

# 복부 CT(Computed Tomography) 이미지 객체 분할 작업 성능 개선에 관한 연구

---

한국통신학회 하계종합학술발표회2024

Computer Software

장기태

1st

**What is SAM**

2nd

**Primary  
Contribution**

3rd

**Sam experiment**

4th

**HQ-SAM**

5th

**Our proposed  
Method**

6th

**Experiment**

# 1. What is SAM



1. GPT trained on Next token prediction performs well on various tasks.
2. Develop a new task, model, and dataset in Computer Vision with the aim of creating such a versatile model.
3. Segmentation is naturally well-executed, and it shows decent performance on other tasks as well.



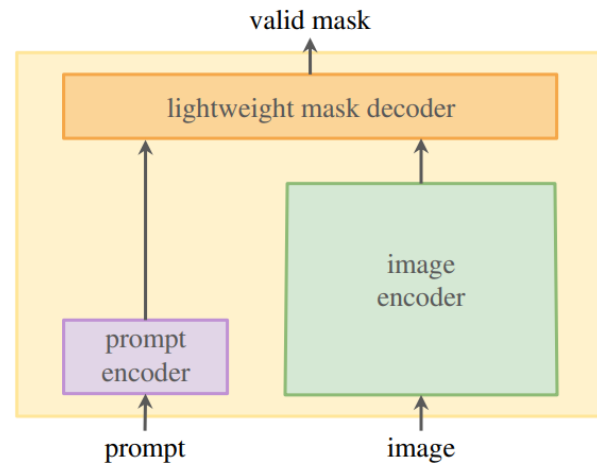
<https://segment-anything.com/demo>

# 1. What is SAM

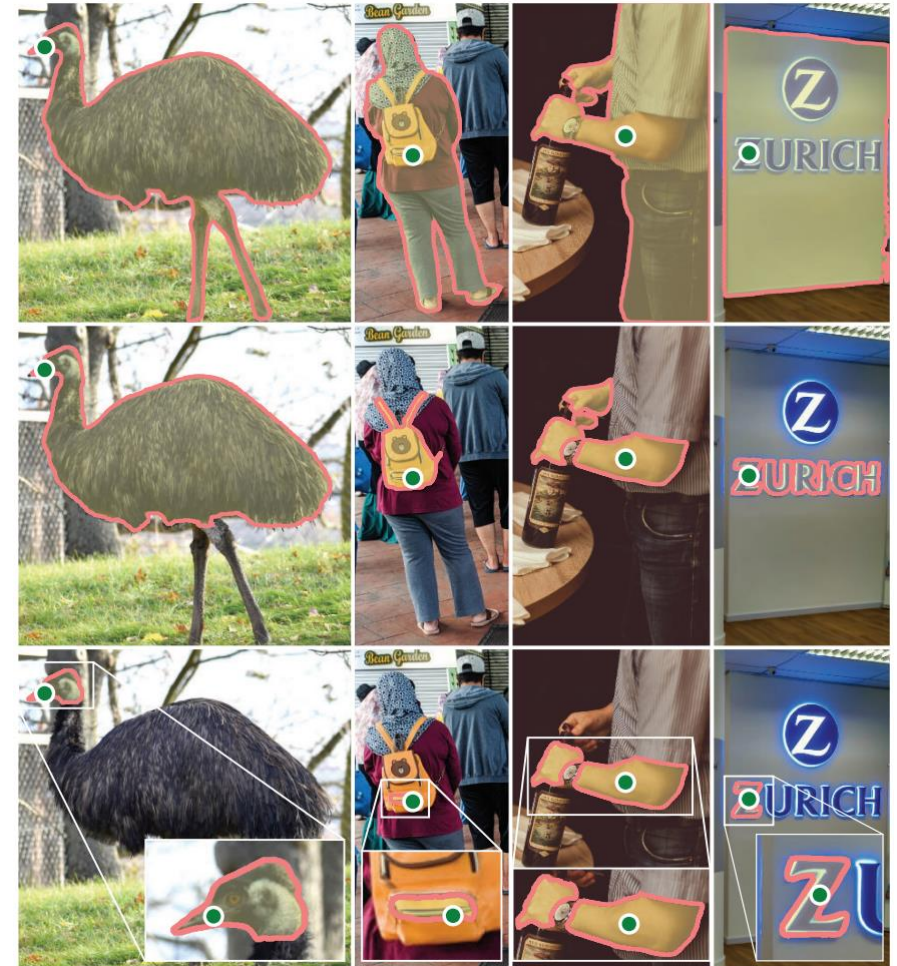
## Segment Anything Model

Refers to a deep learning model that identifies and segments objects in visual data such as images or videos

SAM is primarily used for image segmentation tasks, accurately delineating the boundaries of objects within an image to segment them.



