? Yes or no.	1.00	0.64	0.64
5	Whether the [subject] is [relation] [object]? True or not.	0.80	0.45	0.56
6	Whether the [subject] is [relation] [object]? True or false.	1.00	0.67	0.67
7	Is the [subject] [relation] [object]? yes or no.	1.00	0.65	0.65
8	Is the [subject] [relation] [object]? True or not.	0.91	0.61	0.67
9	Is the [subject] [relation] [object]? True or false.	0.98	0.65	0.66
10	Whether [caption]? Answer the question with yes or no.	0.77	0.47	0.61
11	Is the [subject] [relation] [object]? Answer the question with yes or no.	0.80	0.48	0.59
12	Whether [caption]? A. yes B. no Answer with the option's letter from the given choices directly.	1.00	0.62	0.62
13	Is the [subject] [relation] [object]? A. yes B. no Answer with the option's letter from the given choices directly.	1.00	0.68	0.68
14	Answer the following binary question with the capital letter of the answer list below. Whether [caption]? A.yes B.no.	1.00	0.65	0.65
15	Answer the following binary question with the capital letter of the answer list below. Is the [subject] [relation] [object]? A.yes B.no.	1.00	0.68	0.68

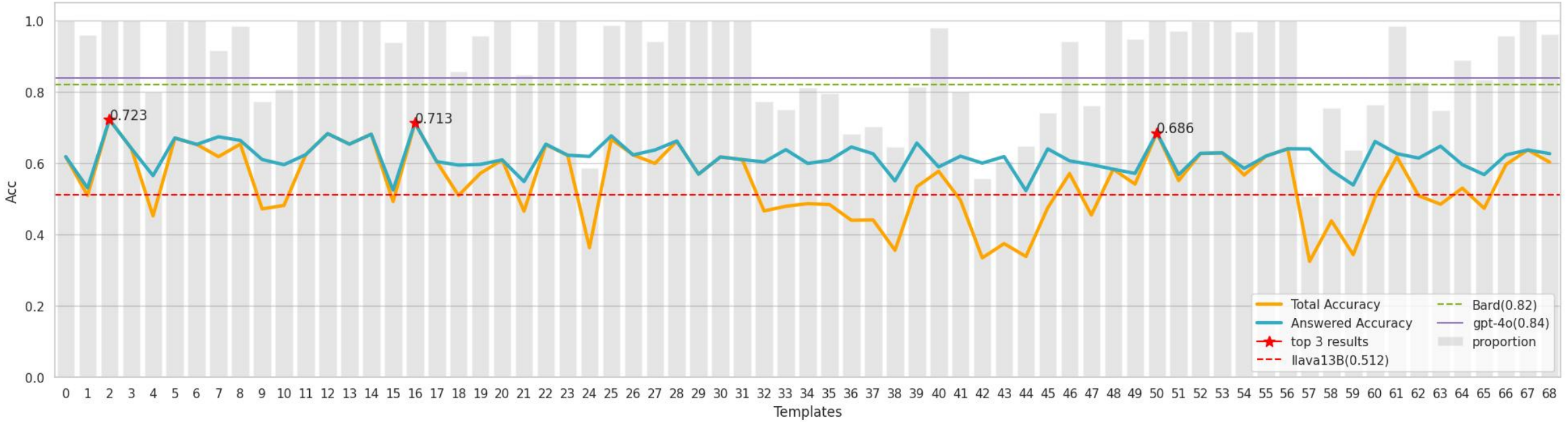
Table 3: Templates of phase 1 which use the simplest questioning style avoiding any unnecessary words and covering most binary question formats. This includes two types of questioning methods: general declarative statements and general interrogative sentences. It covers questions formulated with "is" and "whether", as well as multiple-choice question formats. The replaceable parts in the templates are highlighted in colored fonts.

Index	Templates of phase 4 (bounding box added)	answered	acc_t	acc_a
49	The [subject] [subject box] is [relation] [object] [object box], yes or no.	1.00	0.58	0.58
50	The [subject] [subject box] is [relation] [object] [object box], True or not.	0.94	0.54	0.57
51	The [subject] [subject box] is [relation] [object] [object box], True or false.	1.00	0.68	0.68
52	Whether the [subject] [subject box] is [relation] [object] [object box]? Yes or no.	0.97	0.55	0.56
53	Whether the [subject] [subject box] is [relation] [object] [object box]? True or not.	1.00	0.62	0.62
54	Whether the [subject] [subject box] is [relation] [object] [object box]? True or false.	1.00	0.62	0.62
55	Is the [subject] [subject box] [relation] [object] [object box]? yes or no.	0.96	0.56	0.58
56	Is the [subject] [subject box] [relation] [object] [object box]? True or not.	1.00	0.62	0.62
57	Is the [subject] [subject box] [relation] [object] [object box]? True or false.	1.00	0.64	0.64
58	Distinguish the positional relation between the [subject] [subject box] and the [object] [object box] in the image, whether [caption]?	0.50	0.32	0.64
59	Focus on the positional relation between the [subject] [subject box] and the [object] [object box] in the image, whether [caption]?	0.75	0.43	0.58
60	Reason on the spatial relation between the [subject] [subject box] and the [object] [object box] in the image, whether [caption]?	0.63	0.34	0.53
61	Distinguish the positional relation between the [subject] [subject box] and the [object] [object box] in the image, is the [subject] [relation] the [object]?	0.76	0.50	0.66
62	Focus on the positional relation between the [subject] [subject box] and the [object] [object box] in the image, is the [subject] [relation] the [object]?	0.98	0.61	0.62
63	Reason on the spatial relation between the [subject] [subject box] and the [object] [object box] in the image, is the [subject] [relation] the [object]?	0.82	0.50	0.61
64	Distinguish the positional relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], whether [caption]?	0.74	0.48	0.64
65	Focus on the positional relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], whether [caption]?	0.88	0.53	0.59
66	Reason on the spatial relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], whether [caption]?	0.83	0.47	0.56
67	Distinguish the positional relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], is the [subject] [relation] the [object]?	0.95	0.59	0.62
68	Focus on the positional relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], is the [subject] [relation] the [object]?	1.00	0.63	0.63
69	Reason on the spatial relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], is the [subject] [relation] the [object]?	0.96	0.60	0.62

Table 6: Templates of phase 4 with bounding box of related entities added. We further provided the ground truth [subject box] and [object box] to assist the model in its reasoning process.

