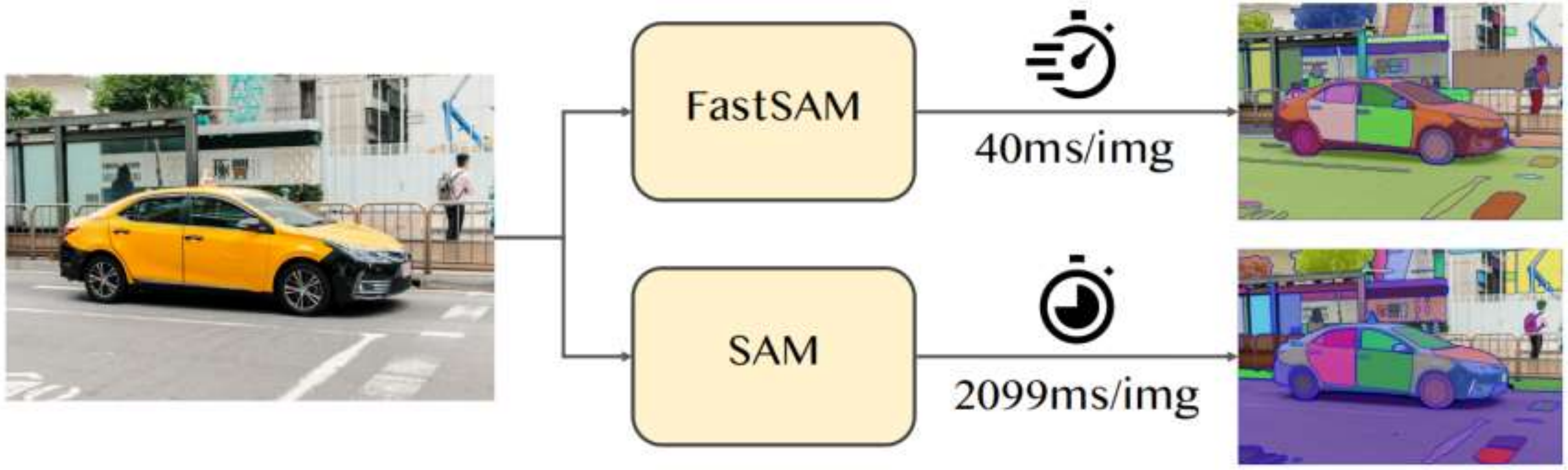
Fast Segment Anything

Xu Zhao [1,3] Wenchao Ding [1,2] Yongqi An [1,2] Yinglong Du [1,2]
Tao Yu [1,2] Min Li [1,2] Ming Tang [1,2] Jinqiao Wang [1,2,3,4]
Institute of Automation, Chinese Academy of Sciences, Beijing, China[1]
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China[2]
Objecteye Inc., Beijing, China[3]
Wuhan AI Research, Wuhan, China[4]

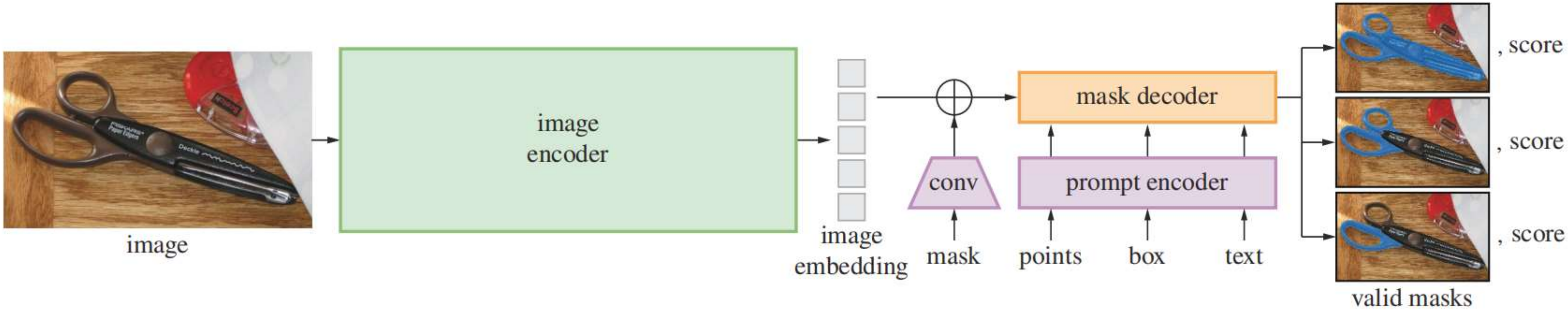


Motivation

The substantial computational resource requirements associated with ViT models make SAM's practical applications are still challenging