(b) **Model:** Segment Anything Model (SAM)



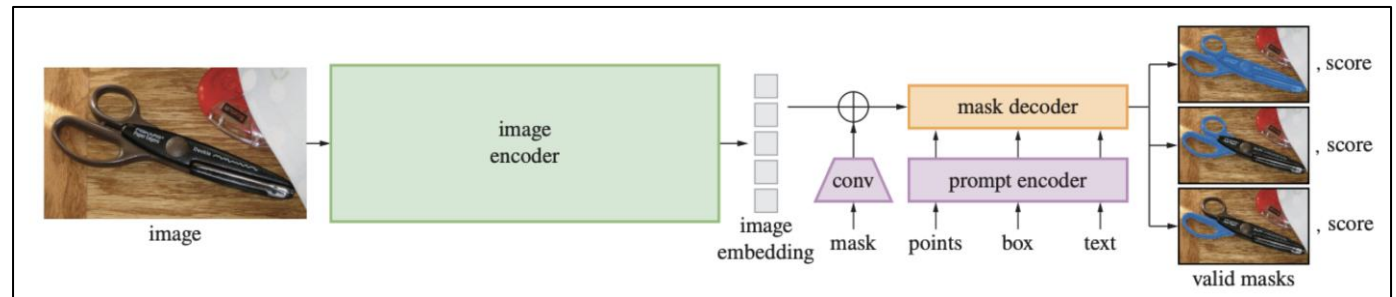
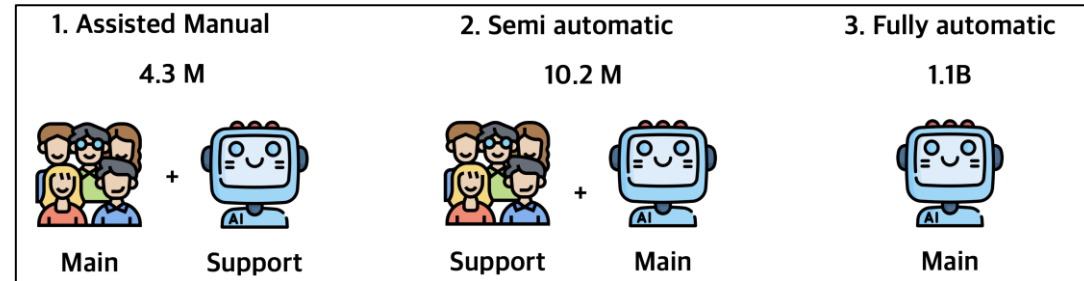
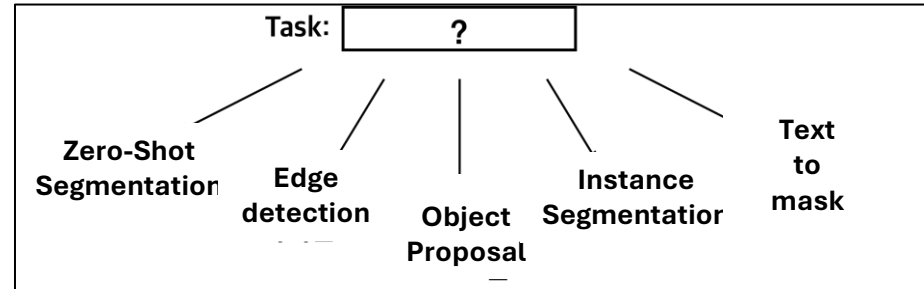
# 2. Primary Contribution

