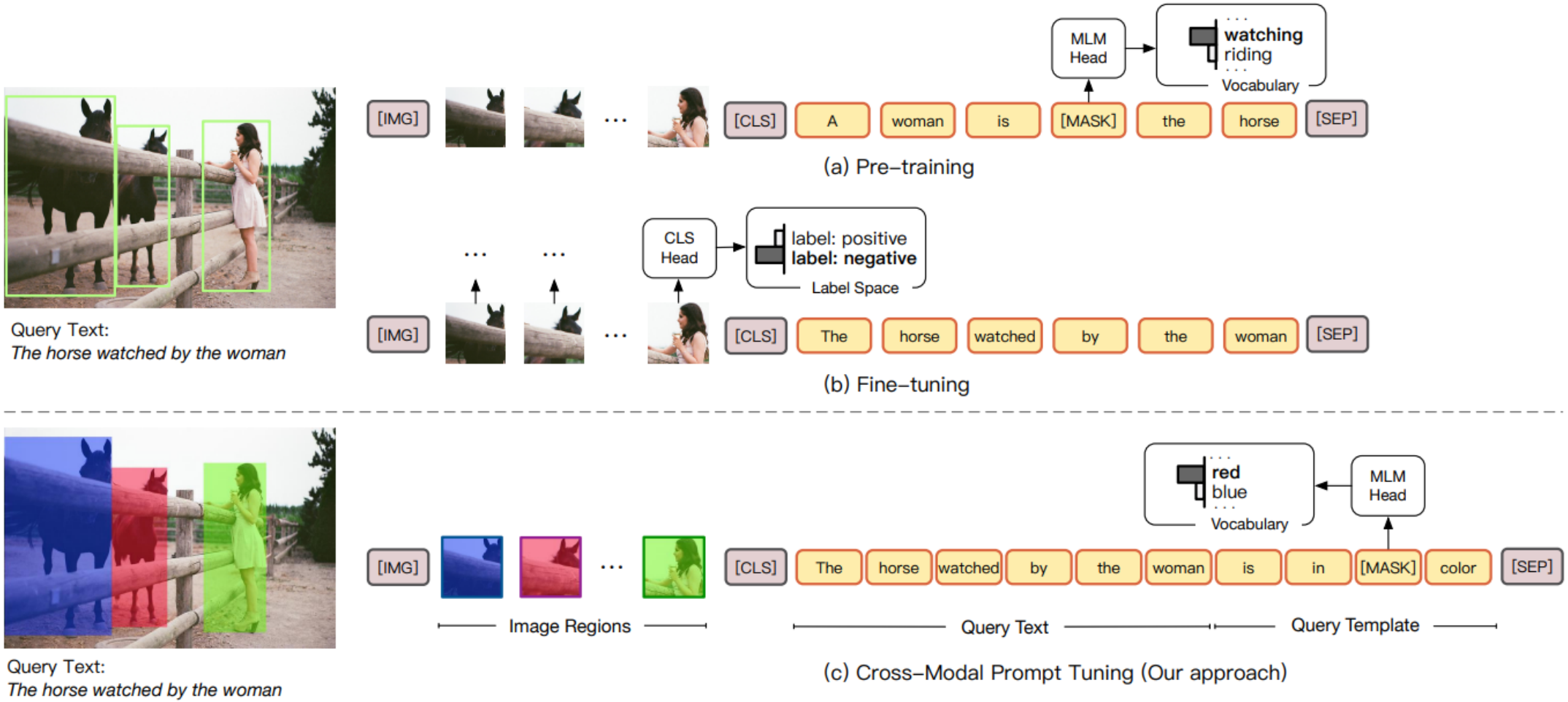
A Brief Intro to *Visual Prompting*

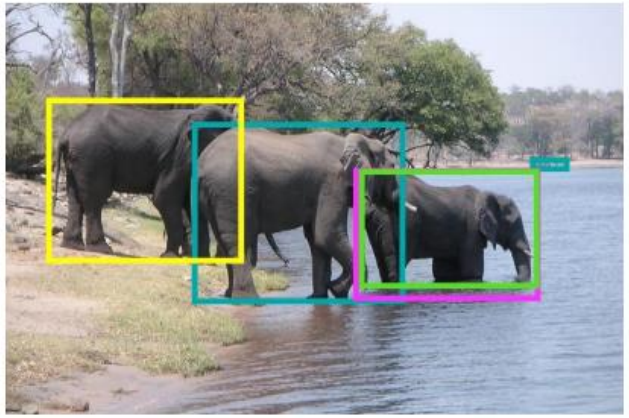
Haoquan Zhang 张皓泉

CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models



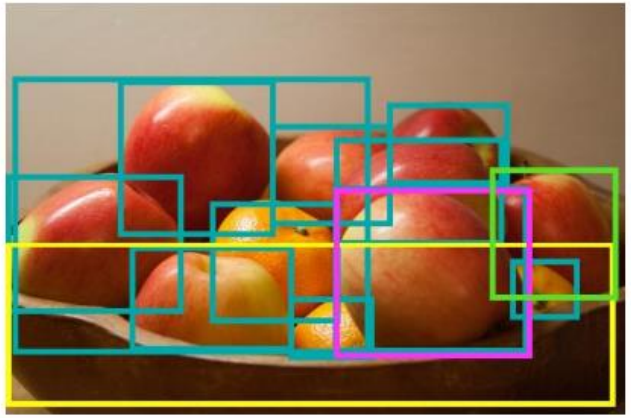
https://arxiv.org/pdf/2109.11797

CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models



Query Text: right elephant in water

(a) Correctly predicted



Query Text: apple on the bottom to the right of the orange in middle

(b) Disturbed by objects of the same type, but still reasonable

Query Text: food in red bowl

(c) Disturbed by colors in raw image regions and text

Figure 4: Case study. The bounding boxes given by image region proposals (olive), ground-truth annotation (pink), CPT (green), and fine-tuning baseline (yellow) are highlighted accordingly.