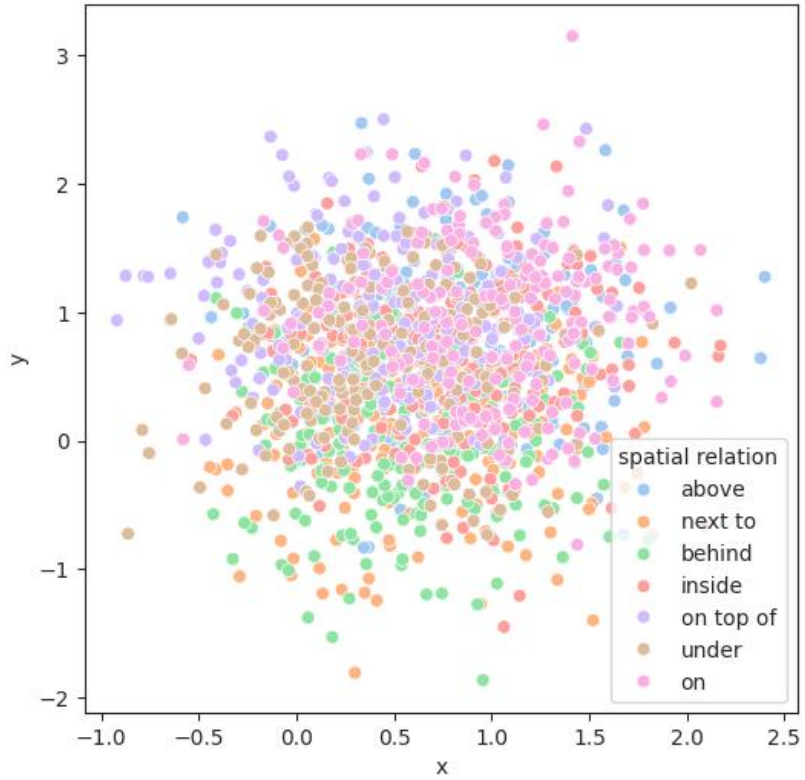
Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

- Re-evaluation on VSR
  - Variance and In-consistency on VSR
  - Hyper-senstivity on Language Prompt



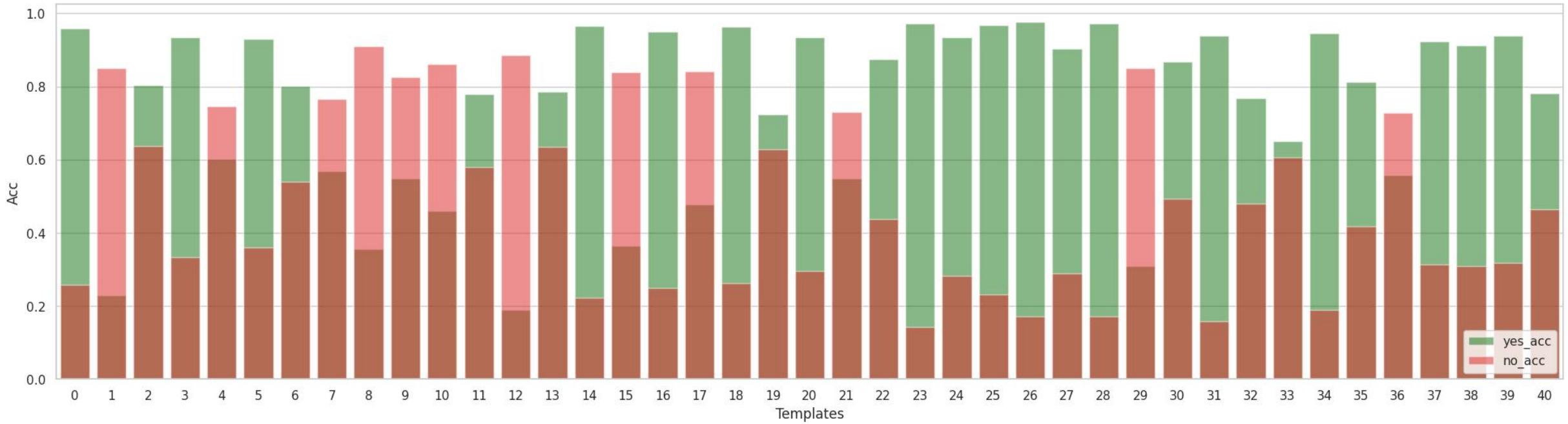
Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

- Re-evaluation on VSR
  - Variance and In-consistency on VSR
  - Hyper-senstivity on Language Prompt
  - Under-Senstivity on Vision Details



Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

- Re-evaluation on VSR
  - Variance and In-consistency on VSR
  - Hyper-senstivity on Language Prompt
  - Under-Senstivity on Vision Details
  - Bias on Binary QA Senario