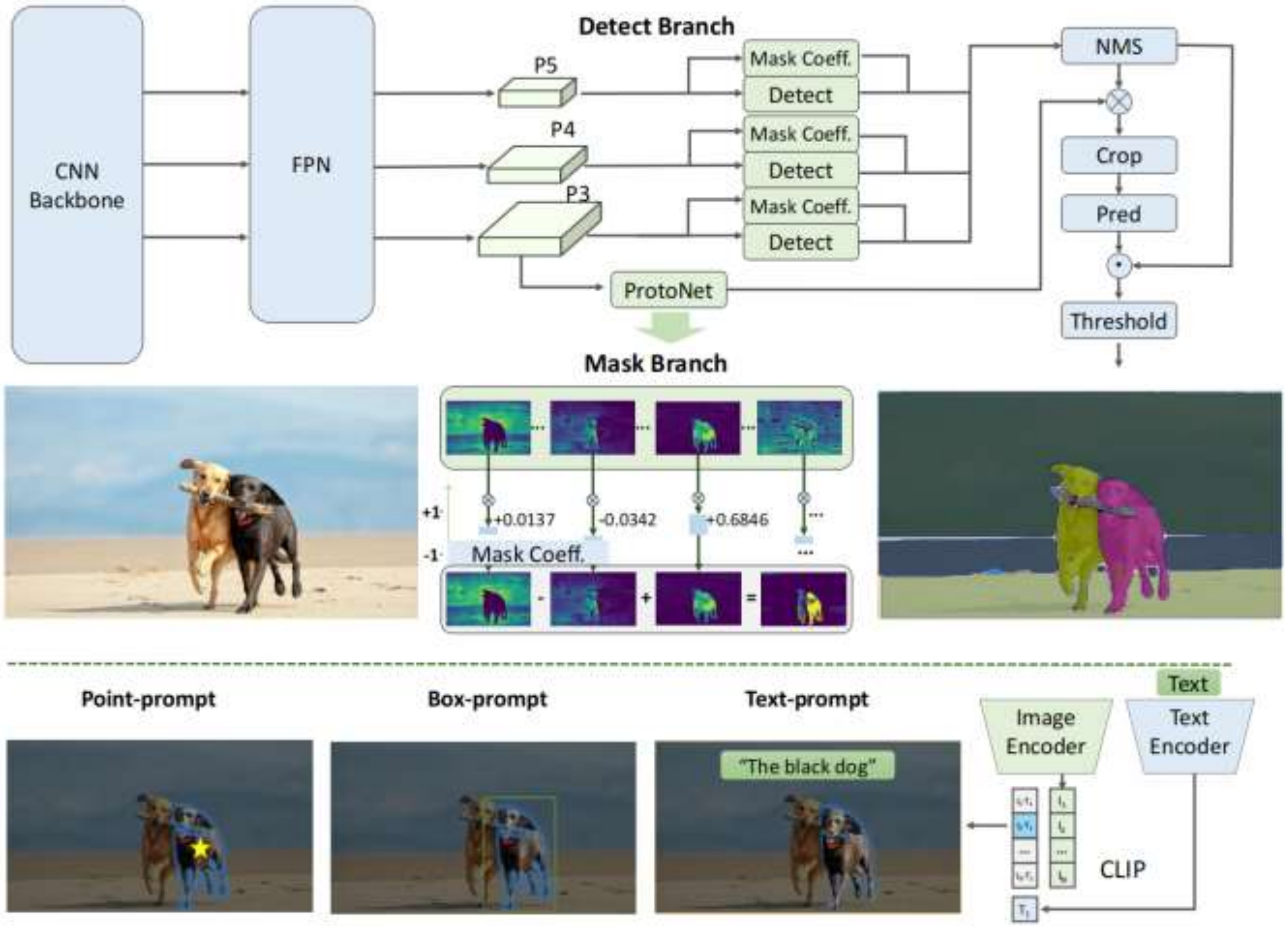
RTX 3090		Running Speed under Different Point Prompt Numbers (ms)					
method	params	1	10	100	E(16×16)	E(32×32*)	E(64×64)
SAM-H [20]	0.6G	446	464	627	852	2099	6972
SAM-B [20]	136M	110	125	230	432	1383	5417
		2 fps			< 1 fps	0.5 fps	