**SAM**

Task

Data

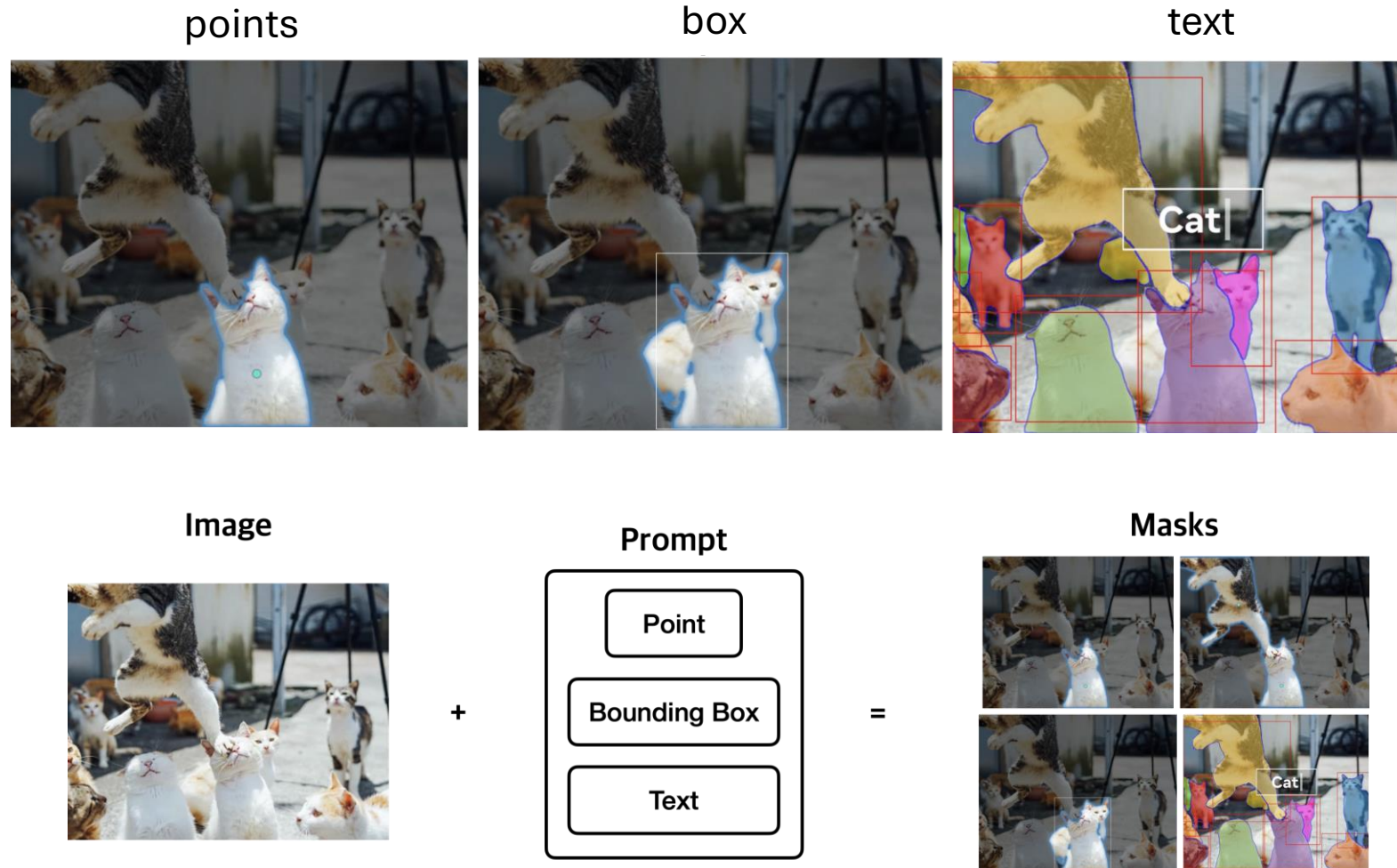
Model





## 2. Primary Contribution(Task)

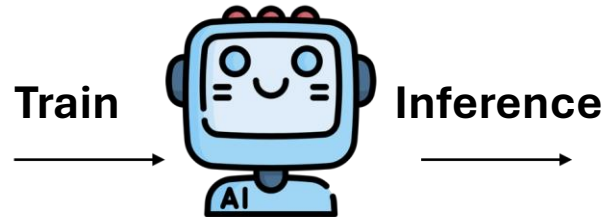
### Promptable Segmentation



## 2. Primary Contribution(Data)

### 1. Assisted Manual

Dataset
LVIS v1
MS COCO
ADE20
Open Images

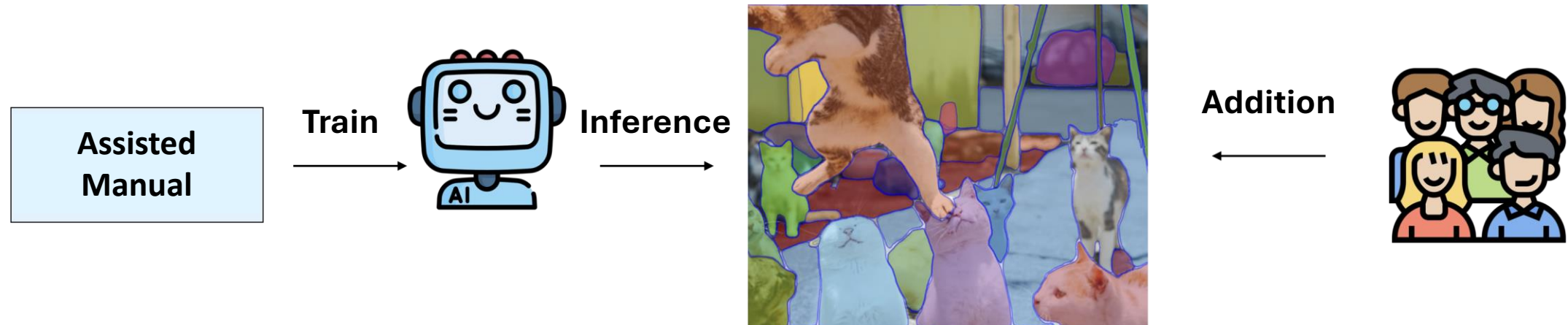


Modification  
Addition



# 2. Primary Contribution(Data)

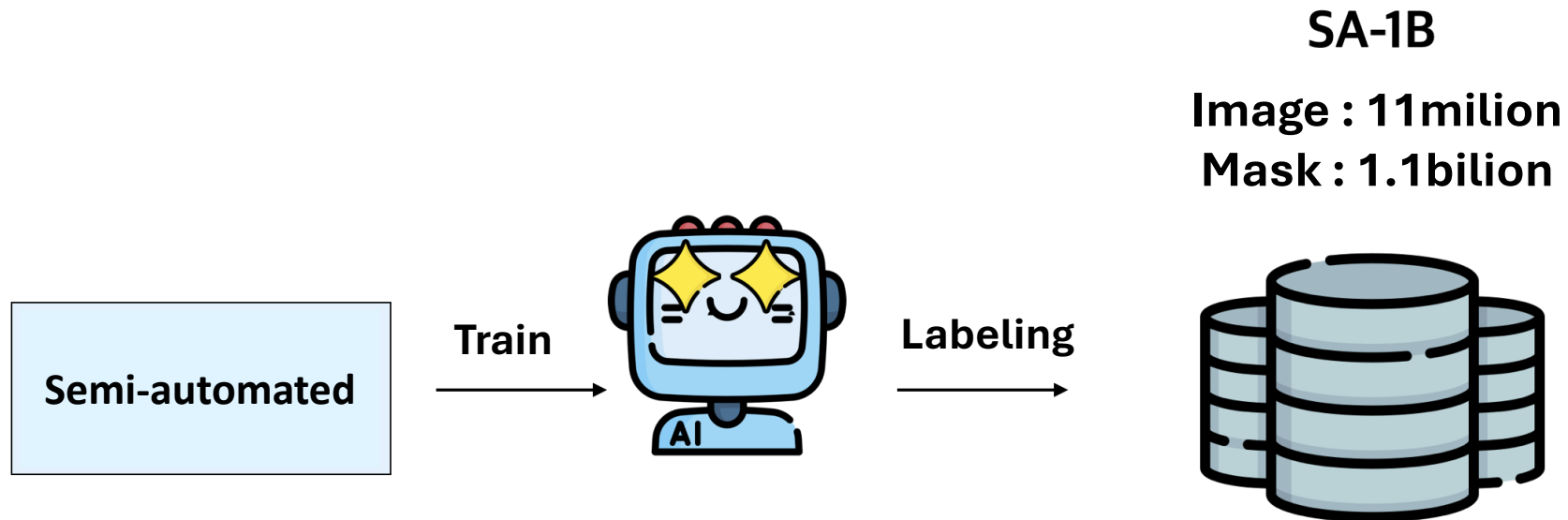