Motivation

Expansion

Experiment

Result

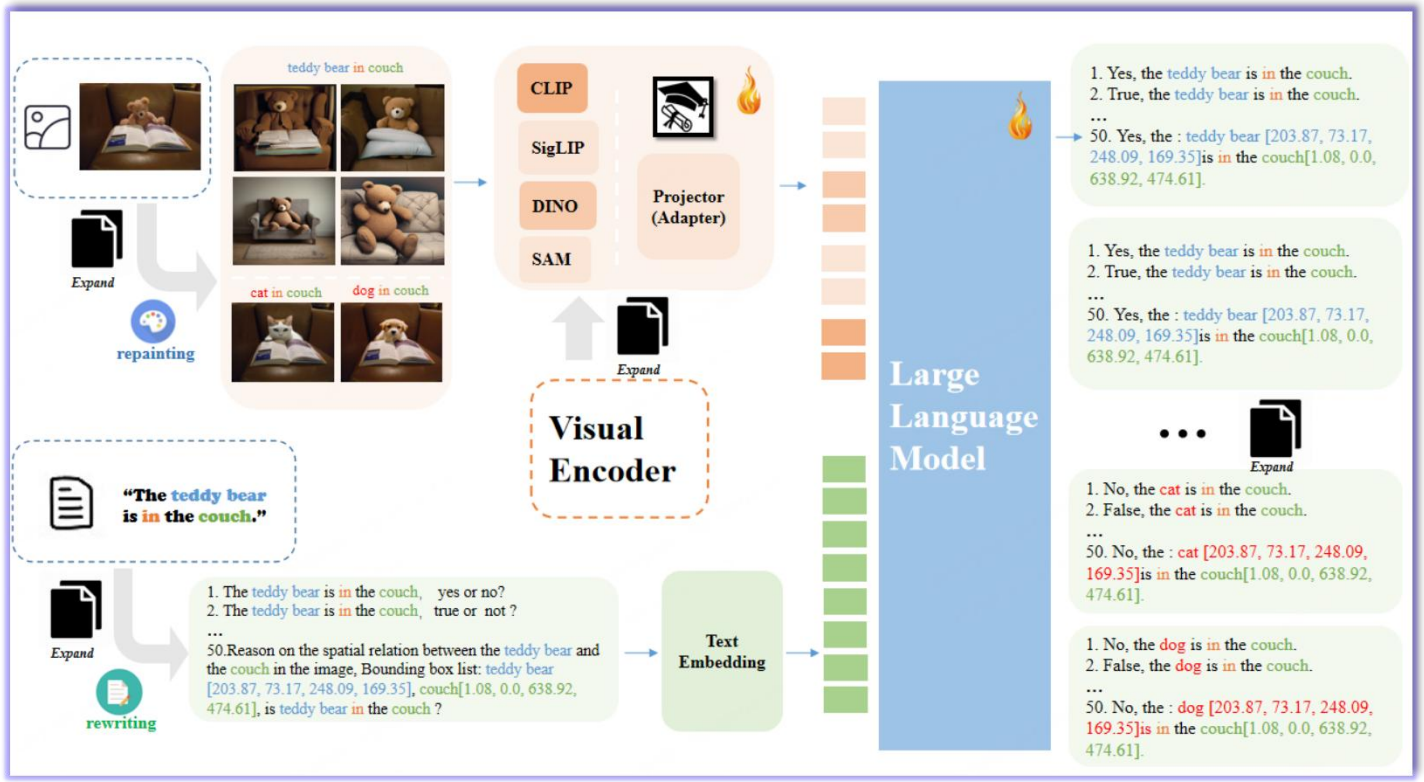
Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

Expansion for Spatial Expert

1. Expansion on Text Data
2. Expansion on Image Data
3. Expansion on Vision Encoder

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. **Expansion on Text Data**
2. **Expansion on Image Data**
3. **Expansion on Vision Encoder**



Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. Expansion on Text Data
2. Expansion on Image Data
3. Expansion on Vision Encoder