Pipeline of SAM



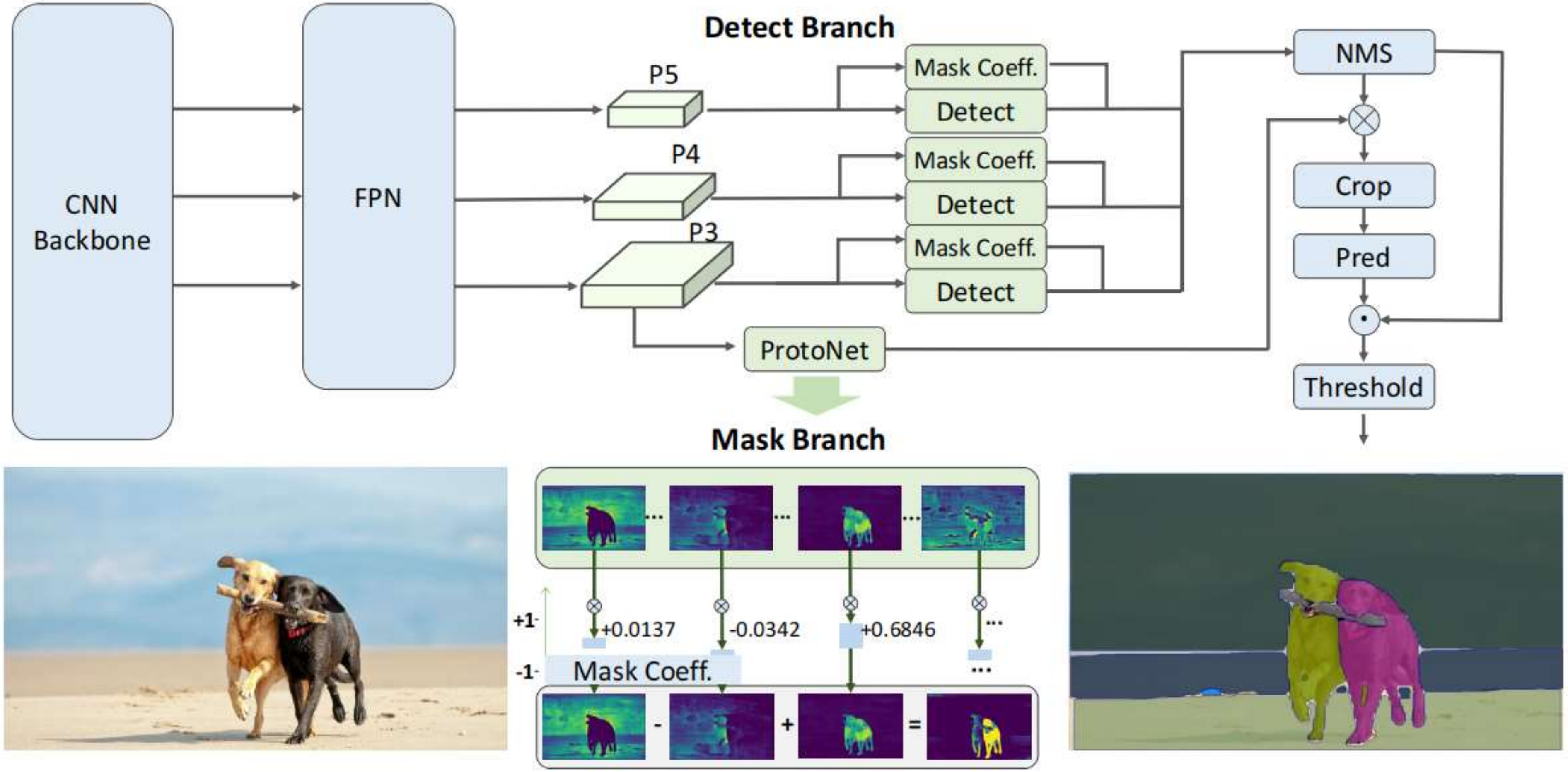
Provide the prompt + Segment the target

Pipeline of Fast-SAM



All-instance segmentation + Prompt-guided selection

YOLOv8-seg

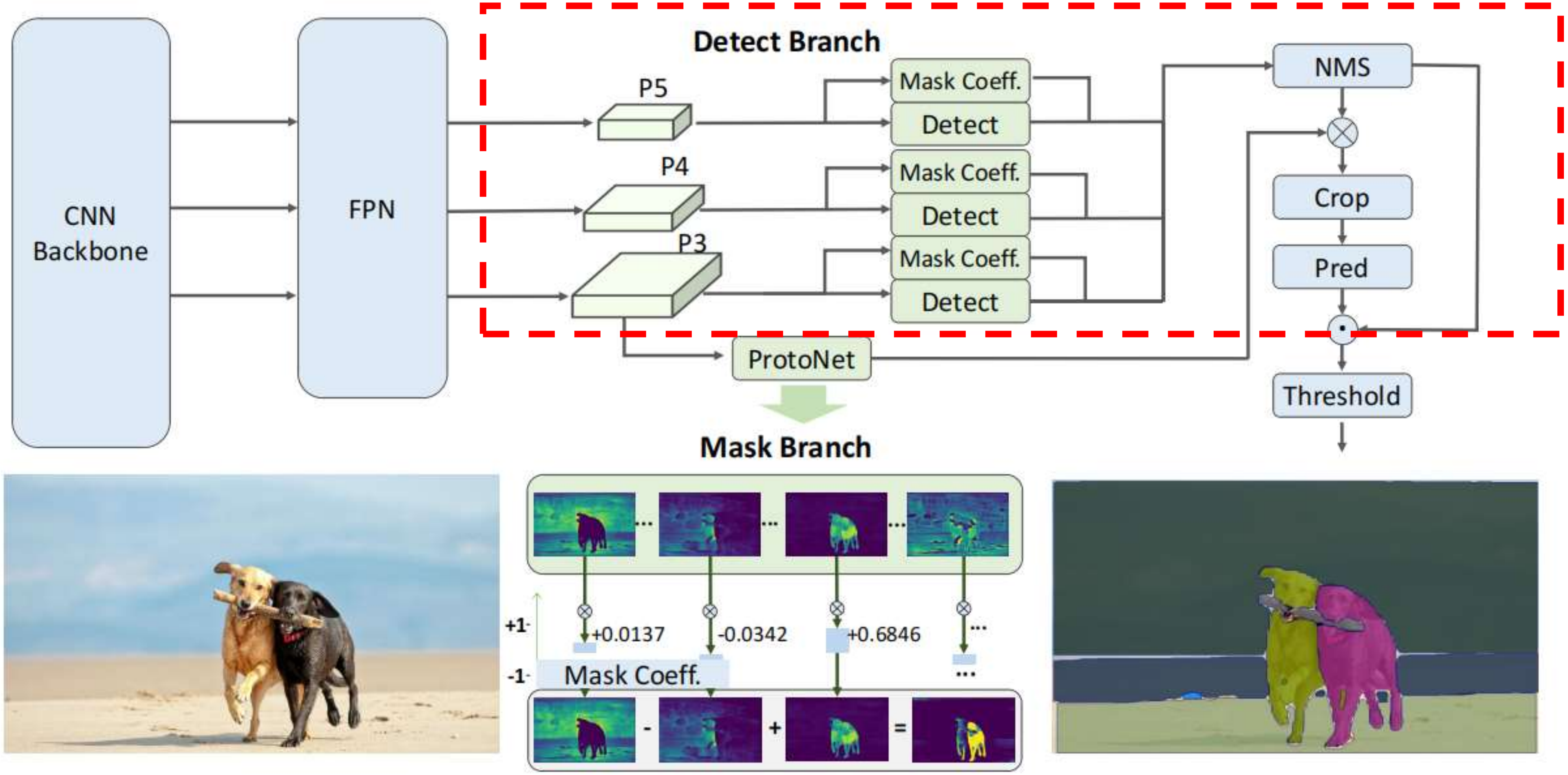


Detect Branch
CNN Backbone
FPN
P5
P4
P3
Mask Coeff.
Detect
Mask Coeff.
Detect
Mask Coeff.
Detect
ProtoNet
NMS
Crop
Pred
Threshold
Mask Branch
+1
-1
+0.0137
-0.0342
+0.6846
Mask Coeff.

Detect-Branch

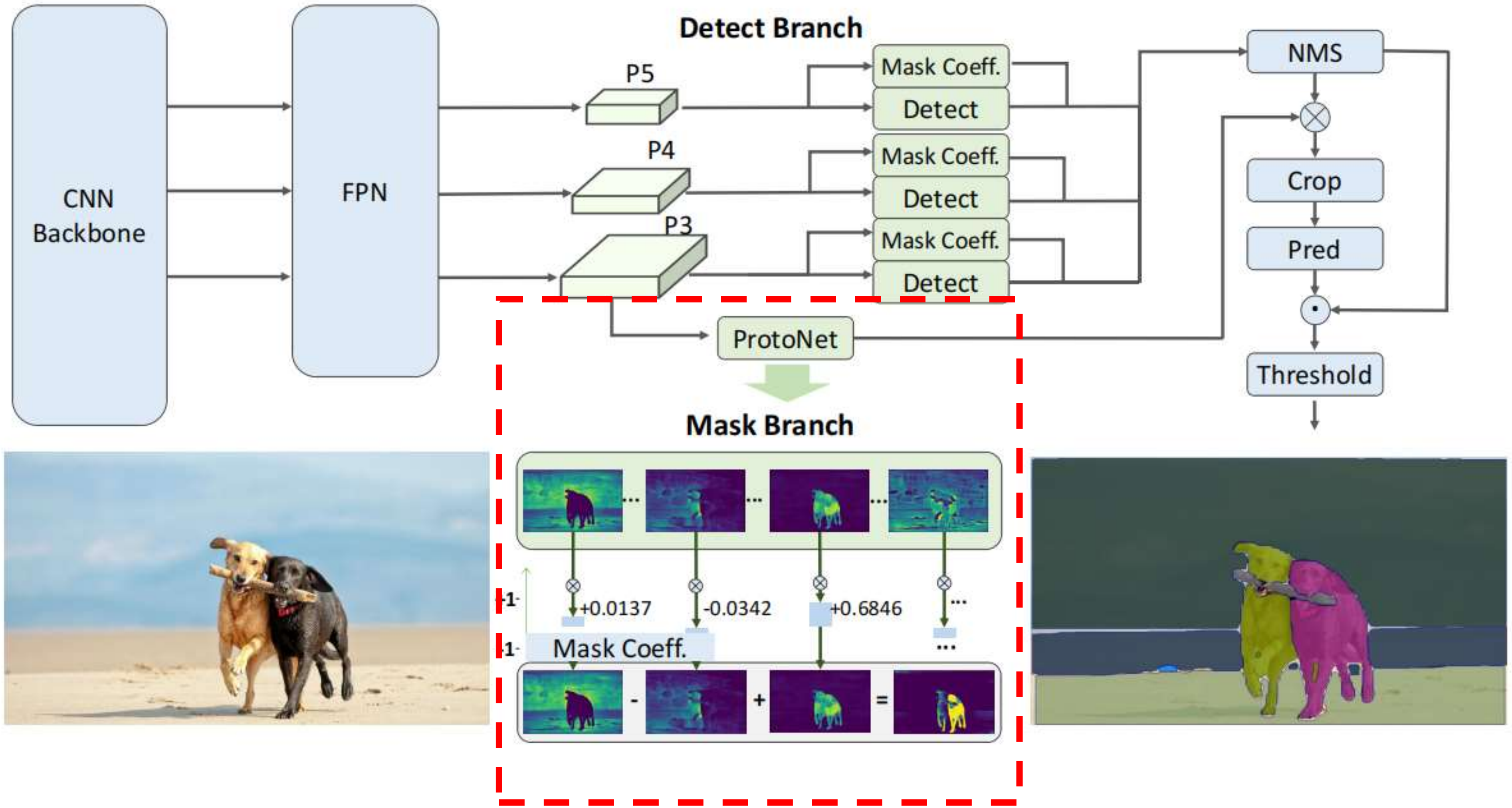
Detect ➞ $W \times H \times [\,x, y, h, w, cls\,]$