https://arxiv.org/pdf/2109.11797

CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models

	Shot	Model	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
ZS	0	Random	15.9 ± 0.2	19.4 ± 0.6	13.4 ± 0.4	16.1 ± 0.1	13.3 ± 0.6	20.0 ± 0.2	18.8 ± 0.4	19.2 ± 0.3
		CPT-Blk	25.7	25.4	27.0	25.9	25.8	25.7	32.9	32.6
		CPT-Seg	**29.5**	**30.6**	**28.7**	**28.8**	**30.3**	**27.4**	**34.6**	**34.8**
Few-Shot	1	Fine-tuning	18.5 ± 3.4	13.7 ± 4.8	25.0 ± 3.7	23.0 ± 6.5	22.8 ± 8.2	23.6 ± 4.5	30.6 ± 7.3	31.5 ± 7.4
		CPT-Blk	36.4 ± 3.5	39.1 ± 4.3	34.3 ± 2.7	34.4 ± 3.8	38.7 ± 5.4	31.2 ± 2.5	38.7 ± 4.8	38.7 ± 4.6
		CPT-Seg	$\mathbf{39.3 \pm 4.2}$	$\mathbf{43.2 \pm 5.6}$	$\mathbf{35.5 \pm 2.4}$	$\mathbf{35.9 \pm 3.8}$	$\mathbf{41.0 \pm 5.0}$	$\mathbf{31.2 \pm 2.8}$	$\mathbf{40.9 \pm 6.0}$	$\mathbf{41.0 \pm 6.1}$
	2	Fine-tuning	23.4 ± 3.5	21.1 ± 5.2	26.7 ± 4.5	28.3 ± 2.3	30.1 ± 5.3	26.4 ± 2.8	33.1 ± 8.3	33.4 ± 8.2
		CPT-Blk	38.3 ± 2.9	40.5 ± 4.2	35.3 ± 1.2	36.2 ± 5.5	41.1 ± 7.6	31.9 ± 3.3	40.6 ± 5.9	41.3 ± 6.1
		CPT-Seg	$\mathbf{41.4 \pm 1.5}$	$\mathbf{45.8 \pm 3.6}$	$\mathbf{36.6 \pm 2.0}$	$\mathbf{38.7 \pm 3.8}$	$\mathbf{44.7 \pm 5.2}$	$\mathbf{33.5 \pm 2.6}$	$\mathbf{43.2 \pm 5.9}$	$\mathbf{43.4 \pm 5.8}$
	4	Fine-tuning	27.8 ± 4.8	26.0 ± 7.8	30.1 ± 3.4	33.4 ± 3.5	36.8 ± 5.1	28.3 ± 2.1	36.9 ± 8.9	37.2 ± 8.7
		CPT-Blk	40.9 ± 1.8	45.0 ± 2.0	36.6 ± 1.6	37.2 ± 3.6	42.4 ± 5.4	33.6 ± 2.3	42.2 ± 6.5	42.7 ± 6.9
		CPT-Seg	$\mathbf{41.3 \pm 5.2}$	$\mathbf{45.9 \pm 7.1}$	$\mathbf{36.5 \pm 3.7}$	$\mathbf{39.8 \pm 3.8}$	$\mathbf{45.7 \pm 5.7}$	$\mathbf{34.1 \pm 1.8}$	$\mathbf{45.7 \pm 7.3}$	$\mathbf{45.8 \pm 7.6}$
	8	Fine-tuning	33.3 ± 4.2	35.6 ± 7.4	31.2 ± 2.7	38.1 ± 3.7	43.5 ± 3.9	31.2 ± 3.8	41.9 ± 8.0	42.5 ± 7.9
		CPT-Blk	42.7 ± 4.1	48.4 ± 5.7	37.3 ± 2.4	39.9 ± 2.2	45.8 ± 3.0	34.6 ± 2.1	44.8 ± 4.1	45.5 ± 4.6
		CPT-Seg	$\mathbf{45.2 \pm 3.6}$	$\mathbf{51.4 \pm 4.9}$	$\mathbf{38.7 \pm 2.4}$	$\mathbf{42.4 \pm 3.8}$	$\mathbf{49.0 \pm 4.9}$	$\mathbf{35.7 \pm 1.8}$	$\mathbf{48.1 \pm 5.4}$	$\mathbf{48.6 \pm 5.8}$
	16	Fine-tuning	38.4 ± 2.4	42.8 ± 4.2	33.4 ± 2.5	40.7 ± 3.2	45.6 ± 3.5	34.7 ± 2.8	48.7 ± 3.5	49.4 ± 3.5
		CPT-Blk	45.7 ± 2.5	53.0 ± 3.2	37.9 ± 1.5	41.8 ± 2.0	48.8 ± 2.6	35.7 ± 1.4	47.7 ± 2.4	48.6 ± 2.8
		CPT-Seg	$\mathbf{48.6 \pm 3.1}$	$\mathbf{55.9 \pm 3.5}$	$\mathbf{40.3 \pm 2.0}$	$\mathbf{43.8 \pm 2.0}$	$\mathbf{50.9 \pm 2.5}$	$\mathbf{36.5 \pm 1.3}$	$\mathbf{50.8 \pm 3.6}$	$\mathbf{51.6 \pm 3.7}$
Fully Supervised	$\lvert\mathcal{D}_{\text{train}}\rvert$	MAttNet	76.7	81.1	70.0	65.3	71.6	52.0	66.6	67.3
		ViLBERT	-	-	-	72.3	78.5	62.6	-	-
		VLBERT	-	-	-	72.6	78.6	62.3	-	-
		ERNIE-ViL	-	-	-	**76.0**	**82.1**	**66.9**	-	-
		UNITER	81.4	87.0	74.2	75.9	81.5	66.7	**74.9**	**75.8**
		Fine-tuning	81.8	**87.5**	73.7	74.8	81.0	64.1	74.7	**75.8**
		CPT-Blk	81.5	87.0	**74.3**	73.6	80.1	64.1	74.1	75.2
		CPT-Seg	**81.8**	87.3	74.1	74.1	79.5	63.8	73.6	74.7

