Fine-Grained Visual Prompting

Lingfeng Yang

paper

Background

- **Visual-language models (VLMs) have impressive zero/few-shot transfer capabilities in image-level visual understanding**

  - Image Editing
  - Image Captioning
  - Image Counting
  - Visual Question Answering

Input	Completion
 Question: Describe the image. Answer:	A German shepherd wearing a harness and holding a leash.
Question: On which suitcases is the dog's paw resting? Answer:	brown suitcase

Results are generated by **open flamingo** (https://laion.ai/blog/open-flamingo)

Background

- **However, they are not good at instance-level object grounding**

Input	What we need
 Question: On which suitcases is the dog's paw resting? Answer:	

- **Unless costly additional designing and training**

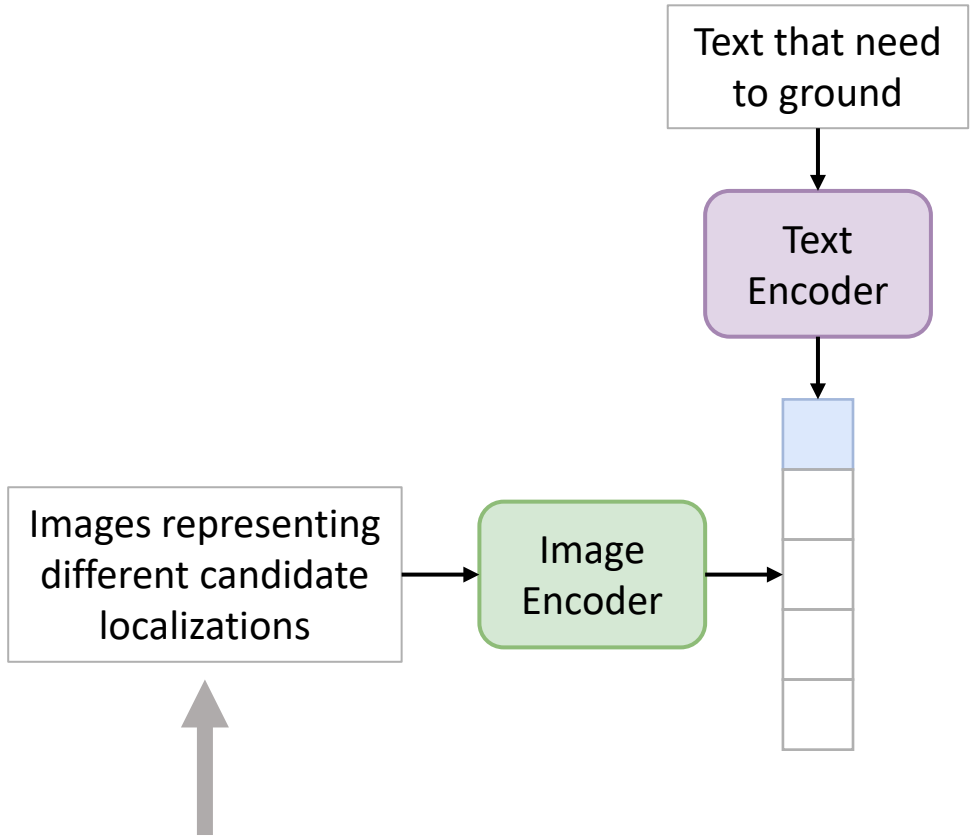
- UNITER
- Pix2Seq
- KOSMOS-2
- VisionLLM

<s> <image> Image Embedding </image> <grounding> It <box><loc44><loc863></box> seats next to a campfire <box><loc4><loc1007></box> </s>