## 2. Semi-automated





## 2. Primary Contribution(Data)

### 3. Fully-automated



# 2. Primary Contribution(Model)

## Model Architecture

### 1. Image Encoder

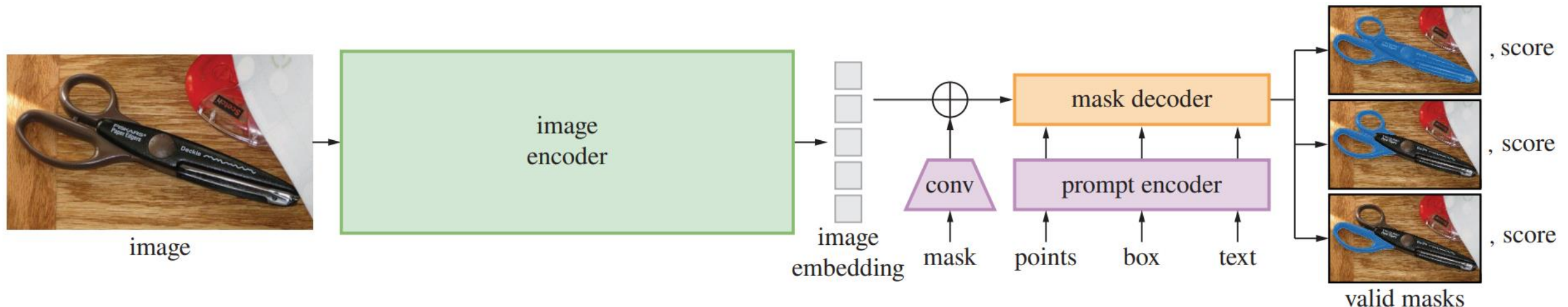
Image encoder extracts features from the given image.  
It is typically based on Convolutional Neural Network

### 2. Prompt Encoder

Prompt encoder converts textual prompts into vector representations.