https://arxiv.org/pdf/2109.11797

Fine-Grained Visual Prompting

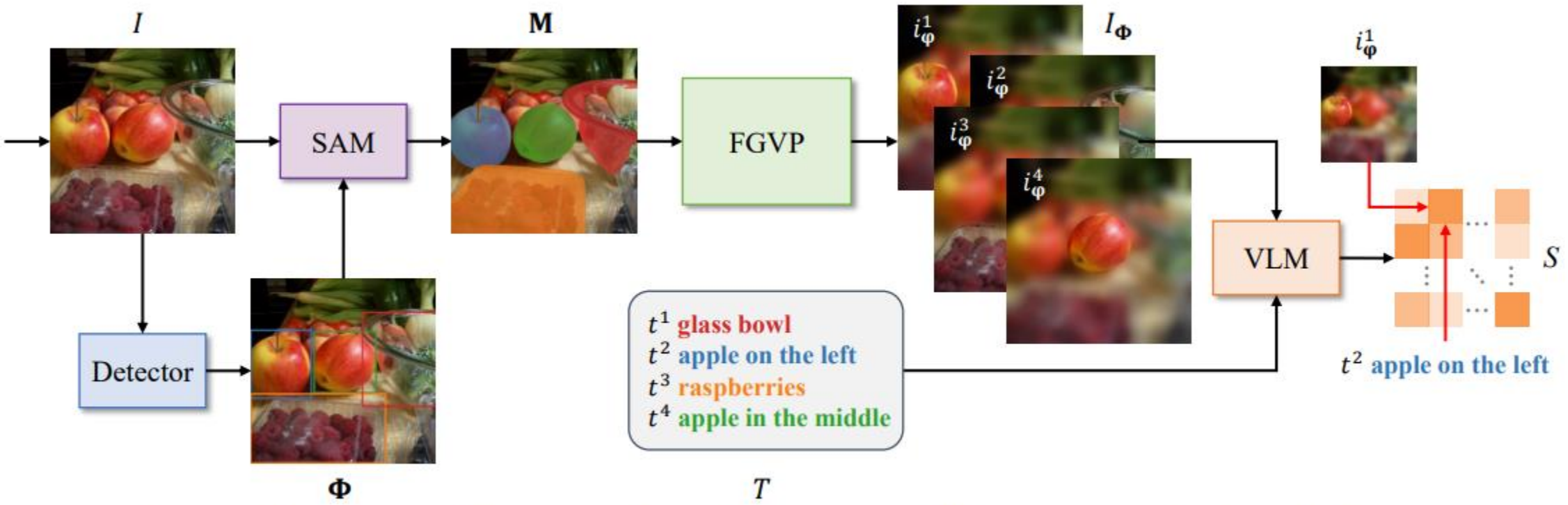


Figure 2: Structure of fine-grained visual prompting with box proposals from a detector.

https://arxiv.org/pdf/2306.04356

Fine-Grained Visual Prompting

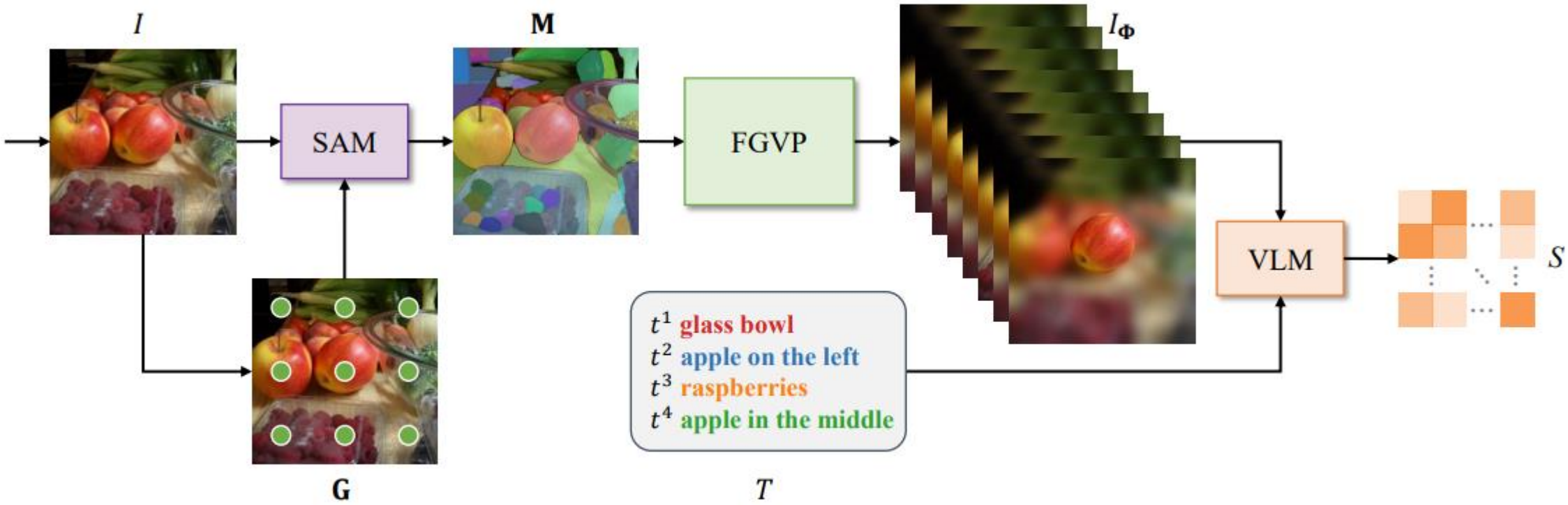


Figure 3: Structure of fine-grained visual prompting with no box proposal. Masks are directly derived via SAM prompted by grid-wise keypoints.