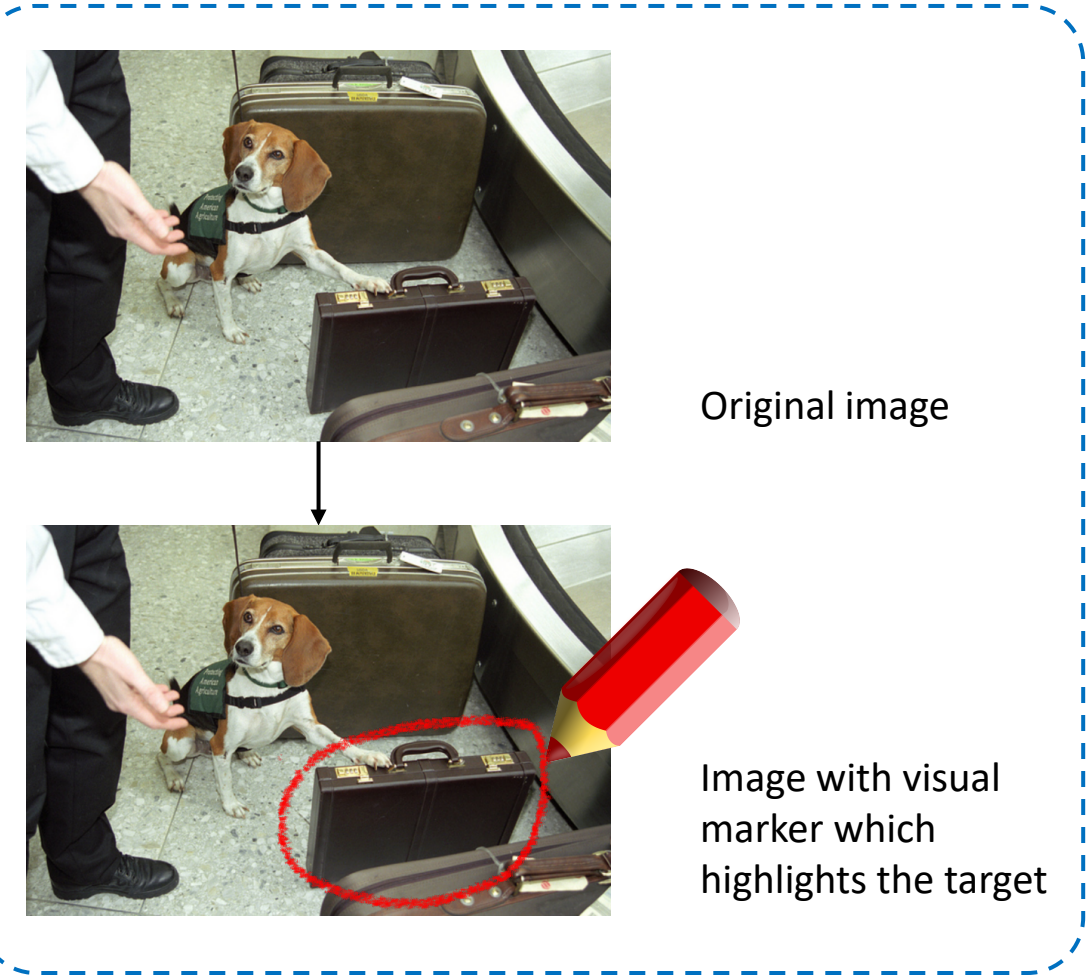
Training Sample

Background

- **Design a SIMPLE zero-shot grounding architecture leveraging CLIP**



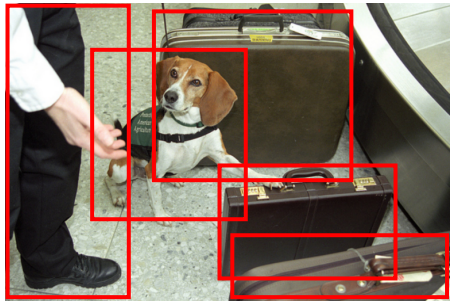
- **How? Through Visual Prompting**



Thinking

- **How to design a better visual prompt?**
- **Pipeline**
  - Detect possible regions as proposal bounding
  - Visual prompting the image with proposals
  - Calculate similarities between texts using CLIP

Detect



Existing Visual Prompt Methods

Crop

CLIP-V

CLIP-T

suitcase dogs paw is on

CPT
(Arxiv 2021)

CLIP-V

CLIP-T

suitcase dogs paw is on
is in red color

RedCircle
(Arxiv 2023)

CLIP-V

CLIP-T

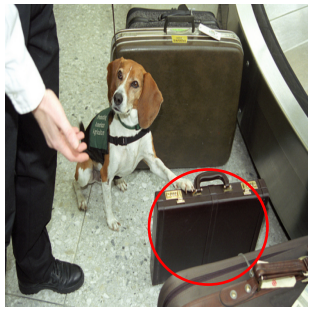
suitcase dogs paw is on

Thinking

- **Using more precise marking as visual prompt**

Crop

CPT



RedCircle

Existing Methods

Drawbacks of existing works

- **Missing global information**
- The marking is too **coarse** to highlight the target
- It brings **unrelated background noise**

FGVP (blur)