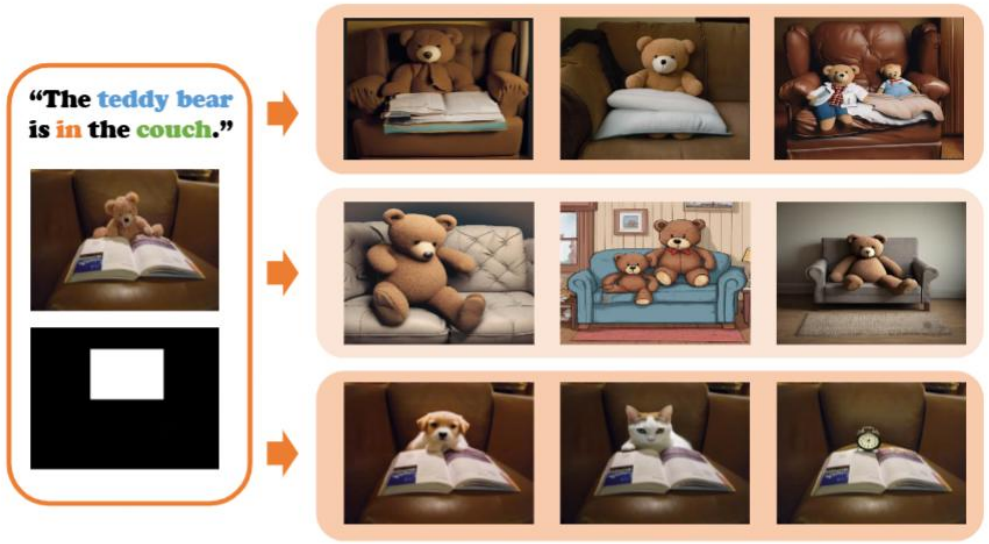
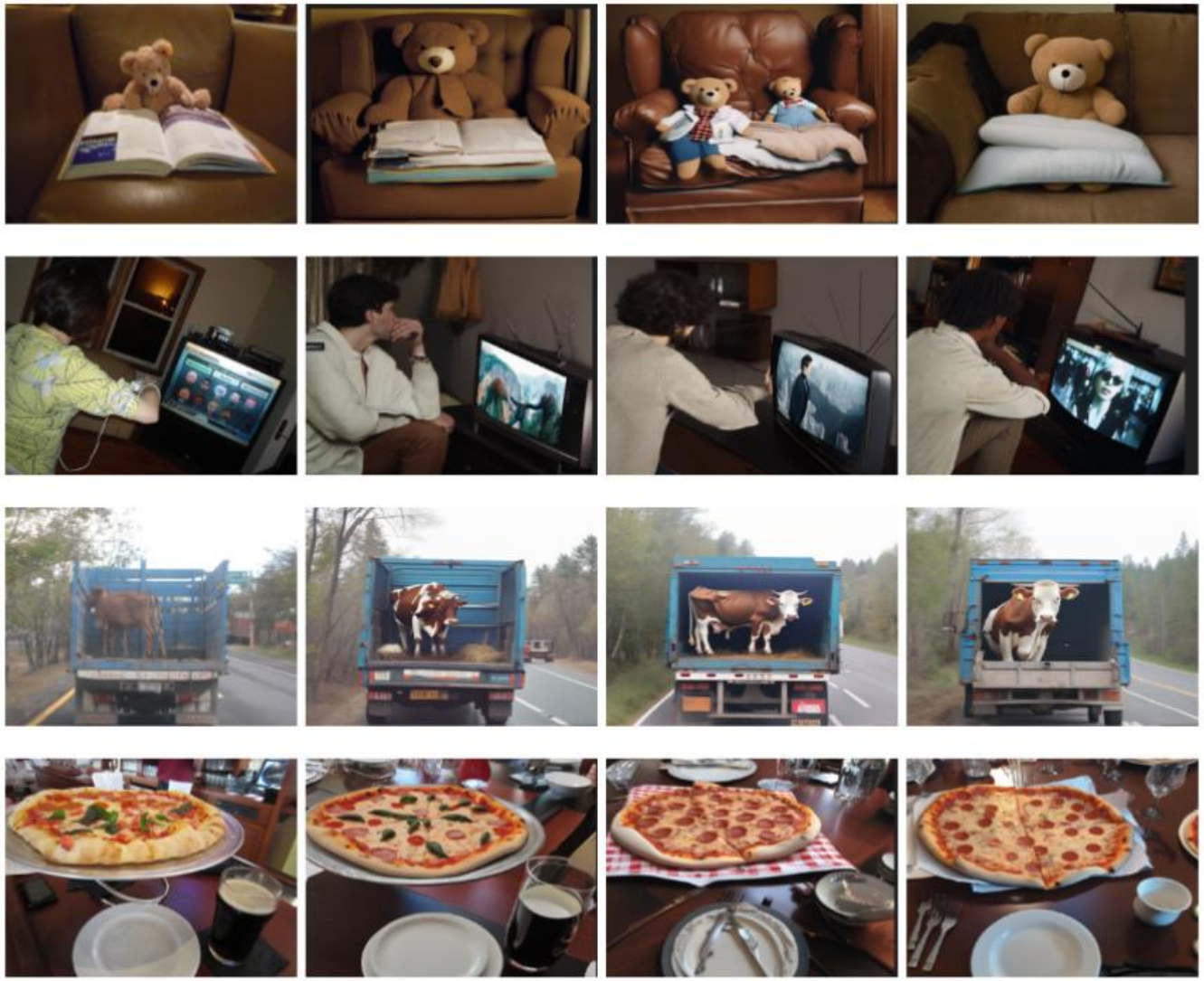


Figure 2: Examples of 3 settings of image-to-image(first row), text-to-image(middle row), and inpainting(last row) through the repainting process with the original image-text pair and mask inputs on the left.

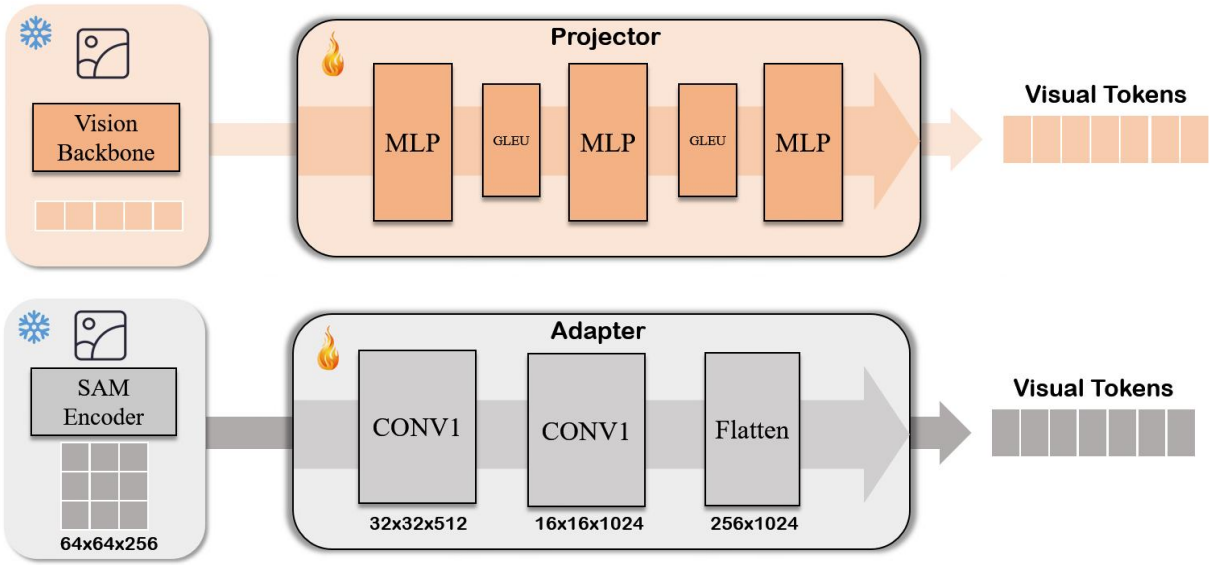


Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. **Expansion on Text Data**
2. **Expansion on Image Data**
3. **Expansion on Vision Encoder**

