Unified Lexical Representation for Interpretable Visual-Language Alignment

Yifan Li, Yikai Wang, Yanwei Fu, Dongyu Ru, Zheng Zhang, Tong He

Fudan University Amazon Web Services

Background

Visual-Language Alignment (VLA)

- Latent representation
- Lexical representation

Lexical representation

Challenges of learning lexical representation:

An non-negative vector in which each dimension explicitly represents the similarity between an image or text and a specific word.

Background

Visual-Language Alignment (VLA)

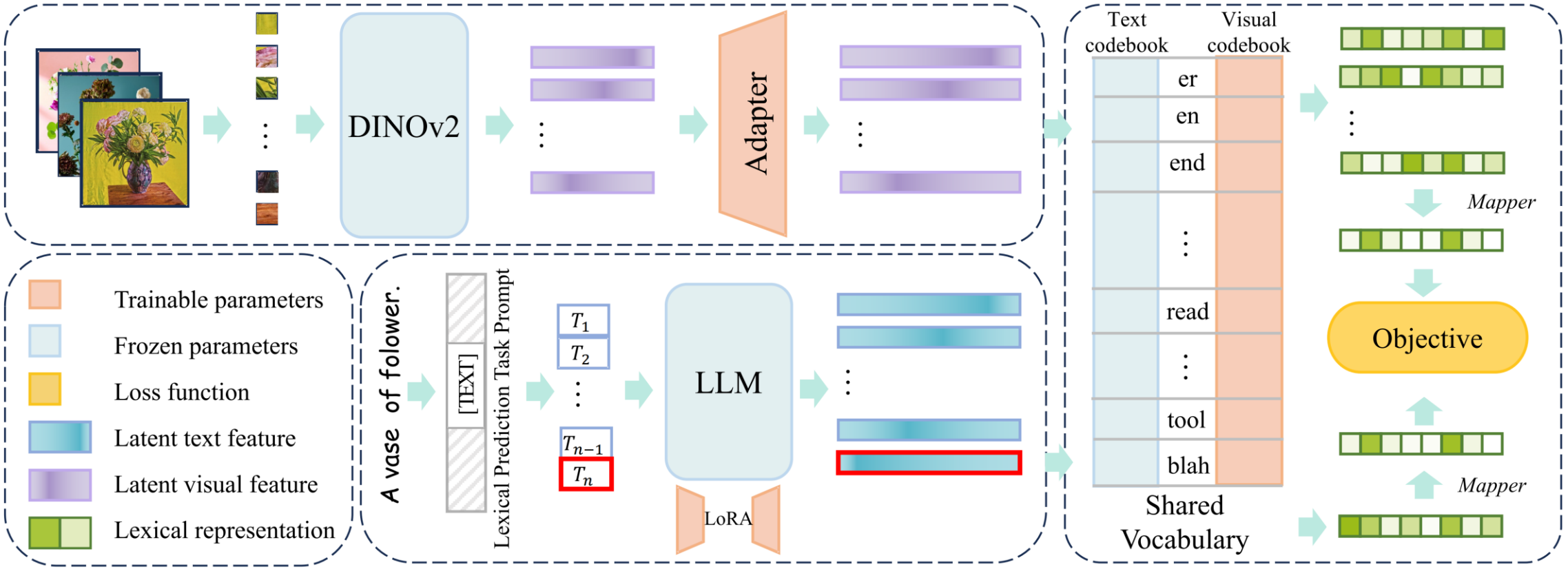
- Latent representation
- Lexical representation