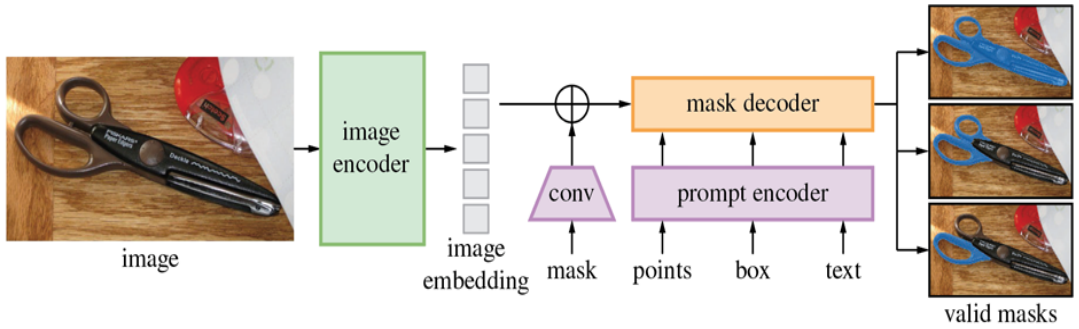
Our method

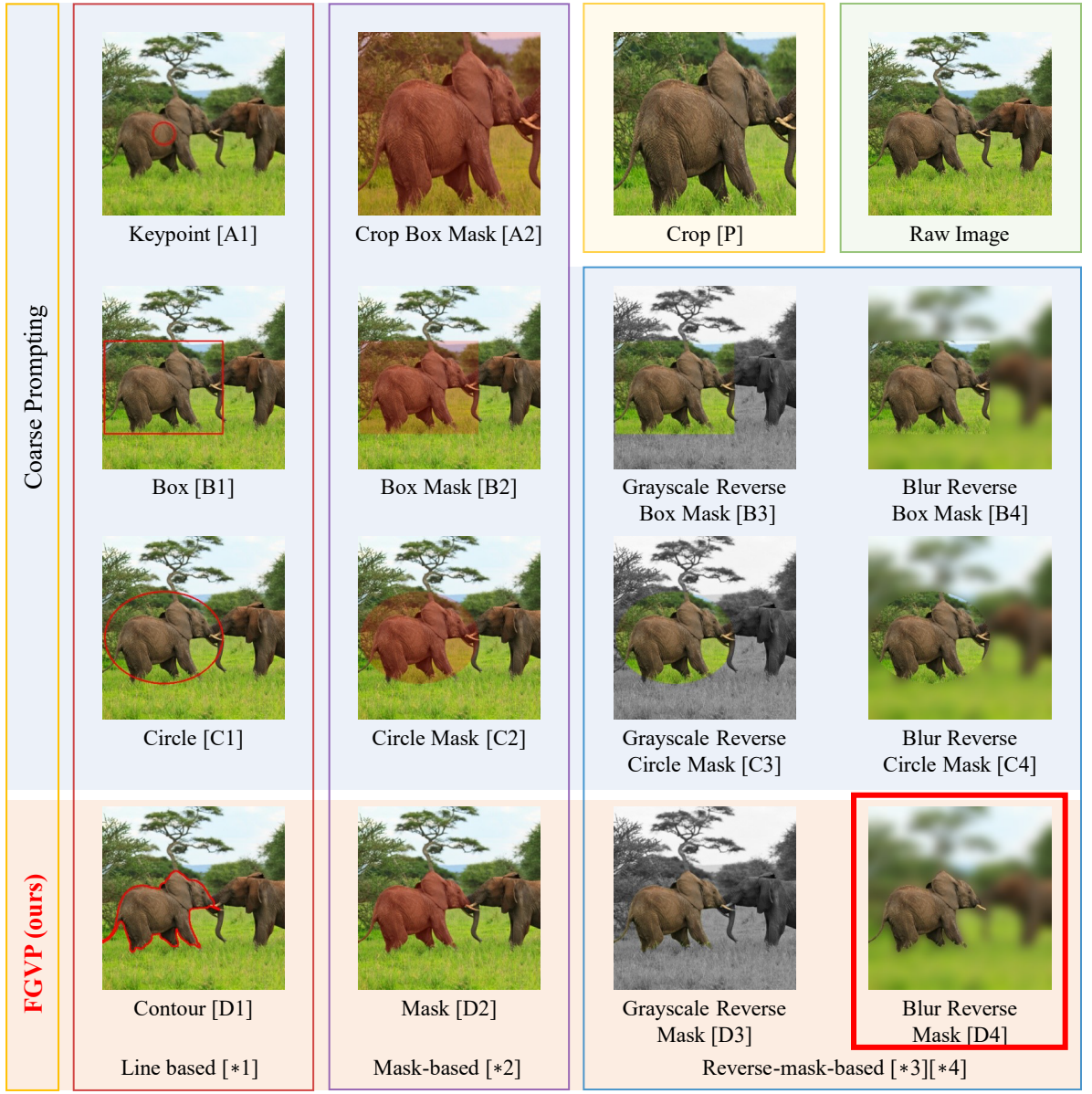
Benefits of our method

- The target is accurately indicated through **fine-grained marking**
- Retain **global information**
- Highlighting the target while keeping **spatial correlation**

Fine-Grained Visual Prompting (FGVP)