Mask Coeff. ➞ $W \times H \times 32$



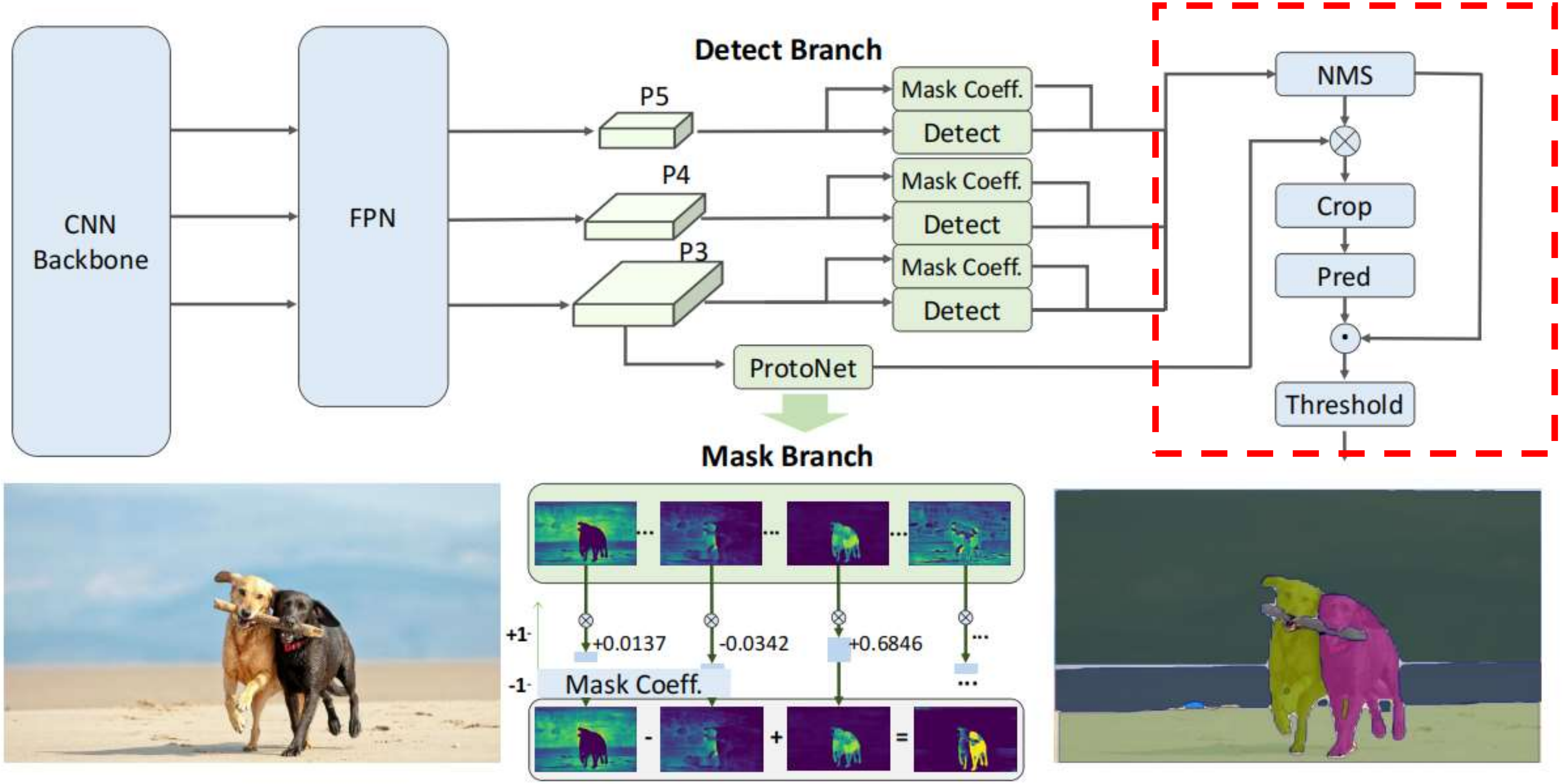
Detect-Branch

[W, H, 320] ➞ [2W , 2H, 32]