Lexical representation

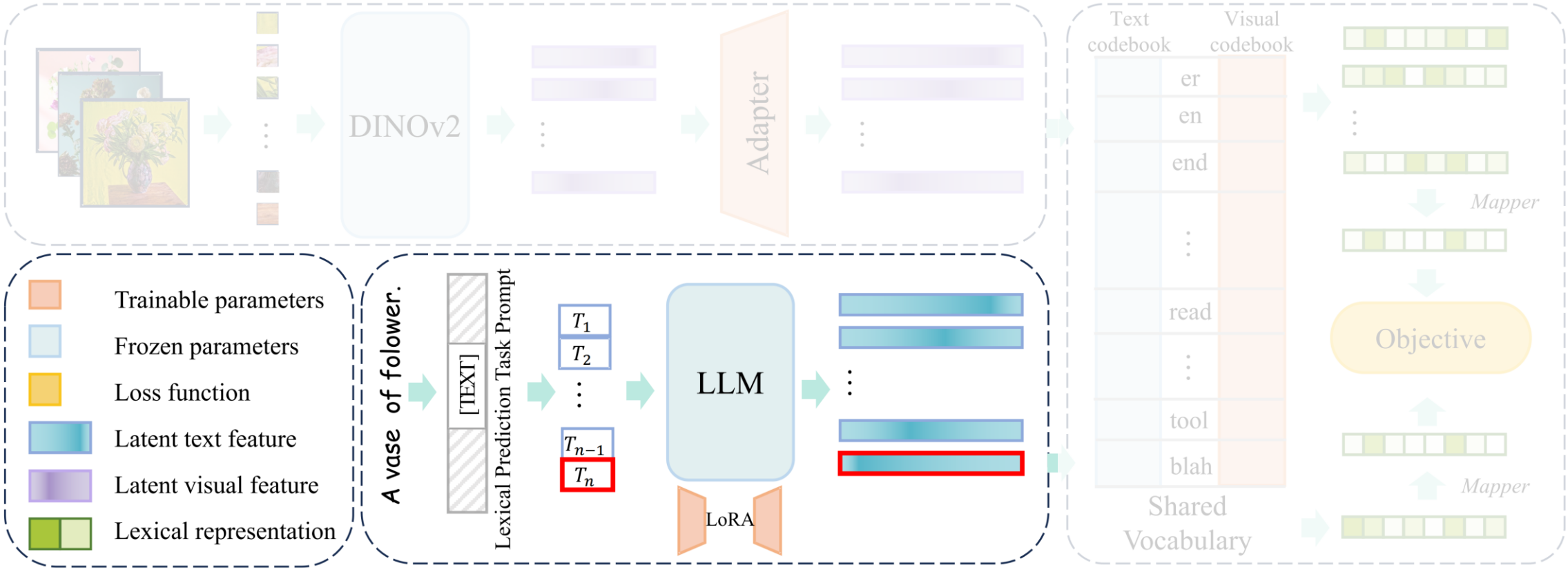
Challenges of learning lexical representation:

1. Lack of precise supervision signals
2. False discovery

Method

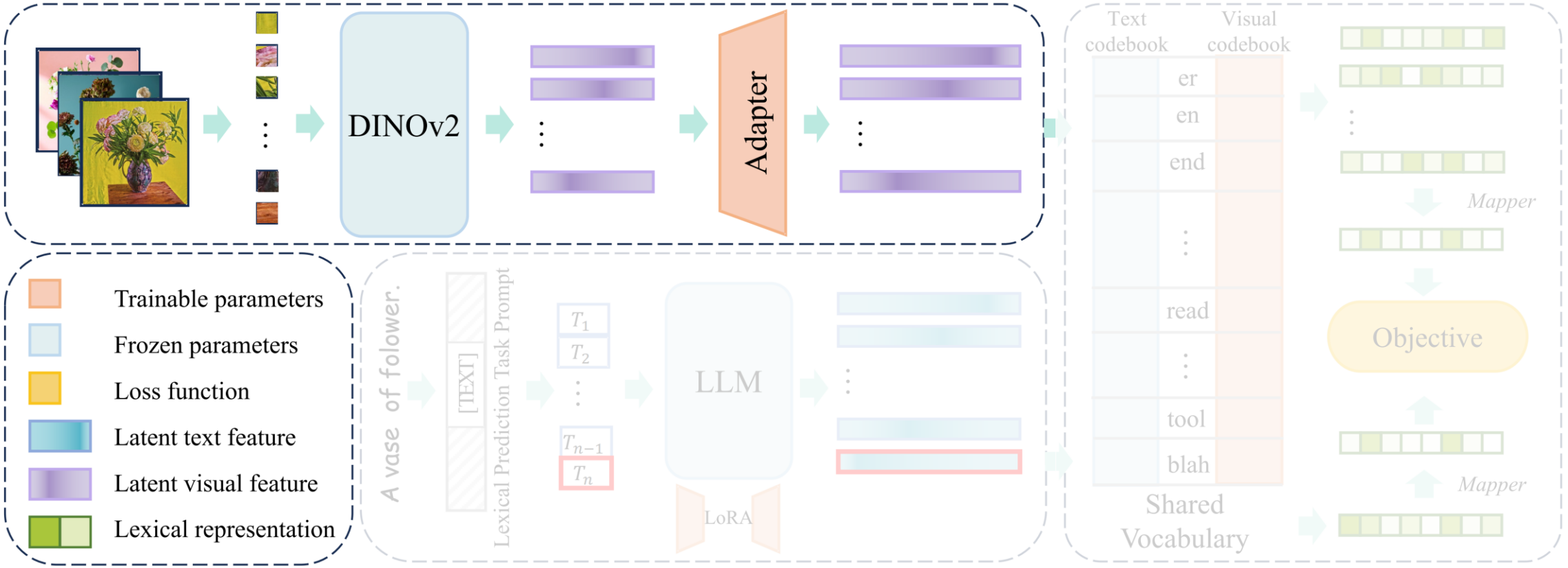

DINOv2
Adapter
Text codebook
Visual codebook
er
en
end
read
tool
blah
Shared Vocabulary
Mapper
Objective
Mapper
Trainable parameters
Frozen parameters
Loss function
Latent text feature
Latent visual feature
Lexical representation
A vase of folower.
[TEXT]
Lexical Prediction Task Prompt
T_1
T_2
T_{n-1}
T_n
LLM
LoRA

Method

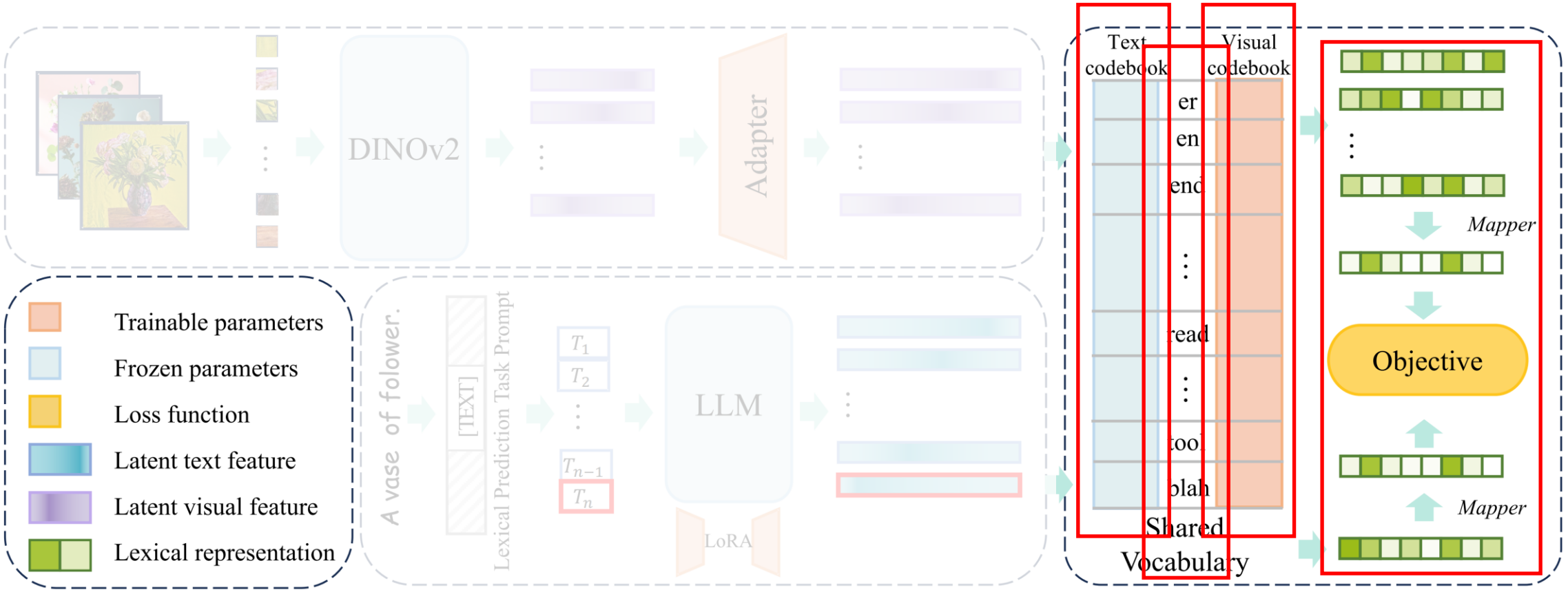


The focus of "The man is riding a white horse." lies on important words:"man", "riding", "white", "horse". The focus of "[TEXT]" lies on important words:

Method


DINOv2
Adapter
Trainable parameters
Frozen parameters
Loss function
Latent text feature
Latent visual feature
Lexical representation
A vase of folower.
[TEXT]
Lexical Prediction Task Prompt
T_1
T_2
T_{n-1}
T_n
LLM
LoRA
Text codebook
Visual codebook
er
en
end
read
tool
blah
Shared Vocabulary
Mapper
Objective
Mapper