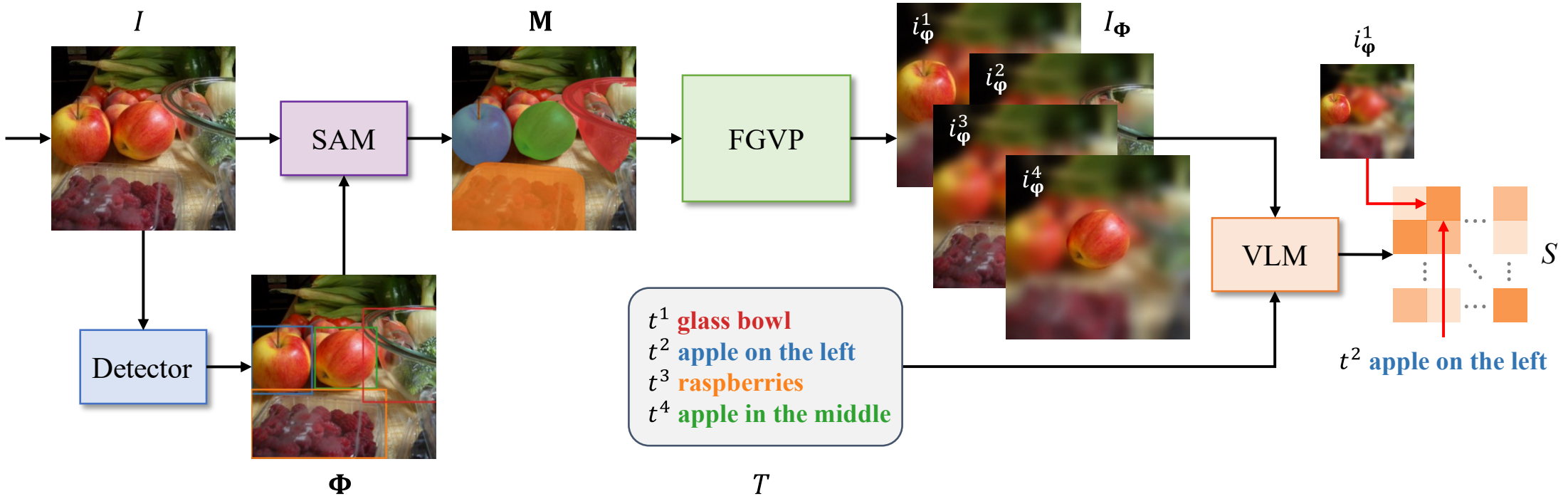
- **Summary of visual prompts**
  - The semantic mask can be generated from segmentors such as Segment Anything Model (SAM)





Method

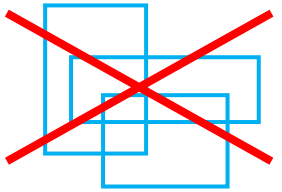
- **Pipeline (proposals available)**
  - Detect possible regions as proposal bounding
  - Generate semantic masks using SAM
  - Fine-grained visual prompting the image with blurred masks
  - Calculate similarities between texts using CLIP



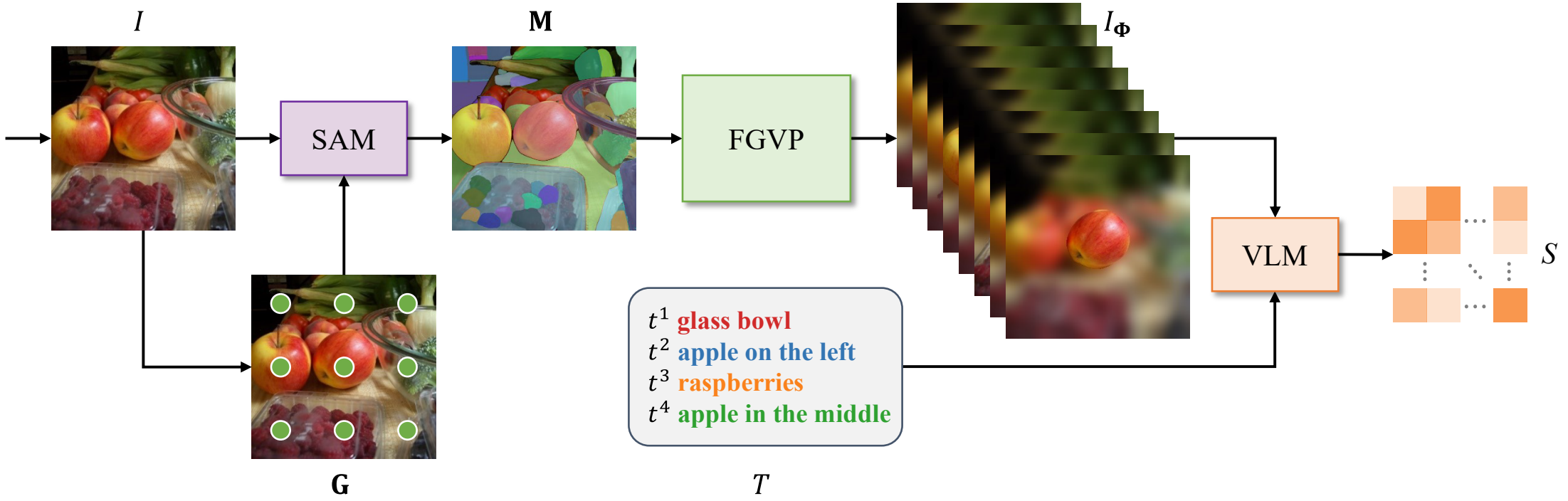
Method

- **Pipeline (no proposals)**
  - Generate grid-wise points
  - Generate semantic masks using SAM
  - Fine-grained visual prompting the image with blurred masks
  - Calculate similarities between texts using CLIP

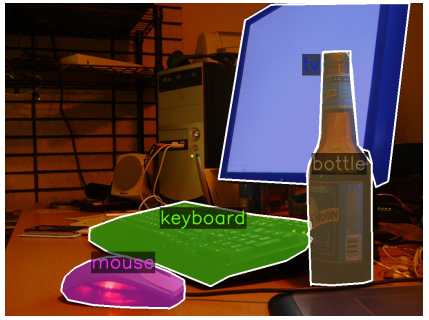
No detector



No proposal bounding boxes



Evaluation

	Text Input	Image Input
Task: Zero-shot **Referring Expression Comprehension (REC)** **Benchmark:** RefCOCO, RefCOCO+, RefCOCOg	suitcase dog not looking at suitcase dogs paw is on briefcase case closest to us	
Task: Zero-shot **Object Grounding** **Benchmark:** COCO	a photo of <object>	
Task: Zero-shot **Object and Part Grounding** **Benchmark:** PACO	The <part> of <object>	

Evaluation

- **Ablation study on individual performance**

  • Zero-shot performance of individual visual prompting in the validation set of COCO, PACO, RefCOCO, RefCOCO+, and RefCOCOg datasets using ground-truth annotations (left) and proposals in referring expression comprehension (right), respectively. VP: Visual Prompt. GT: Ground-Truth. REC: Referring Expression Comprehension.