https://arxiv.org/pdf/2306.04356

Fine-Grained Visual Prompting

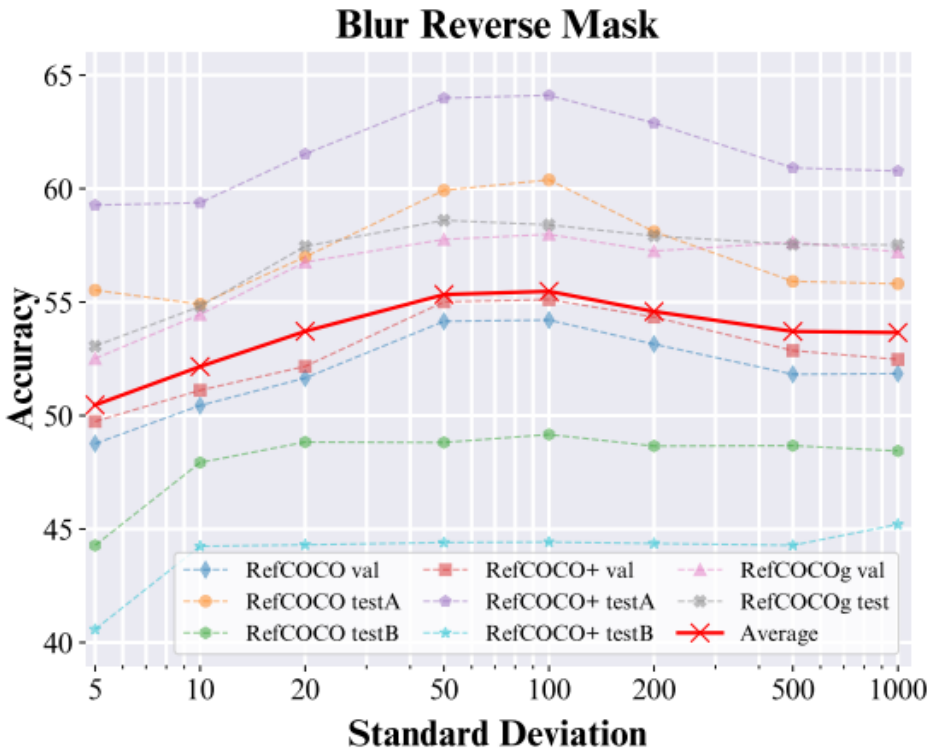
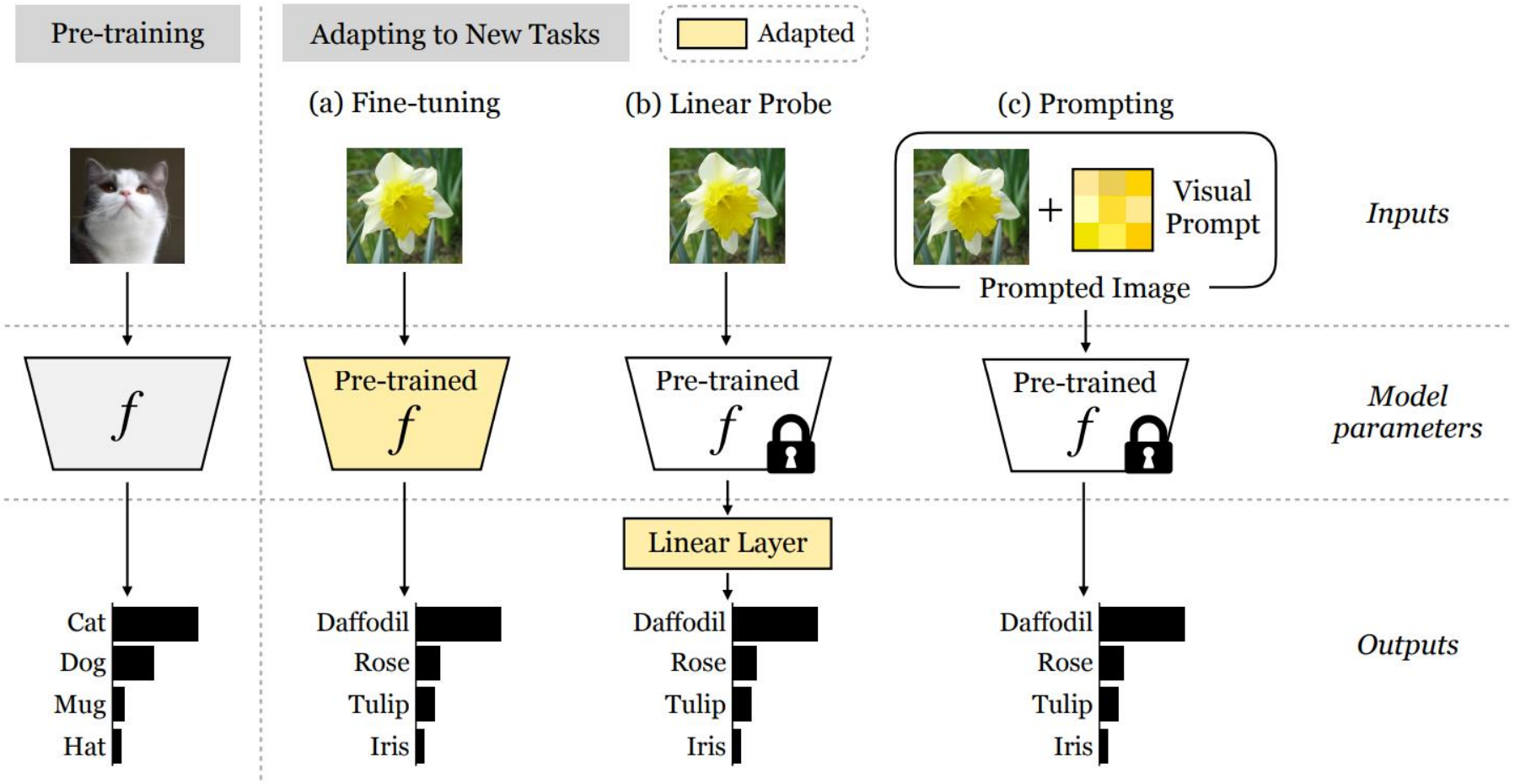