Method



$$\ell_{\text{overuse}} = V \sum_{j=1}^{V} \frac{\bar{s}_{\cdot,j}}{\sum_{k=1}^{V} \bar{s}_{\cdot,k}} \bar{s}_{\cdot,j}^{2} = NV \sum_{j=1}^{V} \left(\sum_{i=1}^{N} s_{i,j}/N \right)^{3} / \sum_{j=1}^{V} \sum_{i=1}^{N} s_{i,j}.$$

PatchDis

horse

mIoU

Ground truth regions

Key Experimental Results

Table 1: Zero-shot cross-modal retrieval. 🔍 indicates variants of our LexVLA. CLIP[1] is the original CLIP [34]; results denoted by $(\cdot)^2$ are reported in VDR [48]; results denoted by $(\cdot)^3$ are reported in STAIR [5]. "Data" is the multi-modal alignment training data size; "Latent" means direct latent feature alignment methods; "Lexical" indicates lexical feature alignment methods. R@K, the recall ratio within top-K items.

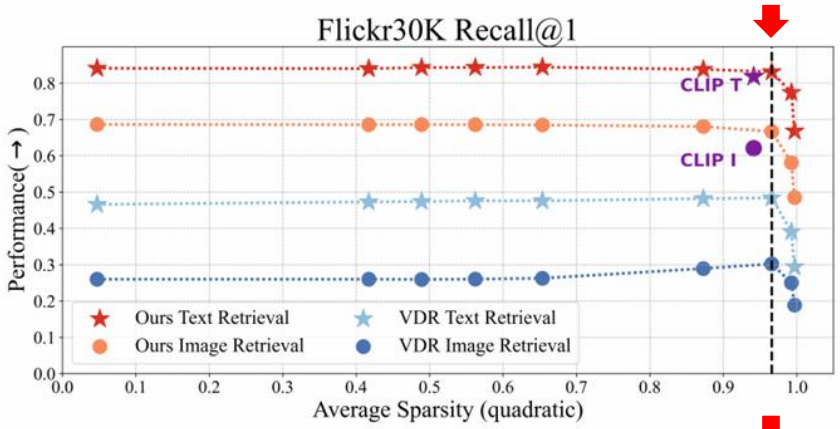
Setting	Model	Data	MSCOCO						Flickr30k					
			image-to-text			text-to-image			image-to-text			text-to-image		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Latent	CLIP[2]	15M	20.8	43.9	55.7	13.0	31.7	42.7	34.9	63.9	75.9	23.4	47.2	58.9
	FILIP[2]	15M	21.6	46.7	59.0	13.7	31.7	41.6	46.3	74.4	83.2	30.7	58.2	68.6
	CLIP-BERT[2]	15M	23.9	47.8	60.3	13.6	33.8	45.1	44.1	71.2	80.7	27.8	54.7	65.9
	DeCLIP[2]	15M	25.3	51.2	63.4	16.6	35.2	45.4	51.3	80.7	88.5	35.5	63.0	73.0
	SLIP[2]	15M	27.7	52.6	63.9	18.2	39.2	51.0	47.8	76.5	85.9	32.3	58.7	68.8
	ProtoCLIP[2]	15M	30.2	55.1	66.5	16.9	37.9	49.4	-	-	-	-	-	-
	CLIP[1]	0.4B	52.4	76.7	84.6	33.1	58.4	69.0	81.8	96.2	98.8	62.1	85.6	91.8
	CLIP[3]	1.1B	53.4	78.3	85.6	36.2	62.2	72.2	79.6	95.5	98.1	63.0	86.7	92.5
Lexical	VDR[2]	15M	30.9	54.5	65.4	17.4	38.1	49.7	51.0	79.3	86.7	32.4	60.1	70.7
	STAIR[3]	1.1B	**57.7**	80.5	87.3	**41.4**	65.4	75.0	81.2	96.1	98.4	66.6	88.7	93.5
Lexical	🔍(BoW)	12M	17.9	34.9	45.2	10.4	24.3	33.1	30.6	56.2	66.3	17.7	36.4	44.9
	🔍(CLIP)	12M	51.8	75.5	84.1	36.8	62.5	72.7	82.9	96.2	98.7	65.2	88.3	93.2
	🔍(FLOPs)	12M	56.2	80.0	87.4	39.0	65.7	75.6	84.2	96.6	98.7	67.4	89.4	94.1
	🔍(512)	12M	56.4	79.9	87.5	38.1	64.6	74.9	**84.5**	97.3	99.0	65.7	89.3	93.8
	LexVLA	12M	55.4	**80.6**	**88.3**	39.8	**66.3**	**76.2**	83.9	**97.5**	**99.1**	**67.8**	**90.2**	**94.2**