VP	GT					REC		
	COCO	PACO	RefCOCO	RefCOCO+	RefCOCOg	RefCOCO	RefCOCO+	RefCOCOg
P	**70.9**	38.5	35.2	40.3	59.1	45.3	46.4	56.4
A1	52.3	39.1	36.9	39.6	43.8	46.7	47.9	48.9
A2	64.2	35.7	37.1	41.9	58.0	48.2	49.0	57.0
B1	48.5	42.7	34.7	39.5	44.6	45.5	46.4	47.0
B2	34.4	37.2	23.9	23.4	22.7	35.4	30.7	30.8
B3	42.4	37.4	34.4	35.9	44.5	45.9	44.0	48.4
B4	62.1	39.2	47.9	51.8	**63.6**	48.8	51.4	54.1
C1	48.9	42.6	43.2	49.3	56.3	48.9	51.7	54.6
C2	36.1	37.2	29.9	29.8	24.5	40.7	37.1	37.9
C3	42.9	36.6	36.9	38.2	47.3	47.8	46.2	50.3
C4	58.1	36.8	49.2	53.1	60.9	49.3	52.1	52.2
D1	47.3	41.0	38.7	41.7	43.5	46.1	45.0	46.3
D2	41.1	**43.7**	29.9	29.1	29.9	41.8	38.5	38.4
D3	45.2	40.4	40.5	43.8	50.9	45.8	45.9	51.2
D4	67.8	43.3	**52.8**	**58.0**	63.5	**52.8**	**55.4**	**57.8**

Evaluation

• Referring Expression Comprehension

• The performance of referring expression comprehension benchmarked with RefCOCO, RefCOCO+, and RefCOCOg datasets. VLM: Vision-Language Model. VP: Visual Prompt. PP: Post Processing, “R” and “S” denote Relations and Subtraction, respectively. * denotes our implementation. FGVP: Fine-Grained Visual Prompting. The best result for each dataset, w.r.t, each codebase is in bold.

Method	CLIP	VP	PP	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val	test
CPT	ViT-B, RN	P | B2	R	41.3	40.6	44.0	41.3	41.8	41.1	51.3	51.2
ReCLIP	ViT-B, RN	P | B4	R	45.8	46.1	47.1	47.9	50.1	45.1	59.3	59.0
RedCircle*	ViT-B, RN	P | C1	R	43.9	46.2	44.1	45.3	47.9	43.1	57.3	56.3
FGVP (ours)	ViT-B, RN	P | D4	R	52.0	55.9	48.8	53.3	60.4	46.7	62.1	61.9
RedCircle	ViT-L, RN	C1 | C3 | C4	S	49.8	58.6	39.9	55.3	63.9	45.4	59.4	58.9
RedCircle*	ViT-L, RN	C1 | C3 | C4	S	51.4	58.3	40.9	56.3	63.6	45.8	58.3	58.0
FGVP (ours)	ViT-L, RN	D1 | D3 | D4	S	52.9	59.6	43.9	57.4	64.8	46.3	58.1	58.3
RedCircle*	ViT-L, RN	P | C1 | C3 | C4	S	51.6	58.0	42.0	58.1	64.5	47.5	60.0	59.3
FGVP (ours)	ViT-L, RN	P | D1 | D3 | D4	S	53.9	60.2	44.3	59.3	66.6	48.8	61.0	61.3
RedCircle*	ViT-L, RN	P | C1 | C3 | C4	RS	56.8	62.4	49.1	58.6	64.7	48.3	62.2	61.0
FGVP (ours)	ViT-L, RN	P | D1 | D3 | D4	RS	**59.6**	**65.0**	**52.0**	**60.0**	**66.8**	**49.7**	**63.3**	**63.4**

Evaluation

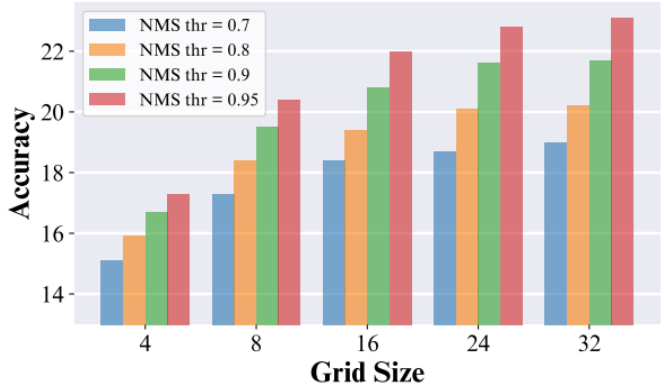
- **Referring Expression Comprehension**
  - Compare with full/weak supervised methods