Detect-Branch

$$[N, 32] \; @ \; [32, W, H] = [N, W, H]$$



Prompt-guided Selection

merging

Compute IoU

Compute similarity

Experiments -using only 2% of the SA-1B

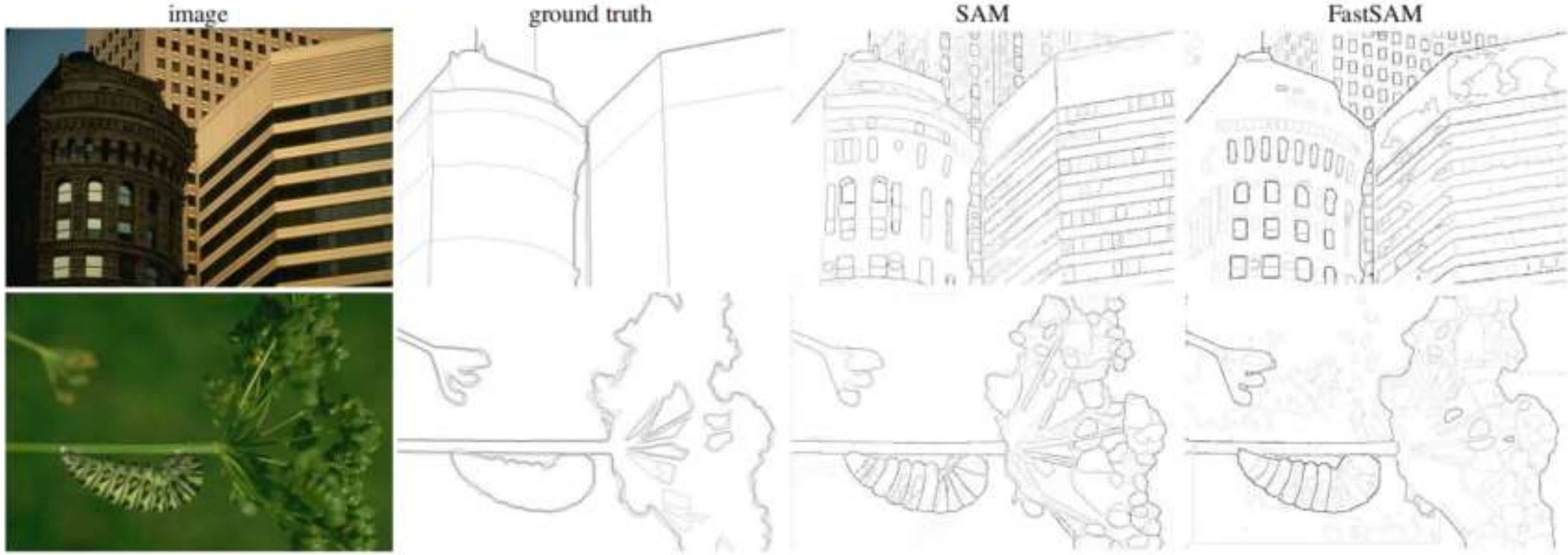
- **Run-time Efficiency**
- **low-level: edge detection**
- **mid-level: object proposal generation**
- **high-level: instance segmentation**
- **high-level: segmenting objects with free-form text input**
- **Real-world Applications**

Run-time Efficiency

method	params	Running Speed under Different Point Prompt Numbers (ms)					
		1	10	100	E(16×16)	E(32×32*)	E(64×64)
SAM-H [20]	0.6G	446	464	627	852	2099	6972
SAM-B [20]	136M	110	125	230	432	1383	5417
FastSAM (Ours)	68M	40					

50x faster than SAM (32×32)
170x faster than SAM (64×64)

Zero-Shot Edge Detection

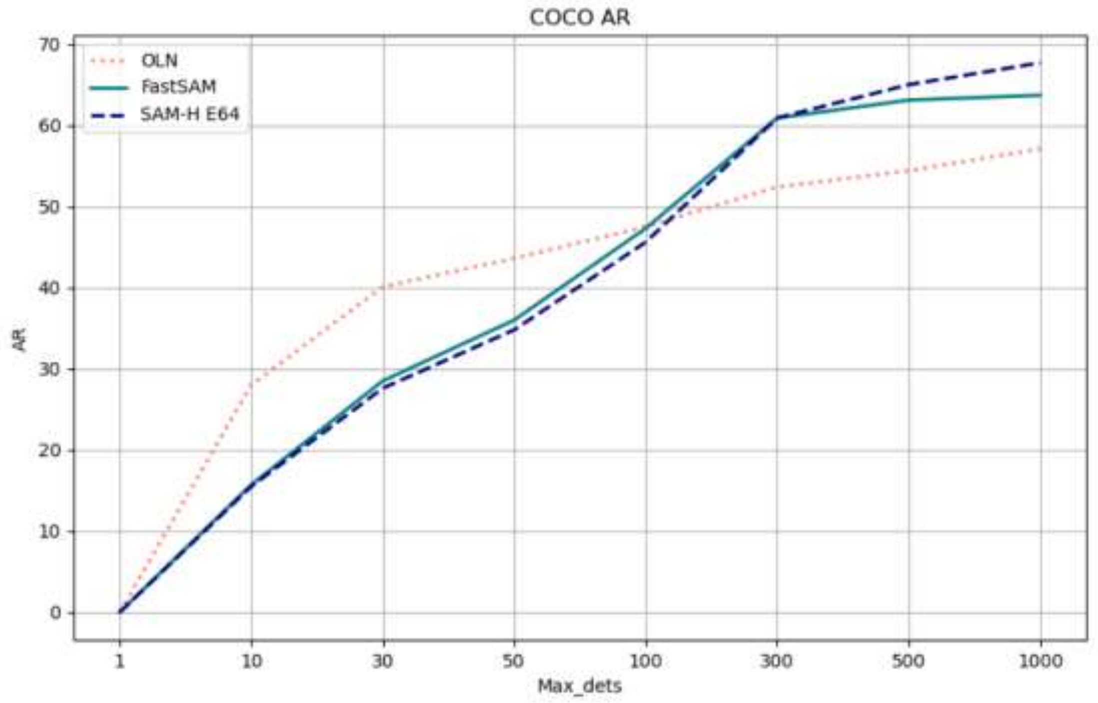


method	year	ODS	OIS	AP	R50
HED [37]	2015	.788	.808	.840	.923
EDETR [30]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [6]	1986	.600	.640	.580	-
Felz-Hutt [9]	2004	.610	.640	.560	-
SAM [19]	2023	.768	.786	.794	.928
FastSAM	2023	.750	.790	.793	.903

object proposal generation

	AR10	AR100	AR1000	AUC
EdgeBoxes [38]	7.4	17.8	33.8	13.9
Geodesic [21]	4.0	18.0	35.9	12.6
Sel.Search [34]	5.2	16.3	35.7	12.6
MCG [2]	10.1	24.6	39.8	18.0
DeepMask [29]	13.9	28.6	43.1	21.7
OLN-Box [17]	27.7	46.1	55.7	34.3
SAM-H E64	15.5	45.6	67.7	32.1
SAM-H E32	18.5	49.5	62.5	33.7
SAM-B E32	11.4	39.6	59.1	27.3
FastSAM (Ours)	15.7	47.3	63.7	32.2

mask proposal generation



method	mask AR@1000						
	all	small	med.	large	freq.	com.	rare
results reported in the SAM paper:							
ViTDet-H [23]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
SAM [20] – single out.	54.9	42.8	76.7	74.4	54.7	59.8	62.0
SAM [20]	59.3	45.5	81.6	86.9	59.1	63.9	65.8
results after our replication:							
ViTDet-H [23]	59.9	48.3	78.1	84.8	-	-	-
SAM-H E64	54.2	39.6	77.9	83.9			
SAM-H E32	51.8	35.2	78.7	85.2	-	-	-
SAM-B E32	45.8	31.1	70.5	73.6	-	-	-
FastSAM (Ours)	49.7	35.6	72.7	77.6	-	-	-

We think this is because the confidence score is defined as the b-box score of YOLOv8, which is not strongly related to the mask quality

Zero-Shot Instance Segmentation

	COCO [26]				LVIS v1 [13]			
method	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L
ViTDet-H [23]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5
FastSAM	37.9	23.9	43.4	50.0	34.5	24.6	46.2	50.8