Vision Backbone	Version	Selected Feature Size	Alignment Module
CLIP	openai/clip-vit-large-patch14-336	[1, 577, 1024]	projector
SigLIP	google/siglip-so400m-patch14-384	[1, 729, 1152]	projector
DINOv2	facebook/dinov2-base	[1, 257, 768]	projector
SAM	facebook/sam-vit-base	[1,64,64,256]	adaptor

Table 7: The specific version details of the used visual backbone, as well as the size of the features from the last hidden layer. Notably, LLaVA 1.5 uses the penultimate(-2) layer features from CLIP.



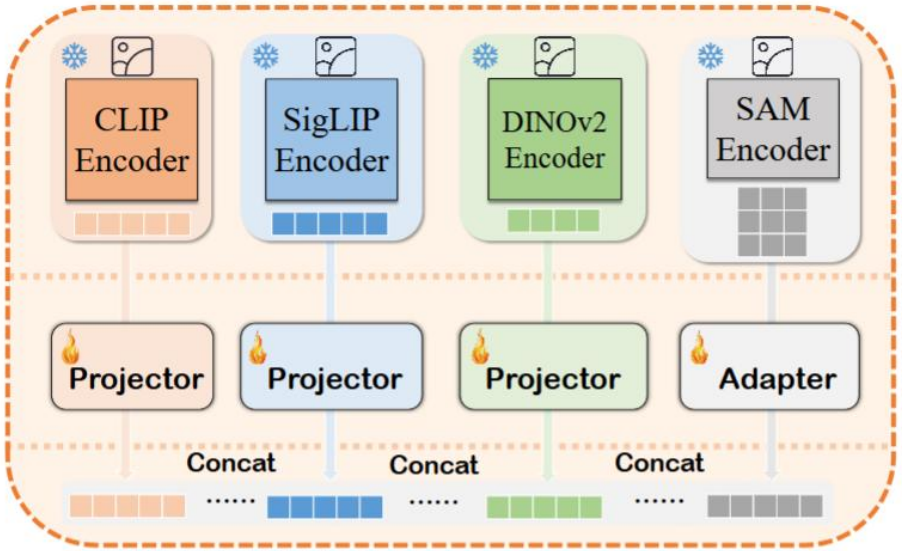


Figure 3: Illustration of the Merged Vision Encoder that concatenate multiple visual features aligned by projector or adapter respectively.

Motivation

Expansion

Experiment

Result

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. Data Detailes
2. Training and Inference

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. Data Detailes

1. Testing Datasets

2. Training Datasets

Data	Usage	Repaint	Rewrite	Amount	Triplet	Images	Templates
Test-G	test	-	-	1222	1222	715	30+20
Test-S	test	-	-	1222	1222	715	1
turn-s 11k	PT / IFT	×	×	3489+7680	3489+7680	5544	1
turn-g 11k	PT / IFT	×	✓	3489+7680	3489+7680	5544	30+20
turn-g 500k	IFT	×	✓	500k	3489+7680	5544	30+20
pre-100k	PT	✓	✓	100k	10k	100k	10
pre-200k	PT	✓	✓	200k	10k	200k	10
pre-300k	PT	✓	✓	300k	10k	300k	10
pre-400k	PT	✓	✓	400k	10k	400k	10
pre-500k	PT	✓	✓	500k	10k	500k	10

Table 8: Statistics of training and testing data. PT refers to pertaining and IFT to instruction fine-tuning. The "Repaint" indicates whether the training data used augmented repainting image data and the "Rewrite" indicates whether used augmented rewriting text data from the template pool. "Template" calculates the number of templates used in the instruction construction.

2. Training and Inference

Motivation

Expansion

Experiment

Result

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. Scaling on Data
2. Scaling on Model
3. Other Benchmarks
4. More Sensitive Vision Features
5. Bias Result

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. **Scaling on Data**
2. **Scaling on Model**
3. **Other Benchmarks**
4. **More Sensitive Vision Features**
5. **Bias Result**

Pretrain Adapter	IFT Adapter+LLM	acc 7B *Test-G/Test-S*	acc 13B *Test-G/Test-S*
-	-	54.3 / 65.3	57.7 / 68.4
-	turn-g 11k	-	-
turn-g 11k	-	57.3 / 65.7	59.2 / 68.2
-	turn-s 11k	-	-
turn-s 11k	-	55.1 / 67.9	56.7 / 70.1
-	turn-g 500k	58.3 / 69.5	62.5 / 71.4
pre-100k	turn-g 500k	61.7 / 71.0	63.2 / 73.7
pre-200k	turn-g 500k	64.1 / 73.3	65.8 / 74.9
pre-300k	turn-g 500k	65.6 / 74.1	69.7 / 75.5
pre-400k	turn-g 500k	66.7 / 73.5	70.3 / 75.7
pre-500k	turn-g 500k	66.2 / 73.6	70.2 / 75.6
pre-400k_1 turn-s 11k_3	turn-g 500k_2	66.4 / 74.7	70.1 / 76.6

Table 1: Result of LLaVA1.5 7B and 13B on scaling training data experiment. We post the Test-G and Test-S accuracy (split through "/") by pretrained the adapter with data of the first column and instruct fine turning(IFT) both the adapter and LLM with the second column data.