Method	Published	Supervision	RefCOCO			RefCOCO+			RefCOCOg	
			val	test-A	test-B	val	test-A	test-B	val	test
MAttNet	CVPR'18	Full	76.7	81.1	70.0	65.3	71.6	56.0	66.6	67.3
NMTree	ICCV'19		76.4	81.2	70.1	66.5	72.0	57.5	65.9	66.4
ReSC	ECCV'20		77.6	80.5	72.3	63.6	68.4	56.8	67.3	67.2
TransVG	ICCV'21		80.3	82.7	78.1	63.5	68.2	55.6	67.7	67.4
VC	CVPR'18	Weak	--	33.3	30.1	--	34.6	31.6	--	--
ARN	ICCV'19		34.3	36.4	33.1	34.5	36.0	33.8	--	--
KPRN	ACMMM'19		35.0	34.7	37.0	36.0	35.2	37.0	--	--
DTWREG	TPAMI'21		39.2	41.1	37.7	39.2	40.1	38.1	--	--
CPT	ArXiv'21	8-shot	41.3	48.2	35.7	42.6	49.3	35.4	47.4	47.4
	ArXiv'21	4-shot	40.7	47.4	35.3	40.3	46.5	34.5	44.4	44.4
	ArXiv'21	2-shot	39.8	45.6	33.9	38.6	44.5	32.8	44.7	44.3
	ArXiv'21	1-shot	37.2	41.5	33.2	37.9	42.3	33.9	43.1	43.4
Pseudo-Q	CVPR'22	zero-shot	56.0	58.3	54.1	38.9	45.1	32.1	46.3	47.4
ReClip	ArXiv'22		45.8	46.1	47.1	47.9	50.1	45.1	59.3	59.0
RedCircle	ArXiv'23		49.8	58.6	39.9	55.3	63.9	45.4	59.4	58.9
FGVP (ours)	ArXiv'23		59.6	65.0	52.0	60.0	66.8	49.7	63.3	63.4

Evaluation

- **Object and Part Grounding**
  - Accuracy of the part detection with ViT-L and ablation study on the NMS threshold and grid size.

VP	PACO	RefCOCO	RefCOCO+	RefCOCOg
P	16.5	17.7	21.6	34.3
A1	11.9	16.7	18.5	19.7
C1	17.4	24.9	29.8	32.4
D2	15.2	24.1	21.4	18.6
D4	**18.3**	**40.8**	**44.9**	**49.6**



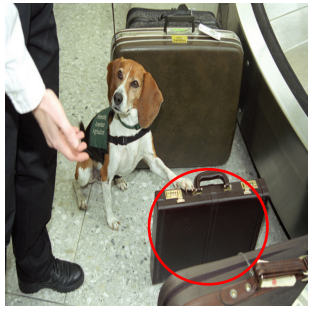
  - Intermediate results

Image | Grid-wise points | Proposals | Prompting (Sample) | Prediction | Ground-Truth

Analysis

Crop

CPT



RedCircle

Existing Methods

FGVP (blur)

Our method

- The blur prompting resembles the natural photography in web-scaled data trained by VLMs.
- A blurred background is similar to “Bokeh”, i.e., the aesthetic quality of the blur produced in out-of-focus parts of an image, caused by circles of confusion in photography.

Analysis

- **Biases on criminal categories**
  - A natural visual prompting reduces the classification biases towards criminal categories.

Original	Red Circle	Red Circle	**FGVP (blur)**	**FGVP (blur)**
This is an image of a **woman** 0.1353	This is an image of a **murderer** 0.1919	This is an image of a **murderer** 0.1899	This is an image of a **man** 0.2194	This is an image of a **woman** 0.2169
This is an image of a **man** 0.1323	This is an image of a **missing person** 0.1814	This is an image of a **woman** 0.1796	This is an image of a **missing person** 0.2109	This is an image of a **missing person** 0.2117
This is an image of a **murderer** 0.1254	This is an image of a **man** 0.1738	This is an image of a **missing person** 0.1776	This is an image of a **murderer** 0.2010	This is an image of a **murderer** 0.1942
This is an image of a **missing person** 0.1072	This is an image of a **woman** 0.1729	This is an image of a **man** 0.1605	This is an image of a **woman** 0.1820	This is an image of a **man** 0.1865

  - Quantitative experiment

Model	Visual Prompt	FairFace	COCO w/ crop	COCO w/o crop
ViT-L/14@336px	Crop	13.0	40.8	43.6
ViT-L/14@336px	RedCircle	20.6 (+7.6)	49.9 (+36.9)	69.3 (+56.3)
ViT-L/14@336px	FGVP	15.9 (+2.9)	34.1 (-6.7)	47.8 (+4.2)
ViT-B/32	Crop	14.5	27.2	34.9
ViT-B/32	RedCircle	22.0 (+7.5)	44.1 (+29.6)	68.6 (+54.1)
ViT-B/32	FGVP	8.2 (-6.3)	19.5 (-7.7)	15.8 (-19.1)