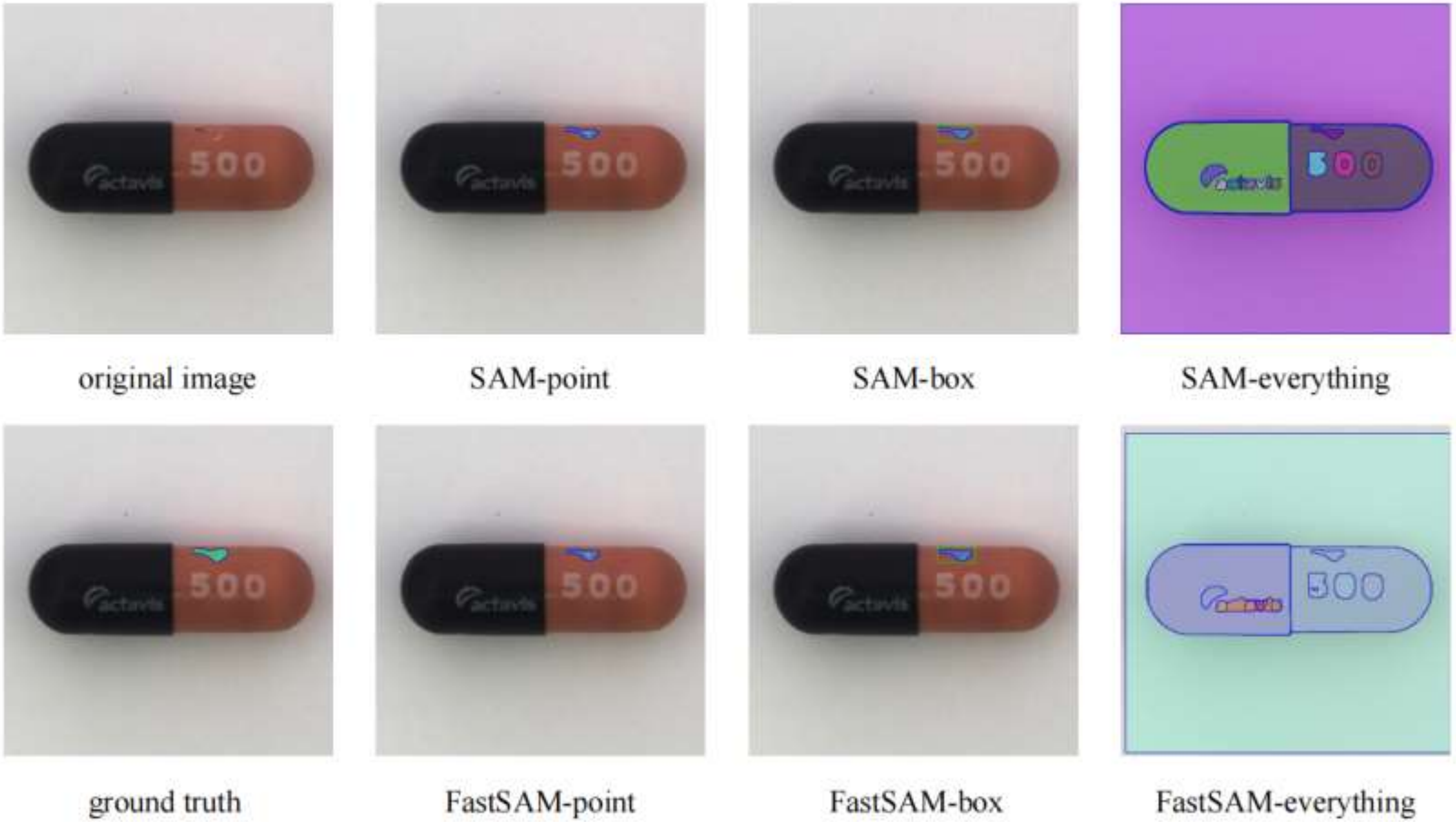
The masks of some of the tiny-sized objects tend to be near the square. Besides, the mask of large objects may have some artifacts on the border of the bounding boxes

Zero-Shot Object Localization with Text Prompts

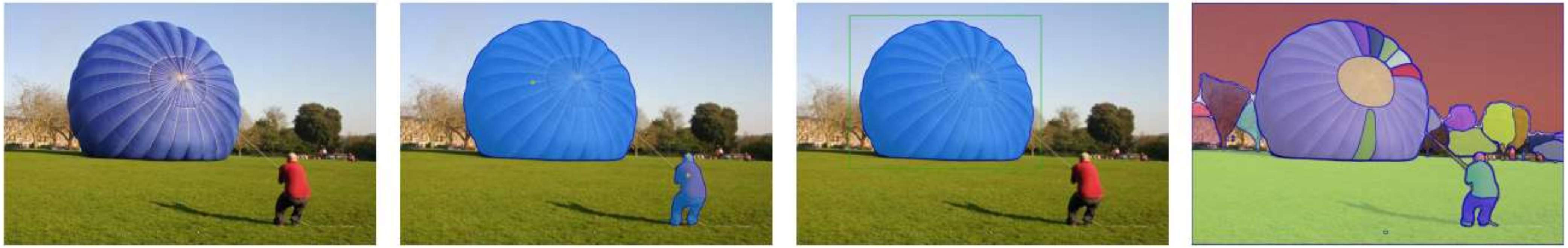
Image Text prompt: “The yellow dog” Text prompt: “The black dog”

the running speed of the text-to-mask segmentation is not satisfying, since each mask region is required to be fed into the CLIP feature extractor