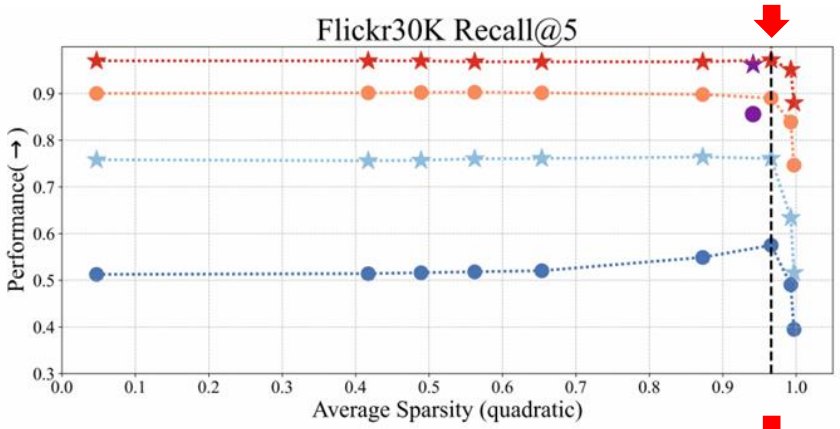
Key Experimental Results

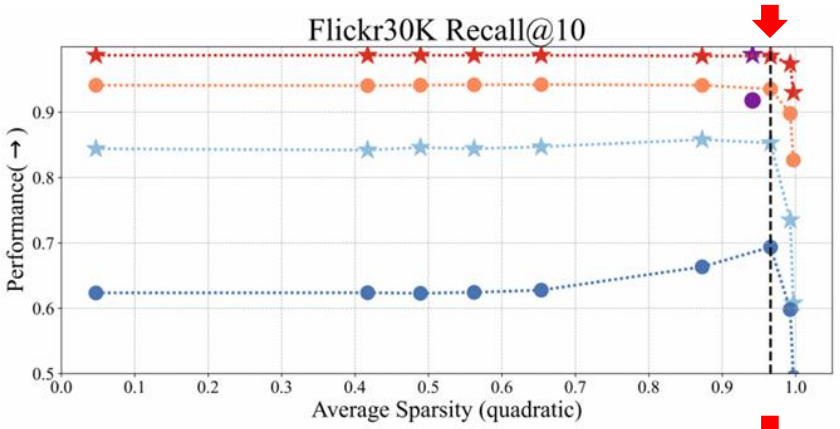
Table 2: PatchDis results.

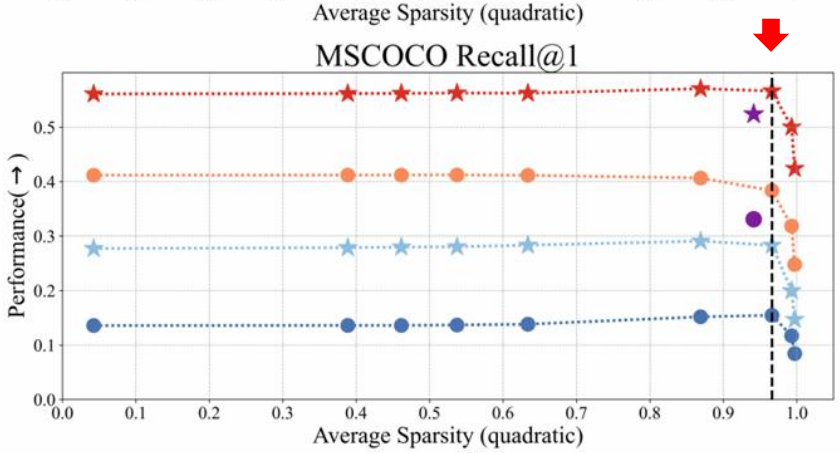
Model	mIoU
Random Dis.	5.0
CLIP	5.3
VDR	12.6
Q(CLIP)	13.9
LexVLA	**36.3**