Visualization

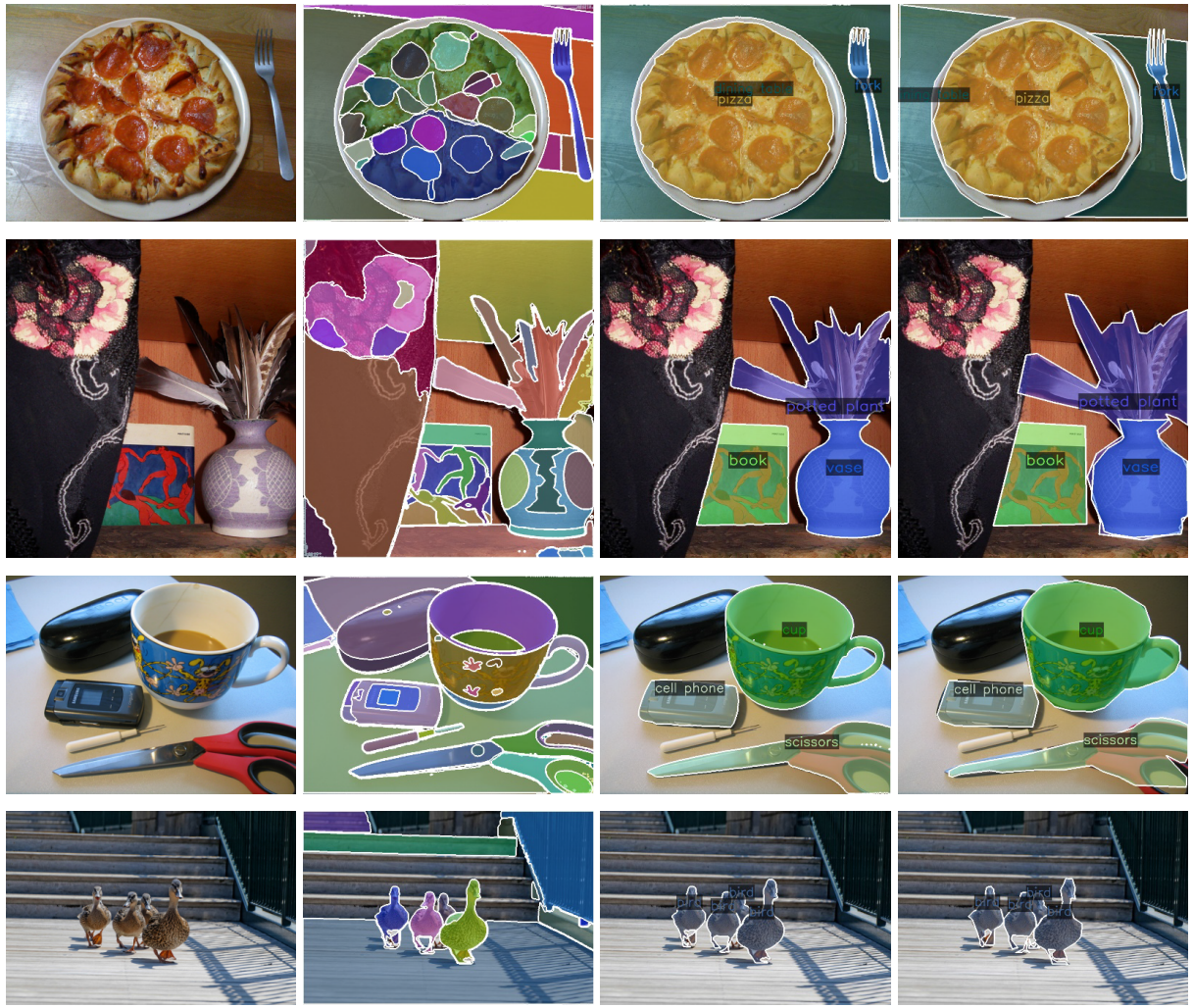


Image Proposals Prediction Ground-Truth

Image Proposals Prediction Ground-Truth

Visualization

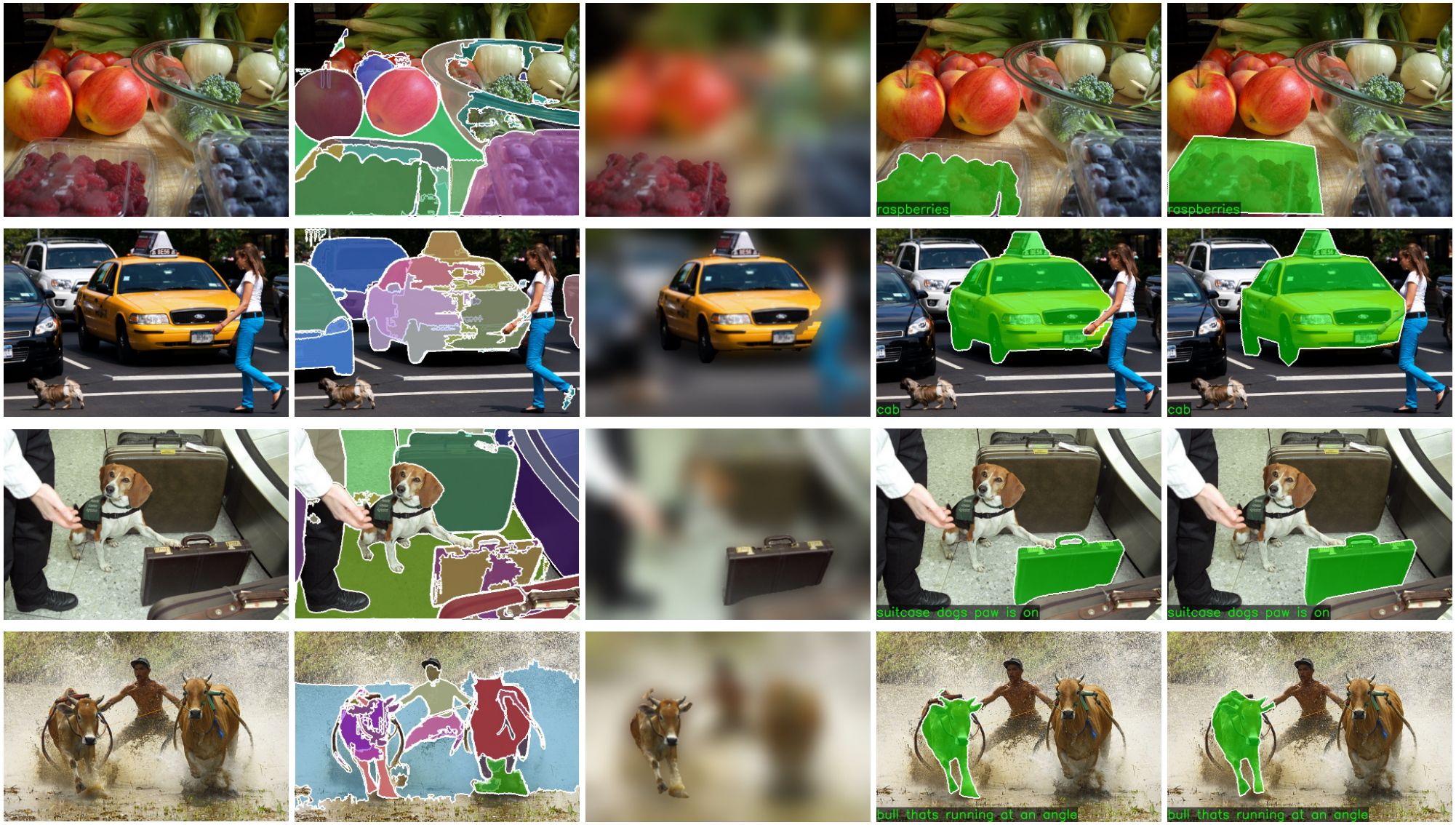


Image | Proposals | Prompting | Prediction | Ground-Truth

Efficiency

- **Available detector proposal**

  - Comparing inference costs in terms of computation and speed between our method and others.
  - Notably, the post-processing technique to filter small disconnected regions and holes in masks can further improve performance at the cost of speed. Disabling the mask-filter post-processing will greatly improve the speed without losing too much performance.
  - Experiments are run on RefCOCO with a CLIP pretrained ViT-L/14@336px on 8×NVIDIA A100.
  - Generally, FGVP takes more inference times than other methods.

Visual Prompt	SAM scale	Mask-filter	CUDA memory (GB)	Inference time (min)	Image per GPU second	Acc
Crop	--	--	0.91	4.49	5.03	45.3
RedCircle	--	--	0.91	4.00	5.64	48.9