### 3. Mask Decoder

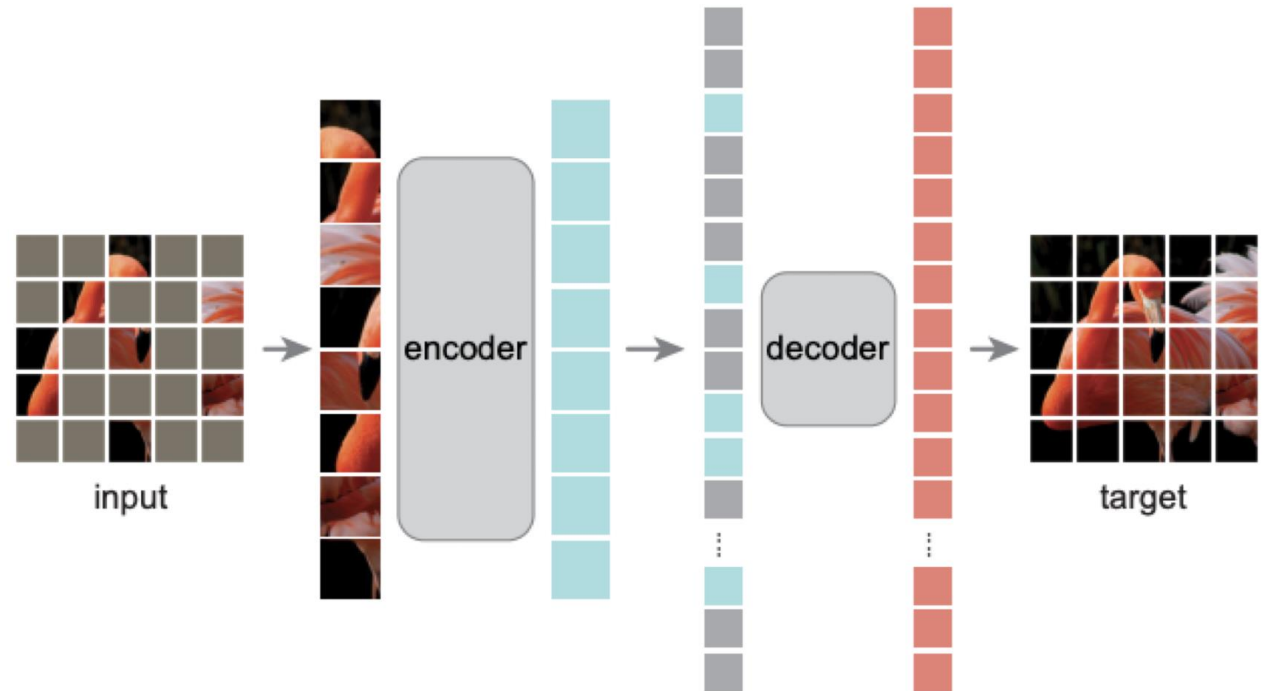
Mask decoder rapidly segments objects based on images and textual prompts.  
This decoder generates masks used for object segmentation by combining image and text information.



## 2. Primary Contribution(Model)

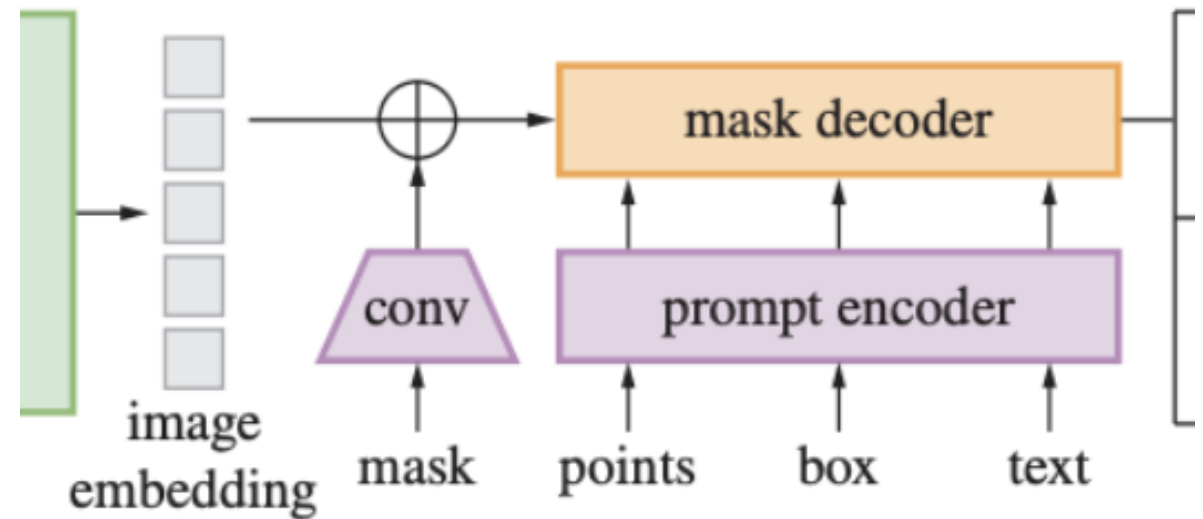
### 1. Image Encoder

- The Image Encoder utilizes a Vision transformer trained with Masked Auto-Encoder (MAE) approach.
- MAE divides the image into a grid of fixed size, randomly masks parts of it, and trains the model to reconstruct.
- Only the encoder is used in the model, excluding the decoder.