Figure 5: Ablation on the standard deviation in Gaussian blur kernel from the Blur Reverse Mask [$D4$] prompting. A larger deviation presents a more blurred background.

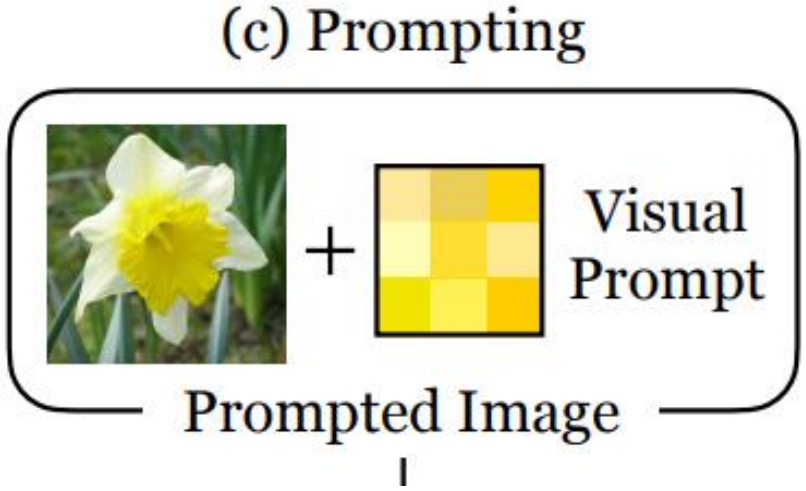
https://arxiv.org/pdf/2306.04356

Exploring Visual Prompts for Adapting Large-Scale Models



https://arxiv.org/pdf/2203.17274

What is Visual Prompt? How does it work?

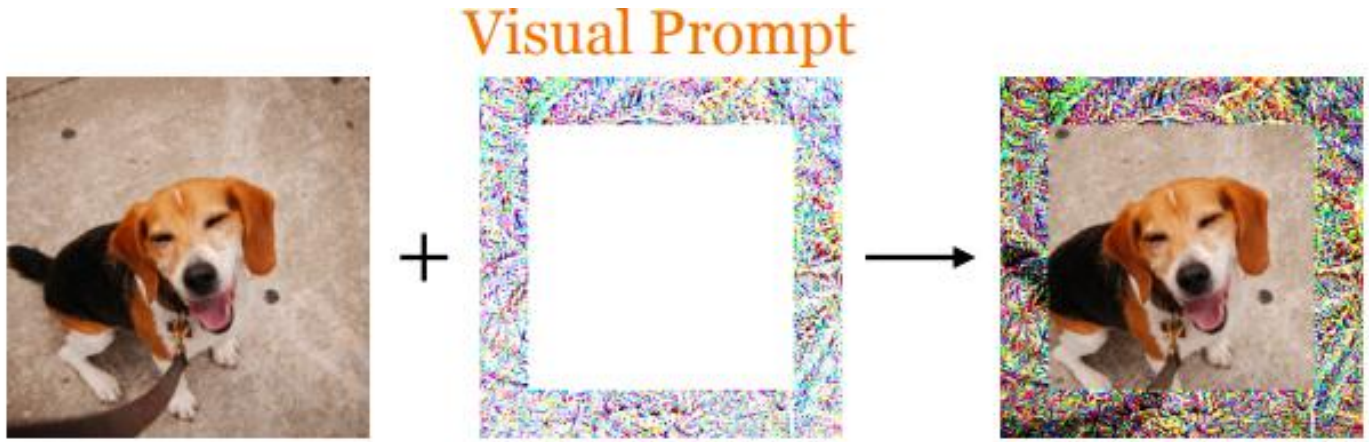


https://arxiv.org/pdf/2203.17274

What is Visual Prompt? How does it work?