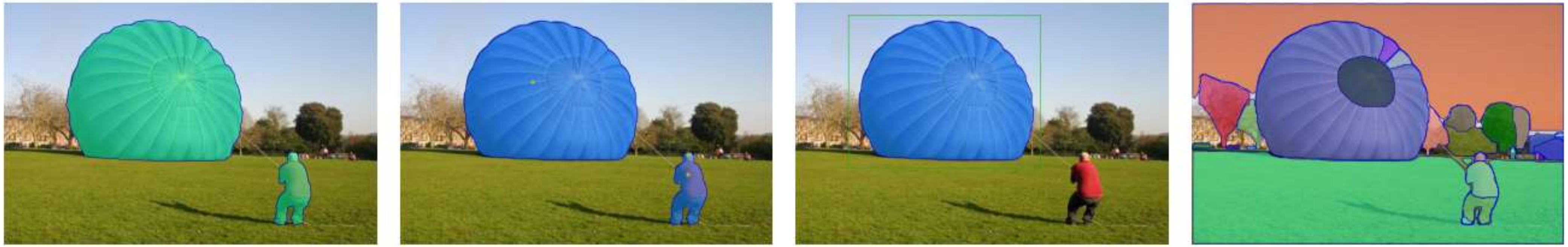
Anomaly Detection



Salient Object Segmentation

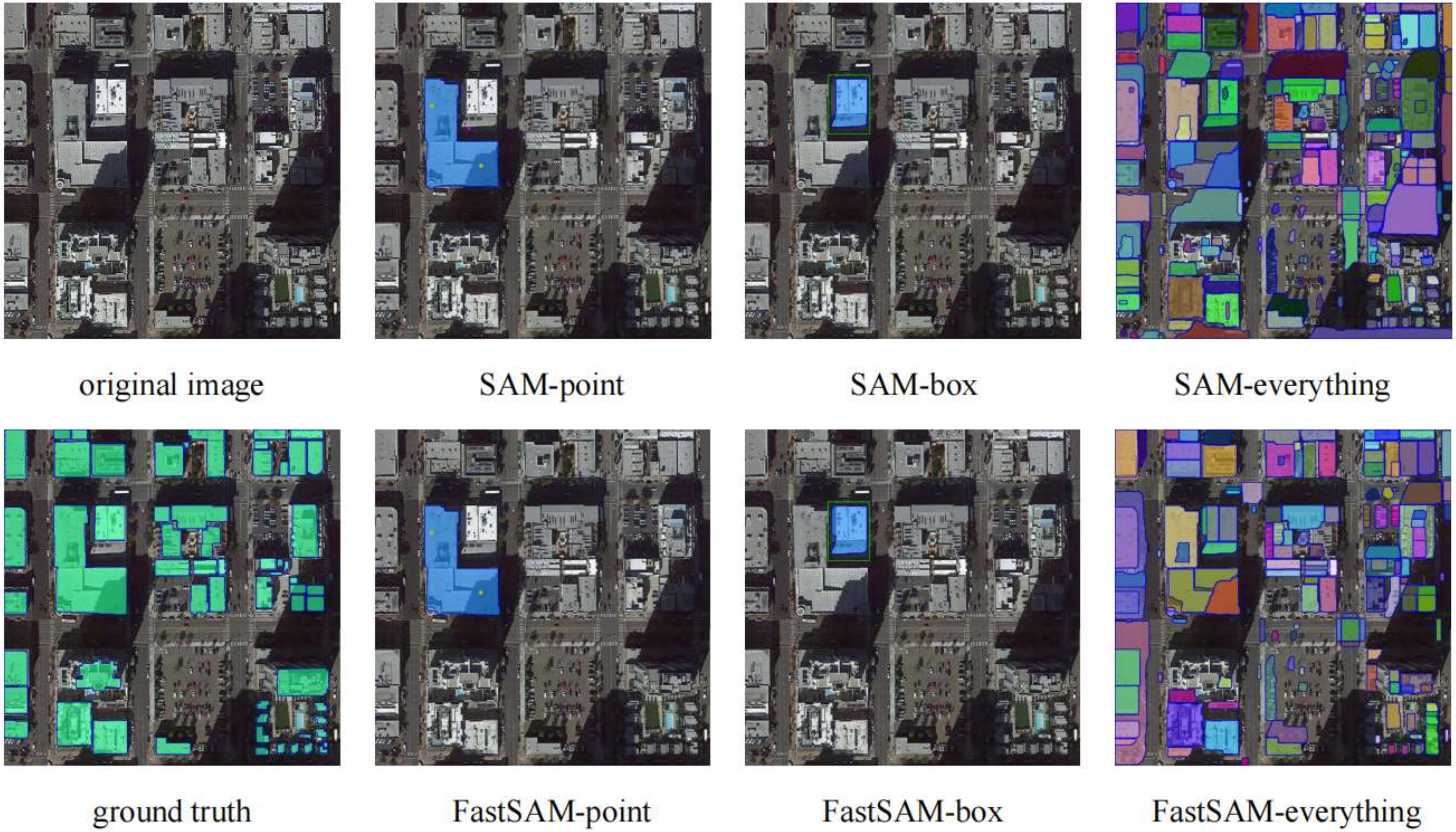


original image | SAM-point | SAM-box | SAM-everything



ground truth | FastSAM-point | FastSAM-box | FastSAM-everything

Building Extracting



Motivation

- The substantial computational resource requirements associated with ViT models make SAM’s practical applications are still challenging
- **Break the limitations of the ViT model provided by SAM**