## 2. Prompt Encoder

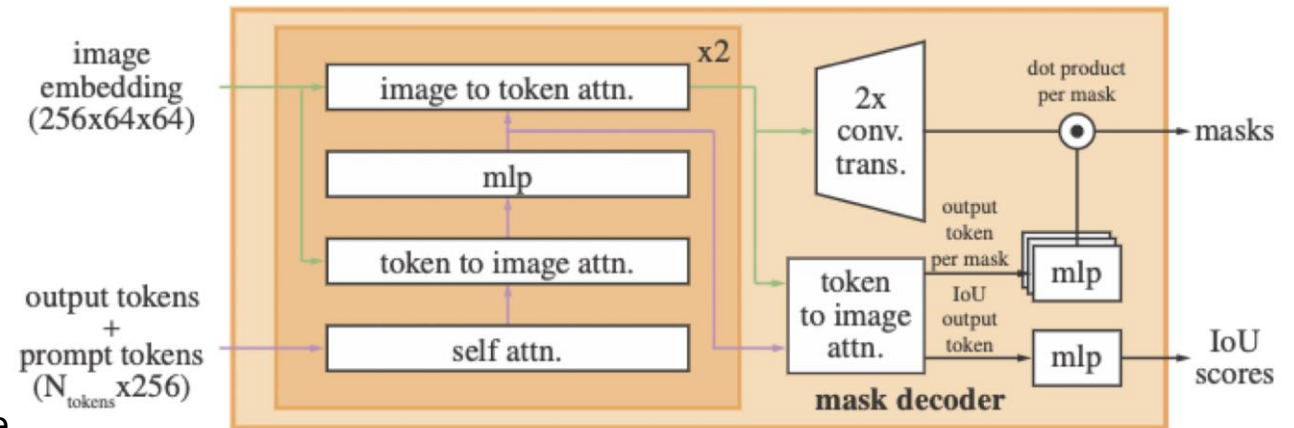
- **Sparse prompts** (points, boxes, text): Embedded into a 256-dimensional vector.
  - Points : The position of the dot is encoded using positional encoding + foreground/background embedding.
  - Boxex : positional encoding + top-left corner / bottom-right corner
  - Text : text encoder from CLIP
- **Dense prompts** (masks): convolutions and summed element-wise with the image embedding



## 2. Primary Contribution(Model)

### 3. Mask Decoder

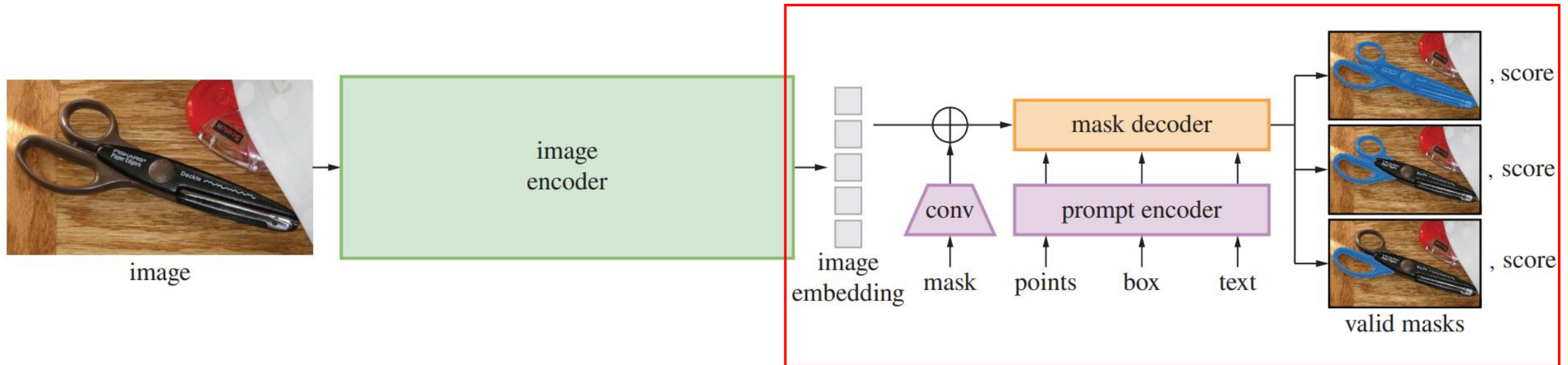
- 1) Self-attention on the tokens
- 2) Cross-attention from tokens to the image embedding  
**(query : token / key, value : image)**
- 3) Point-wise MLP updates each token  
**A linear layer is applied to channels per pixel.**
- 4) Cross-attention from the image embedding to token  
**(query : image / key, value : token)**
- 5) Token to image cross attention -> Output token
- 6) The image embedding, passed through attention, is upsampled using convtranspose to increase the spatial size by 4 size.