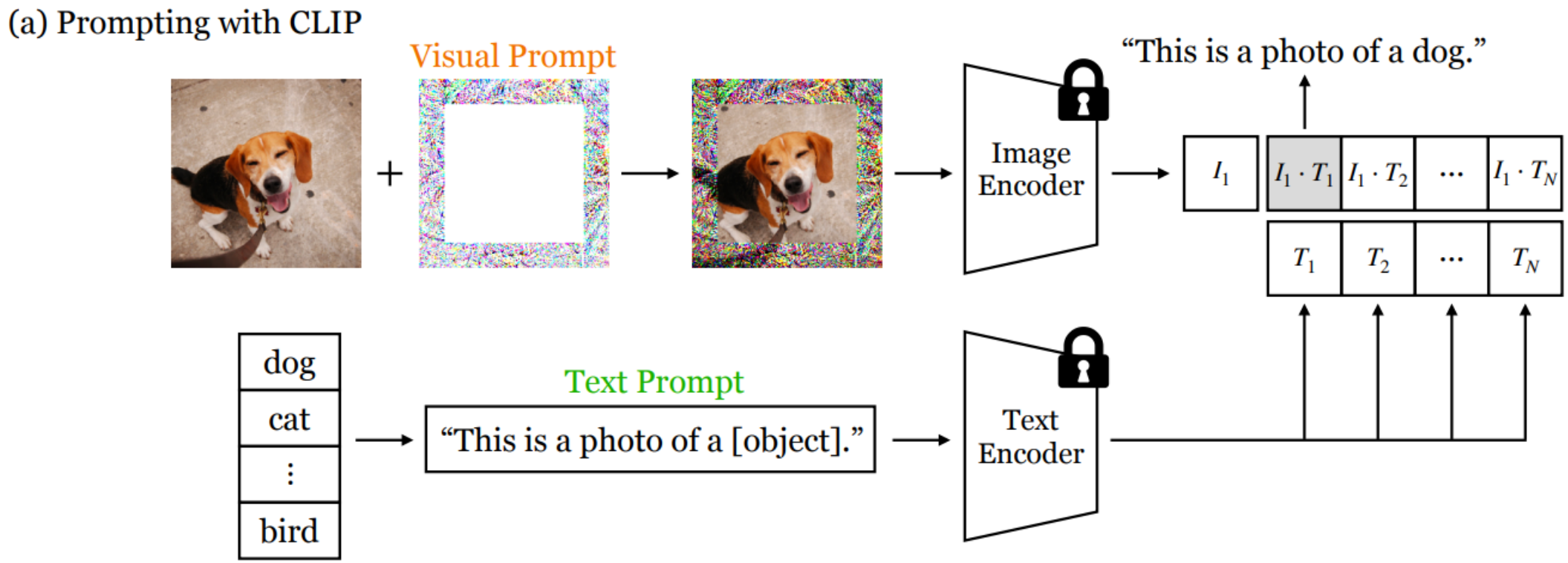
https://arxiv.org/pdf/2203.17274

What is Visual Prompt? How does it work?



https://arxiv.org/pdf/2203.17274

What is Visual Prompt? How does it work?



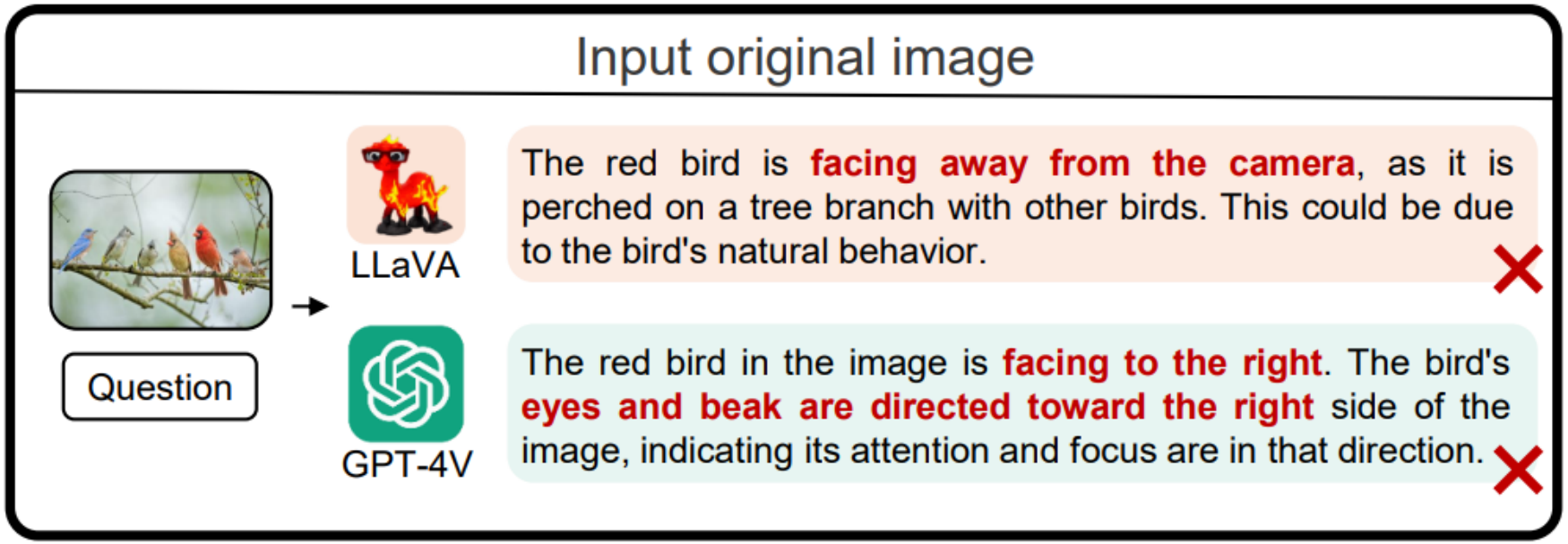
https://arxiv.org/pdf/2203.17274

Attention Prompting on Image for Large Vision-Language Models

Which direction is the red bird facing?