Limitations of SAM ViT-Huge

- Heavy computation
- Promptable model
- Portability

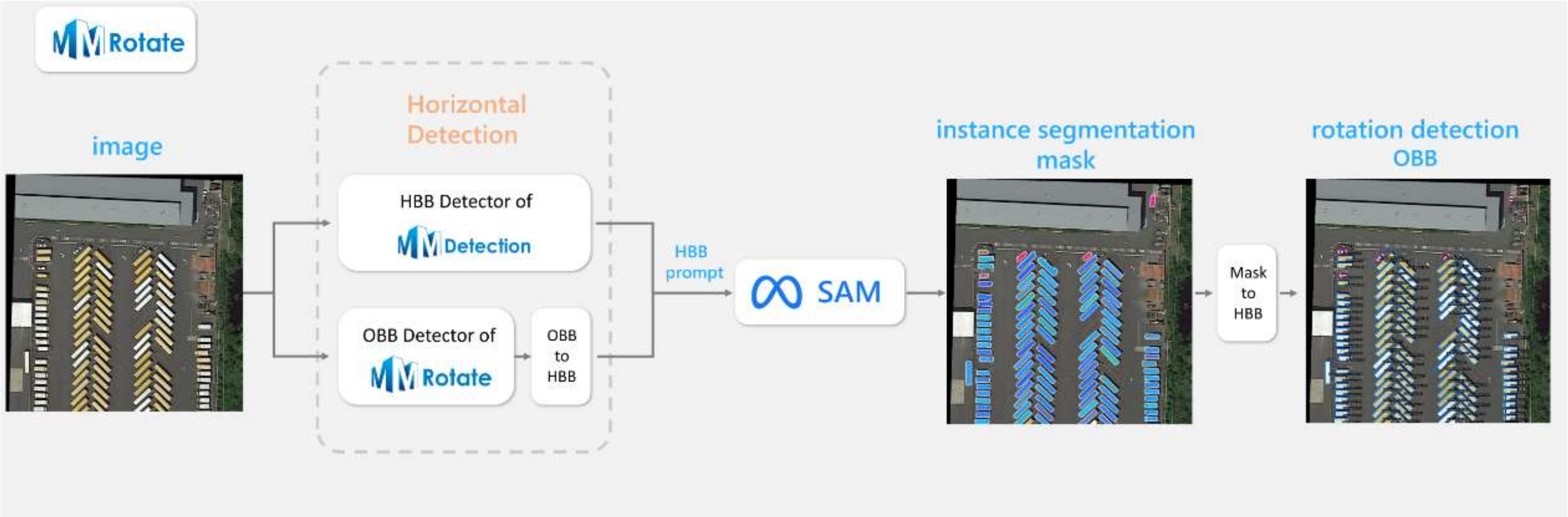
Promptable model

- SAM segments nothing without prompt

- A model that generates prompt + ViT-Huge

- Find a suitable prompt for all downstream tasks

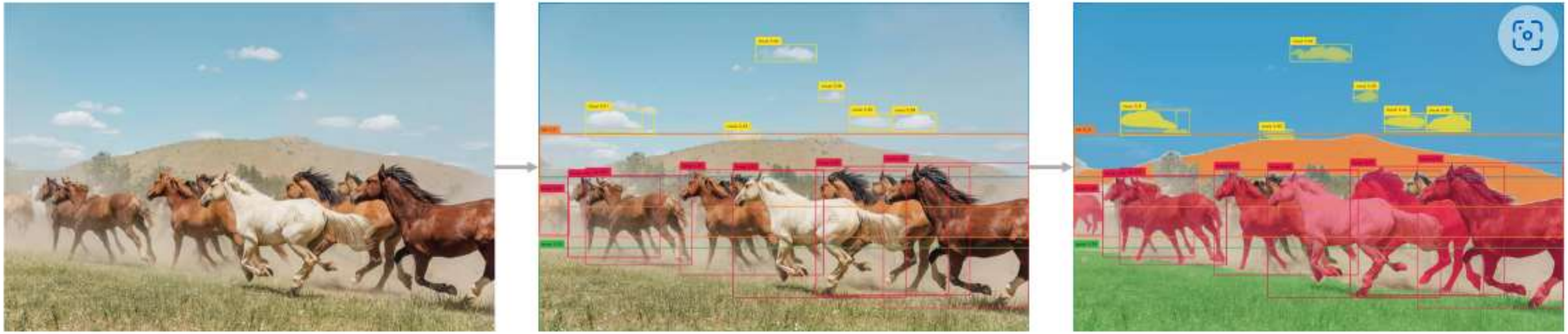
SAM-RBox



MMRotate
image
Horizontal
Detection
HBB Detector of
MMDetection
OBB Detector of
MMRotate
OBB
to
HBB
HBB
prompt
SAM
instance segmentation
mask
Mask
to
HBB
rotation detection
OBB

Grounded-Segment-Anything

combining Grounding DINO and Segment Anything

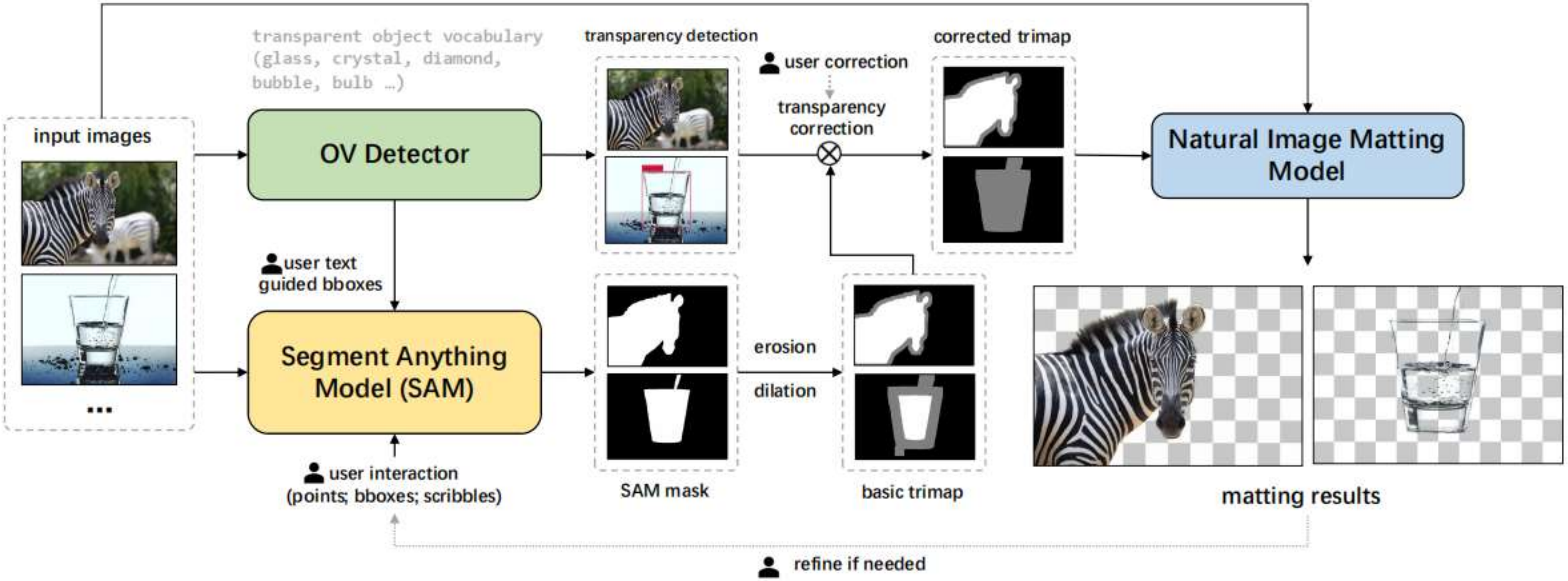


Text Prompt:
"Horse. Clouds. Grasses. Sky. Hill."

Grounding DINO:
Detect Everything

Grounded-SAM:
Detect and Segment Everything

Matte Anything



What does SAM bring us？

- **A ViT-Huge model**
- **SA1B dataset of 1 Billion masks**
- **A data engine**

A Survey on Segment Anything Model (SAM): Vision Foundation Model Meets Prompt Engineering

Chaoning Zhang* — Kyung Hee University
Fachrina Dewi Puspitasari — KAIST
Sheng Zheng — Beijing Institute of Technology
Chenghao Li — KAIST
Yu Qiao — Kyung Hee University
Taegoo Kang — Kyung Hee University
Xinru Shan — Microsoft STCA
Chenshuang Zhang — KAIST
Caiyan Qin — Harbin Institute of Technology
Francois Rameau — State University of New York at Korea
Lik-Hang Lee — Hong Kong Polytechnic University
Sung-Ho Bae — Kyung Hee University
Choong Seon Hong — Kyung Hee University

Based on the task of promptable segmentation, the segment anything model **(SAM) is the first vision foundation model that mimics the human eye to understand the world and its emergence has transformed the computer vision community**.

Our work conducts the first yet comprehensive survey on SAM. We hope our survey helps readers interested in SAM for performing their research.