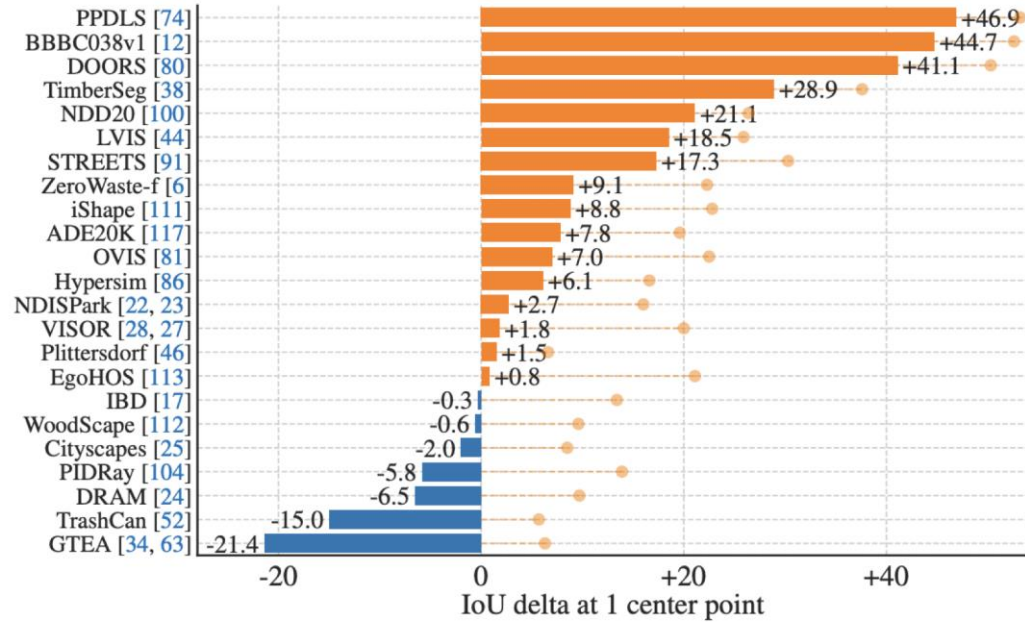


## 2. Primary Contribution(Model)

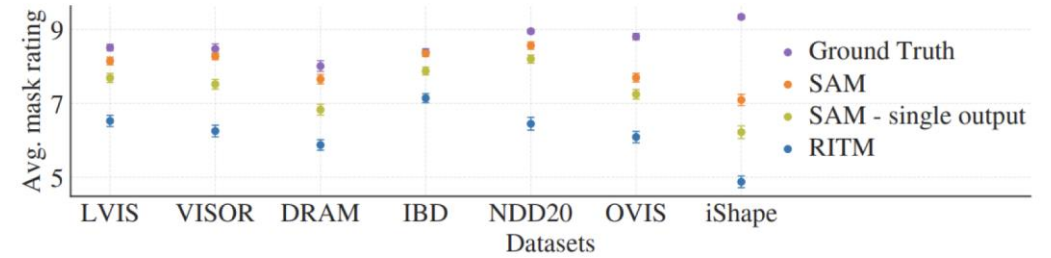
### 3. Mask Decoder



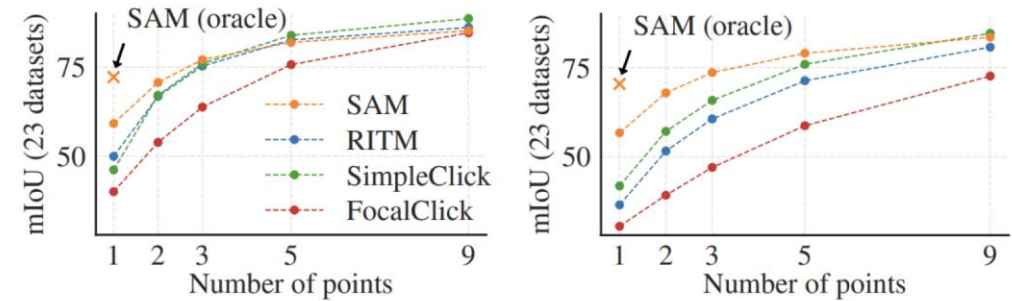
# 3. Sam experiment



(a) SAM vs. RITM [92] on 23 datasets



(b) Mask quality ratings by human annotators



(c) Center points (default)

(d) Random points

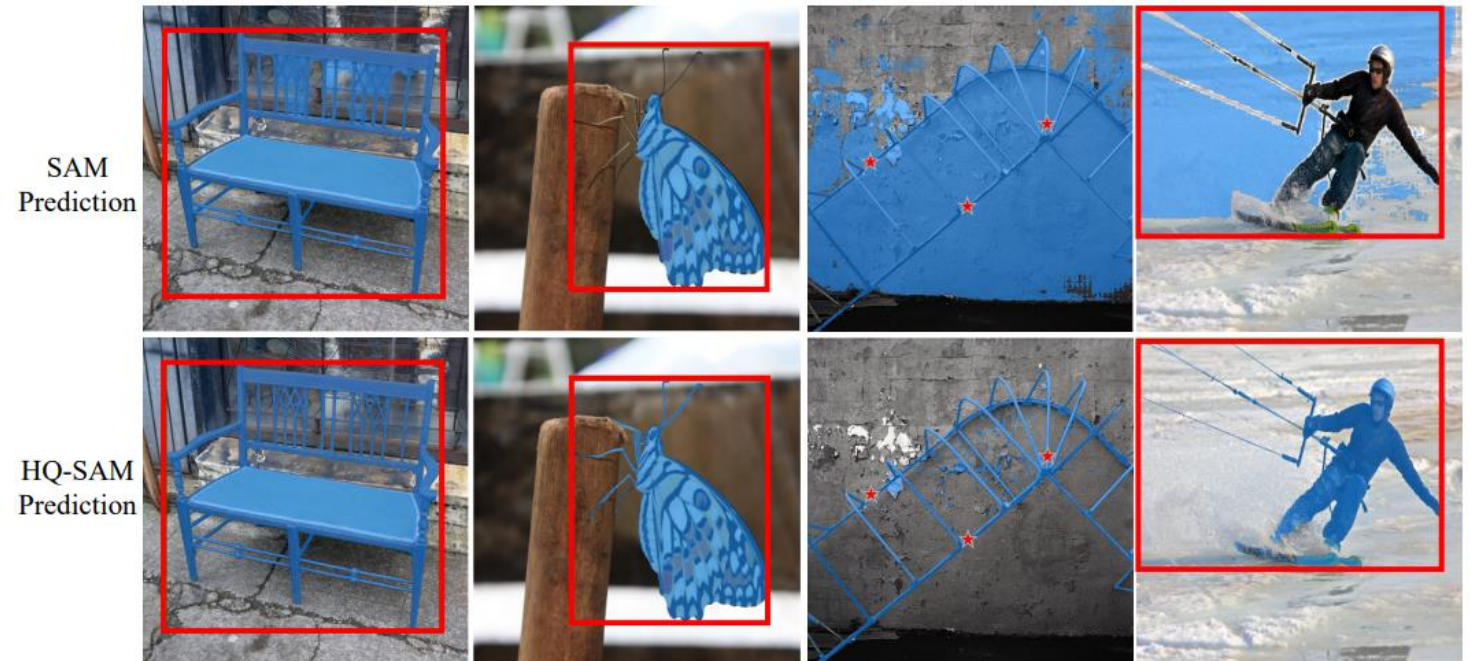
- The segmentation result for the one point prompt.
- Compared to RITM, SAM demonstrates superior performance on 16 out of 23 datasets.