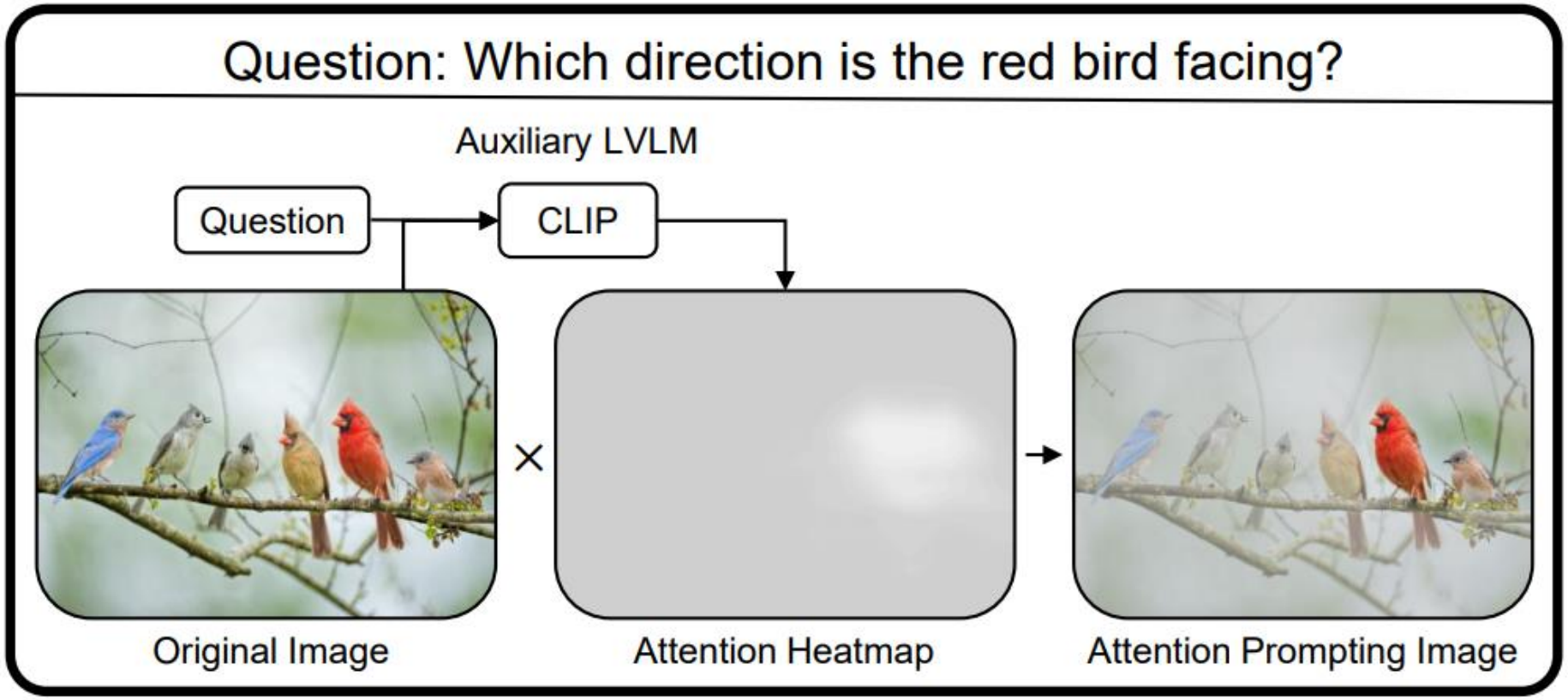
https://arxiv.org/pdf/2409.17143

Attention Prompting on Image for Large Vision-Language Models



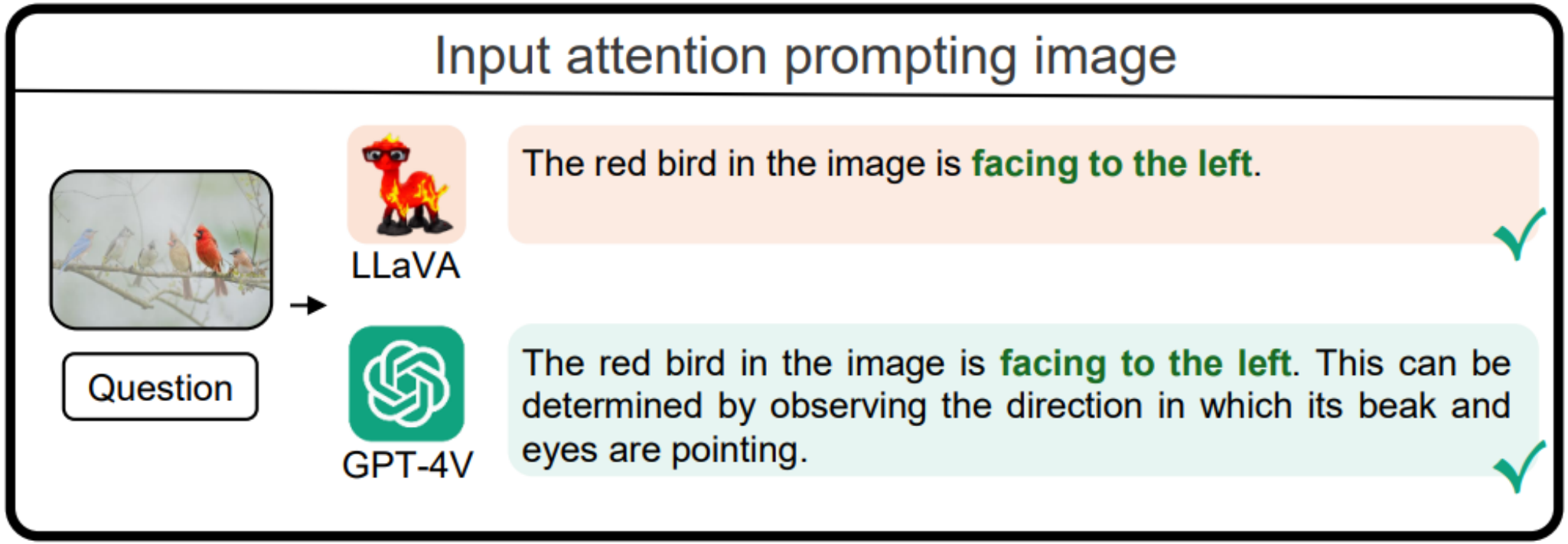
https://arxiv.org/pdf/2409.17143

Attention Prompting on Image for Large Vision-Language Models



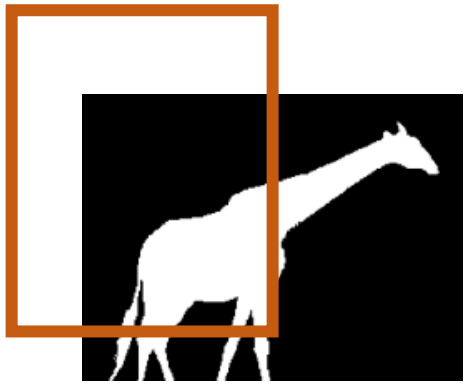
https://arxiv.org/pdf/2409.17143