FGVP	base	no	1.32	5.20	4.34	51.7
FGVP	base	yes	1.32	27.47	0.82	52.1
FGVP	large	no	2.14	6.29	3.59	51.0
FGVP	large	yes	2.14	27.49	0.82	52.2
FGVP	huge	no	3.42	7.34	3.08	51.9
FGVP	huge	yes	3.42	28.02	0.81	52.8

Efficiency

- **Grid keypoints as proposals**
  - We explore speed-performance trade-offs by varying grid sizes and NMS thresholds.
  - Experiments are run on PACO with a CLIP pretrained ViT-L/14@336px and SAM-huge on 8×NVIDIA A100.
  - FGVP could outperform RedCircle in speed and accuracy at grid size 8 and NMS threshold 0.95 trade-off.

Visual Prompt	Grid size	NMS threshold	Inference time (min)	Image per GPU second	Acc
Crop	16	0.7	13.34	3.27	16.5
Crop	32	0.95	37.25	1.17	19.5
RedCircle	16	0.7	12.75	3.42	17.4
RedCircle	32	0.95	34.18	1.28	19.9
FGVP	8	0.7	8.33	5.24	17.3
FGVP	8	0.95	9.17	4.76	20.5
FGVP	16	0.7	14.89	2.93	18.4
FGVP	16	0.95	17.29	2.52	22.0
FGVP	32	0.7	34.73	1.26	19.0
FGVP	32	0.95	39.66	1.10	23.2

Thanks

Lingfeng Yang

paper