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

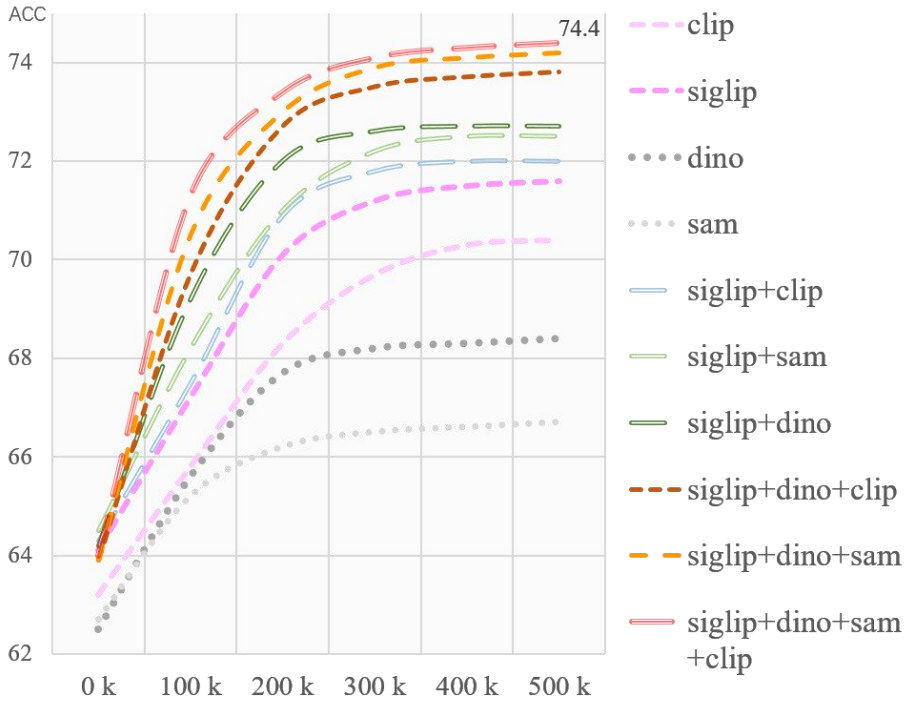
1. **Scaling on Data**
2. **Scaling on Model**
3. **Other Benchmarks**
4. **More Sensitive Vision Features**
5. **Bias Result**

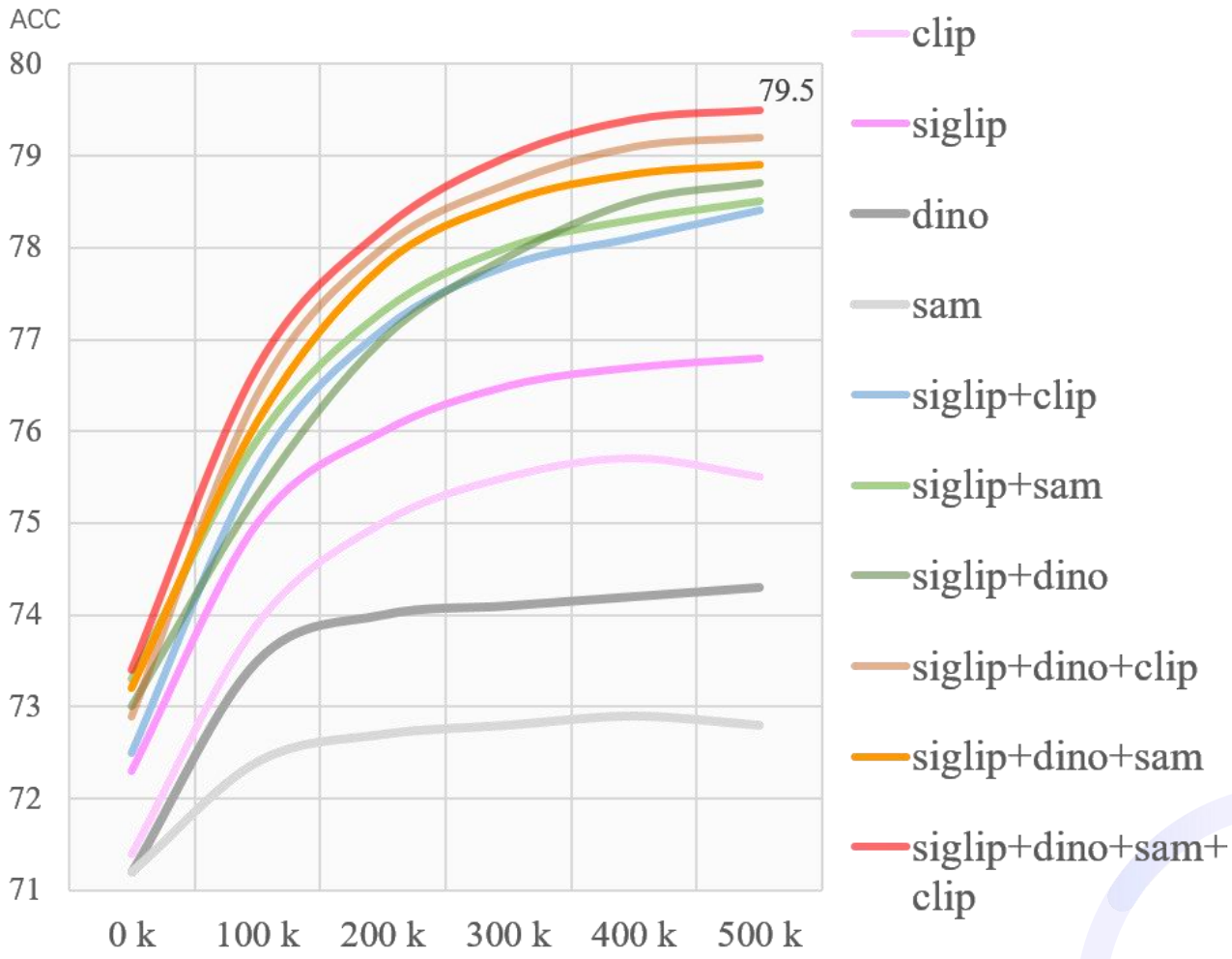
LLM / VLLM	zreo-shot	turn-g 500k	+pre-100k	+pre-200k	+pre-300k	+pre-400k	+pre-500k
vicuna 7B	-	56.4 / 64.0	58.3 / 67.2	60.2 / 68.2	61.6 / 70.1	63.1 / 72.6	63.4 / 72.9
vicuna 13B	-	59.8 / 67.3	62.4 / 69.1	64.7 / 70.8	65.7 / 73.5	68.7 / 74.2	69.2 / 74.2
LLAMA2 7B	-	57.1 / 61.7	57.8 / 64.5	61.3 / 67.3	62.1 / 68.2	63.3 / 69.9	62.4 / 70.1
LLAMA2 13B	-	60.9 / 63.2	61.8 / 67.8	65.5 / 70.6	67.0 / 72.4	69.2 / 73.4	68.8 / 74.4
LLAMA3 8B	-	61.5 / 65.4	62.6 / 68.4	65.8 / 70.5	68.5 / 72.9	70.0 / 74.1	70.0 / 74.3
Qwen-VL 7B	57.8 / 62.7	63.4 / 68.2	65.8 / 69.2	66.1 / 71.7	67.0 / 72.2	67.9 / 73.6	68.2 / 73.6
BILP2 (FlanT5XXL)	59.3 / 66.5	64.2 / 69.6	66.3 / 69.7	68.0 / 71.1	69.3 / 72.5	70.4 / 73.9	70.2 / 74.2
InstructBLIP (FlanT5XXL)	54.4 / 63.1	59.0 / 66.2	62.5 / 67.8	64.3 / 69.7	66.9 / 71.8	67.3 / 71.9	68.2 / 72.3