Attention Prompting on Image for Large Vision-Language Models



https://arxiv.org/pdf/2409.17143

Summary
Visual Prompting

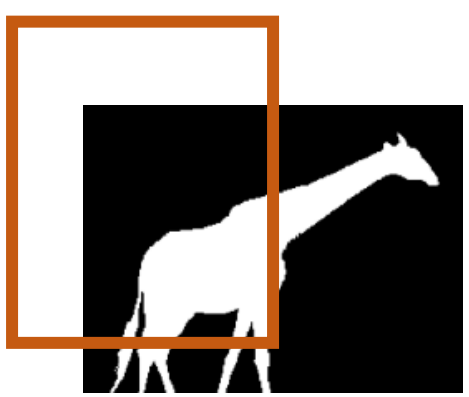


Bbox Mask

Blur

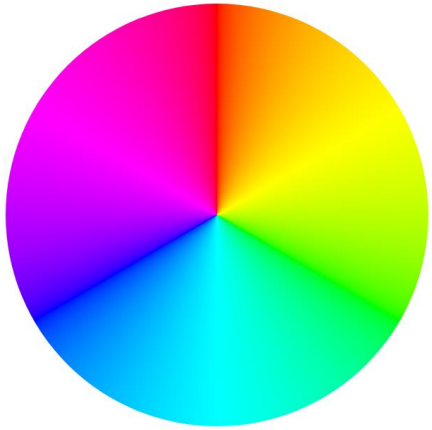
Pixel Prompt

Attention

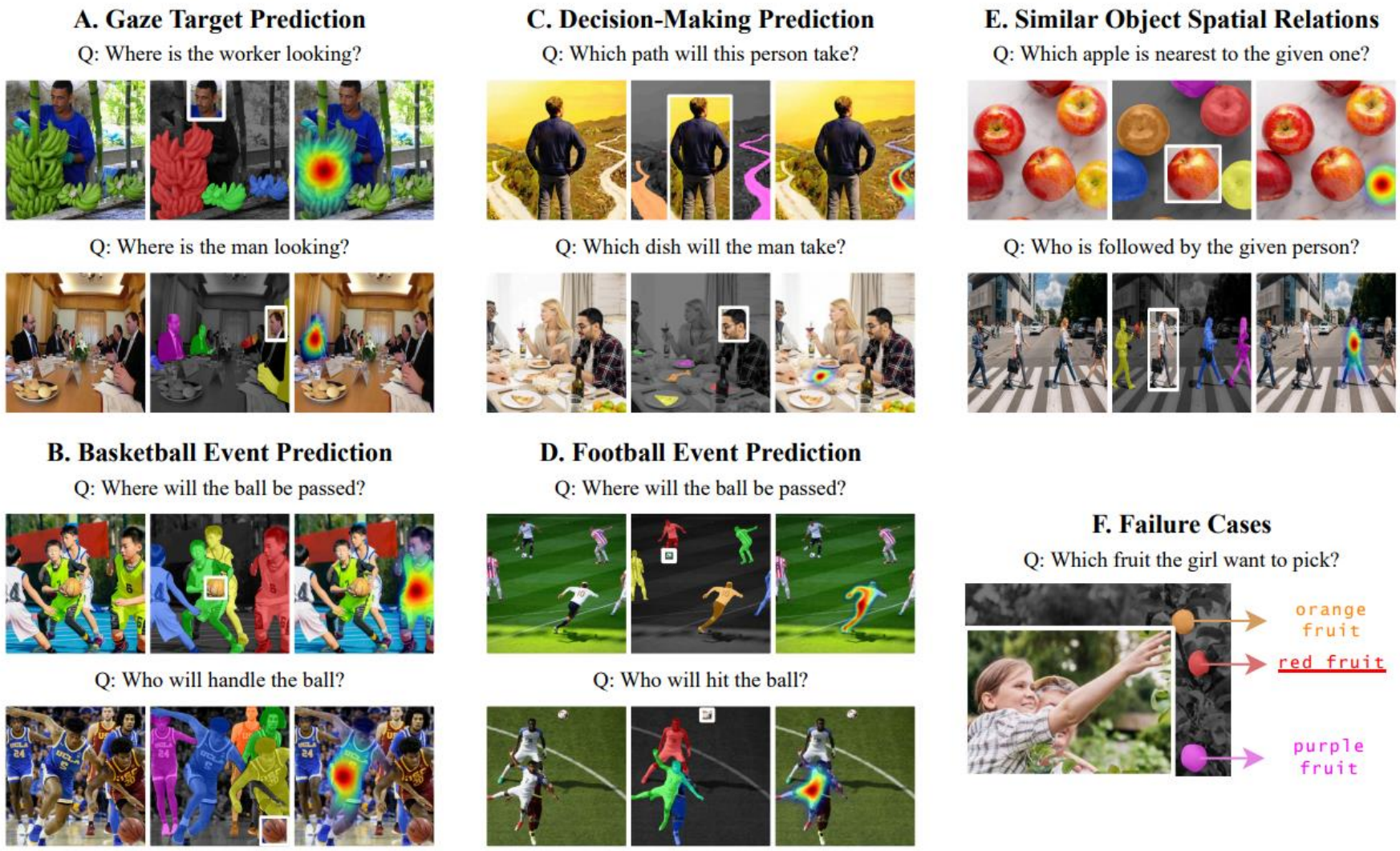


Bbox Mask

X



Color



Haoquan Zhang et.al, Mask4Align, CVPR2024