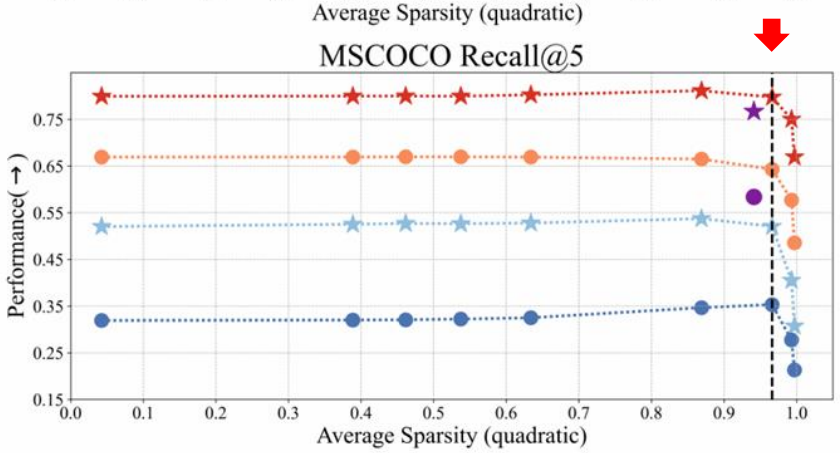
Key Experimental Results

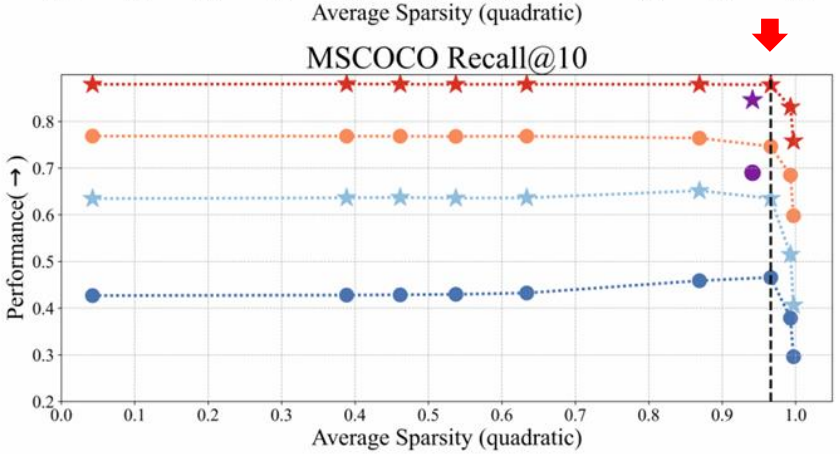






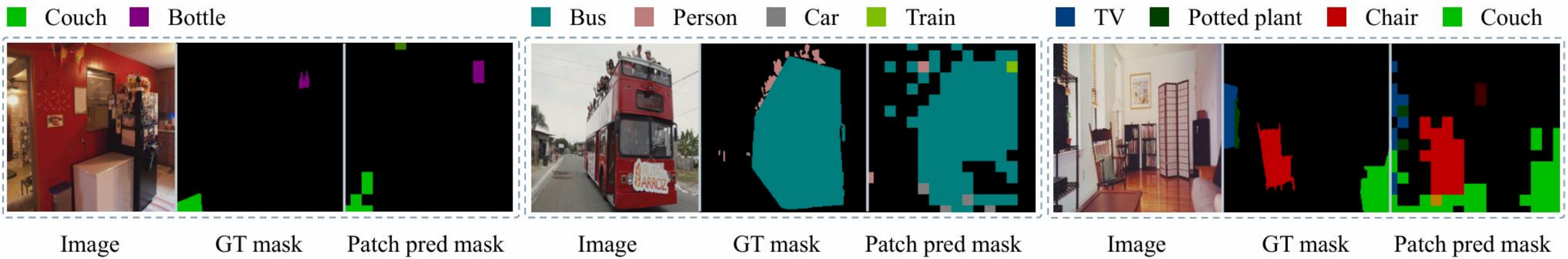




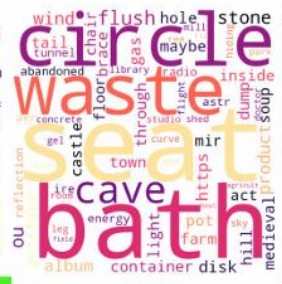


Interpretability

Interpretability



Interpretability

		FLOPs	Overuse
Image			
Caption	Toilet with indefinite end because of bright light.		

		FLOPs	Overuse
Image			
Caption	A cow stands in the back of a large truck.		

		FLOPs	Overuse
Image			
Caption	A picture of a young boy standing on a snowboard.		

Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you for your attention!