Table 2: Result of scaling data across other hot-spot LLM and VLLM on Test-G and Test-S (split through "/"). The colum names represent the data used for training, in sequence of first 3 as: no data for "zero-shot", only 500k turn-g data for "turn-g 500k", turn-g 500k plus pre-100k for "+pre-100k". For LLMs, we randomly initialized the adapter weights and sequentially used the corresponding data to perform the pretrain and fine-tune processes. And for VLLMs, we combined the tuning and pre-train data sequentially for IFT.

Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. **Scaling on Data**
2. **Scaling on Model**
3. **Other Benchmarks**
4. **More Sensitive Vision Features**
5. **Bias Result**





Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. Scaling on Data
2. Scaling on Model
3. Other Benchmarks
4. More Sensitive Vision Features
5. Bias Result

Model	MME	MMBench	SEEDv2
MiniGPT-4v2	43.33	-	32.6
Qwen-VL(chat)	128.33	47.2	40.3
LLaVA1.5 13B	133.33	57.6	38.5
BLIP2	73.33	58.4	36.2
VSRE	155.00	64.8	46.6

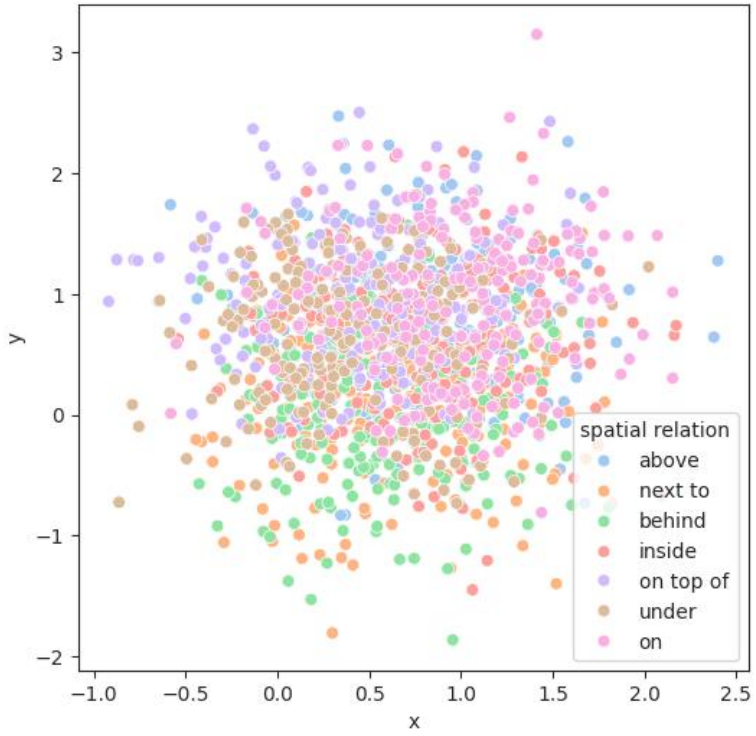
Table 3: The comparison results of VSRE on the related subsets of other datasets including MME, MMBench and SEEDv2.

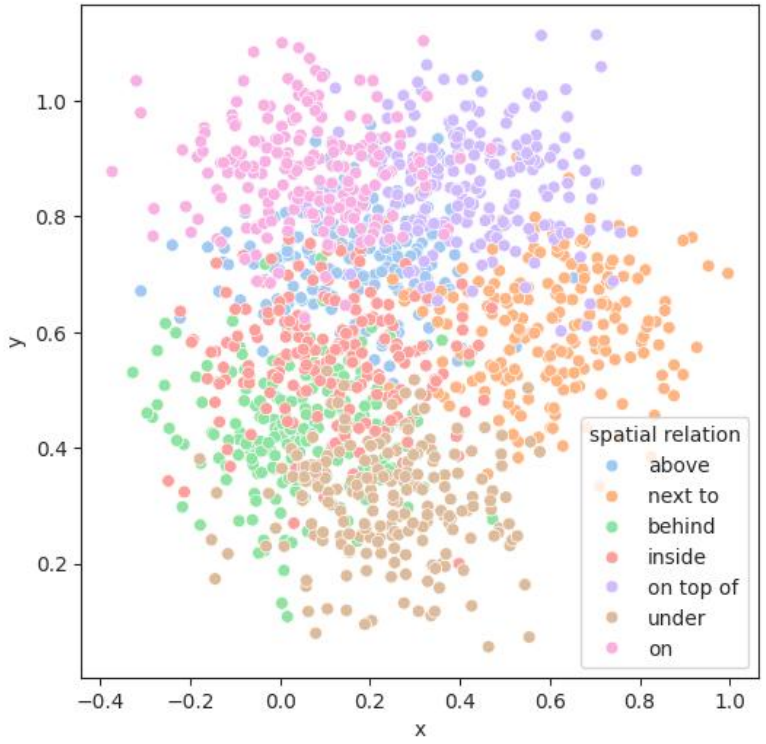
Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. Scaling on Data
2. Scaling on Model
3. Other Benchmarks
4. More Sensitive Vision Features
5. Bias Result

Relations	above	next to	behind	inside
LLaVA(51.2%)	0.54	0.66	0.53	0.42
VSRE(79.5%)	0.24	0.34	0.27	0.21
Relations	on top of	under	on	AVG
LLaVA(51.2%)	0.62	0.57	0.69	0.57
VSRE(79.5%)	0.33	0.29	0.36	0.29

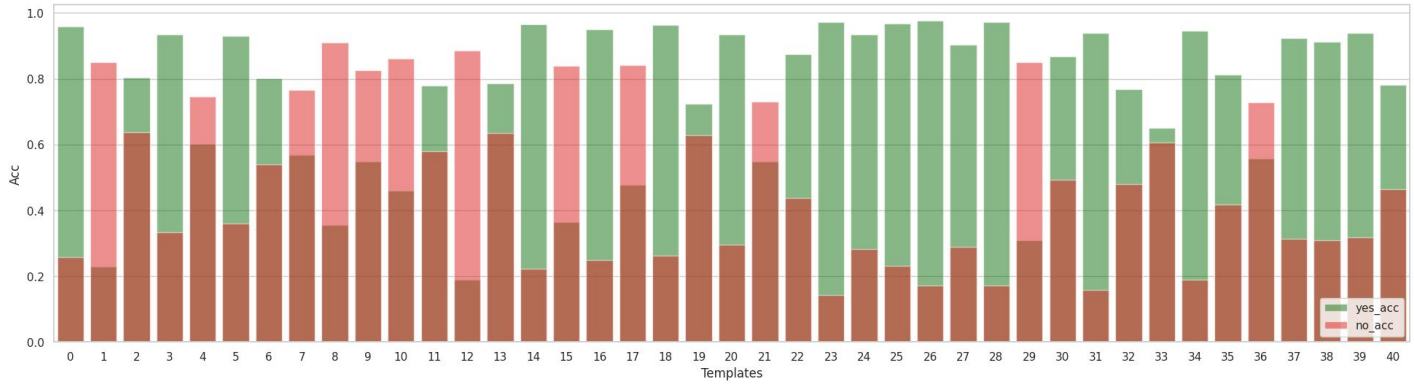
Table 4: Statistic result of average intra-class distance for each spatial relation category on 200 samples by llava1.5 13B (acc 51.2%) and VSRE(acc 79.5%).

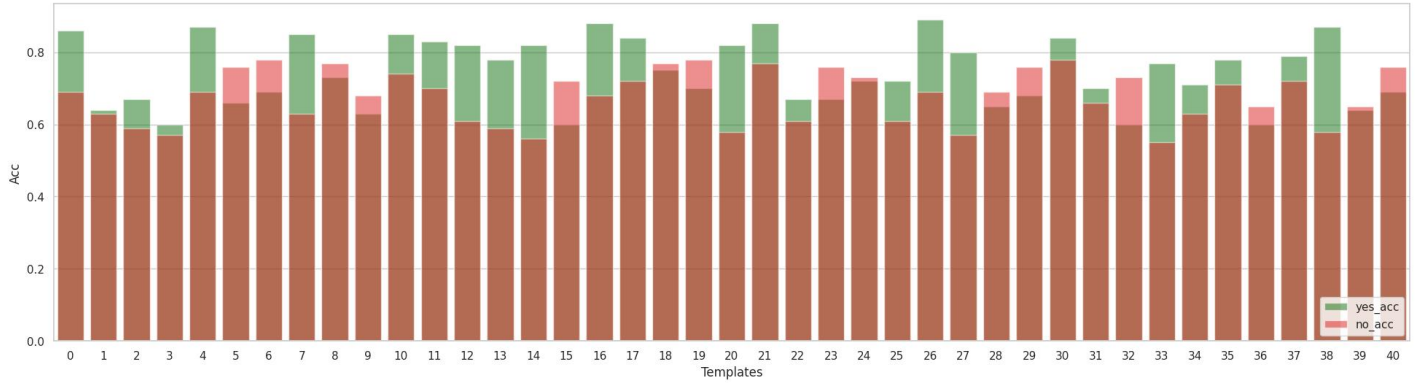




Expand VSR Benchmark for VLLM to Expertize in Spatial Rules

1. Scaling on Data
2. Scaling on Model
3. Other Benchmarks
4. More Sensitive Vision Features
5. Bias Result





Thanks

Peijin Xie