- Comparison based on mask quality scores assigned by human annotators.

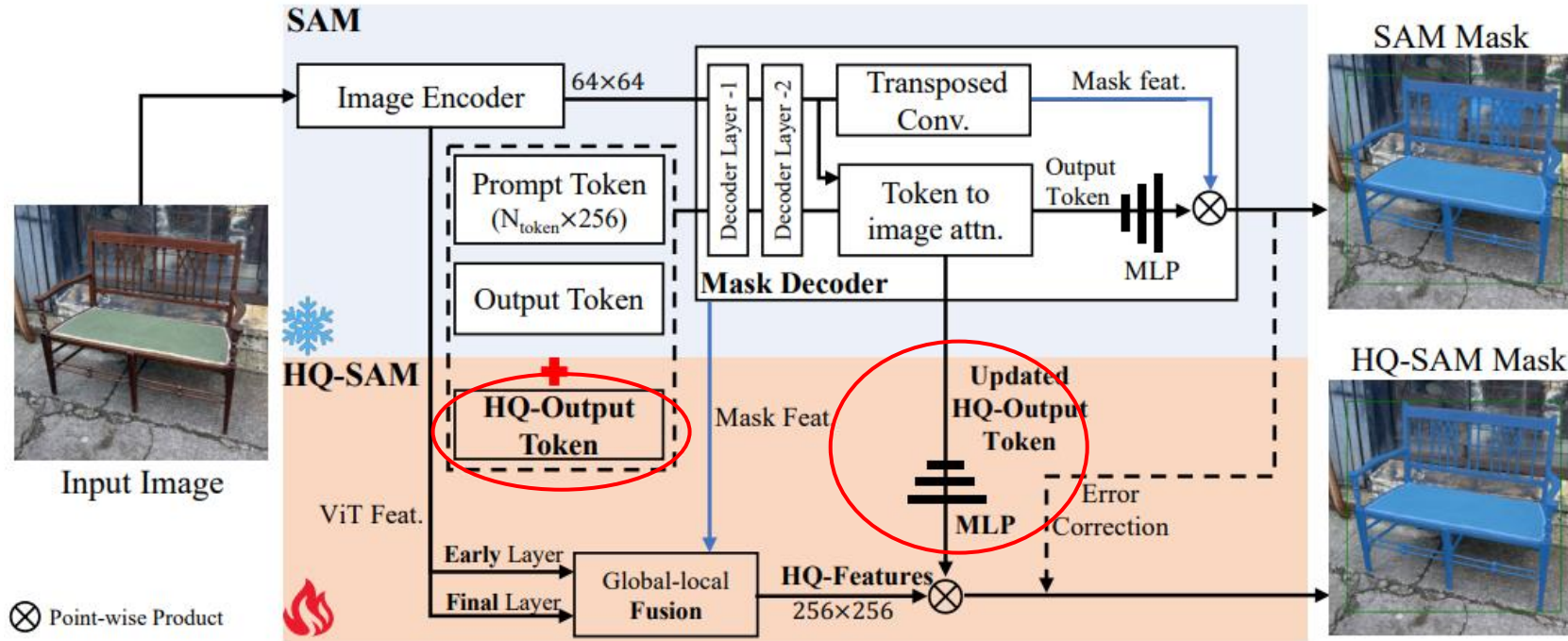
## Segment Anything in High Quality

1. It often ignores segmentation of thin object structures.
2. It introduces large errors for broken masks and tricky cases.

Propose HQ-SAM, which can predict a highly accurate segmentation mask even in very difficult cases without compromising the powerful zero-shot capability and flexibility of SAM.



# 4. HQ-SAM



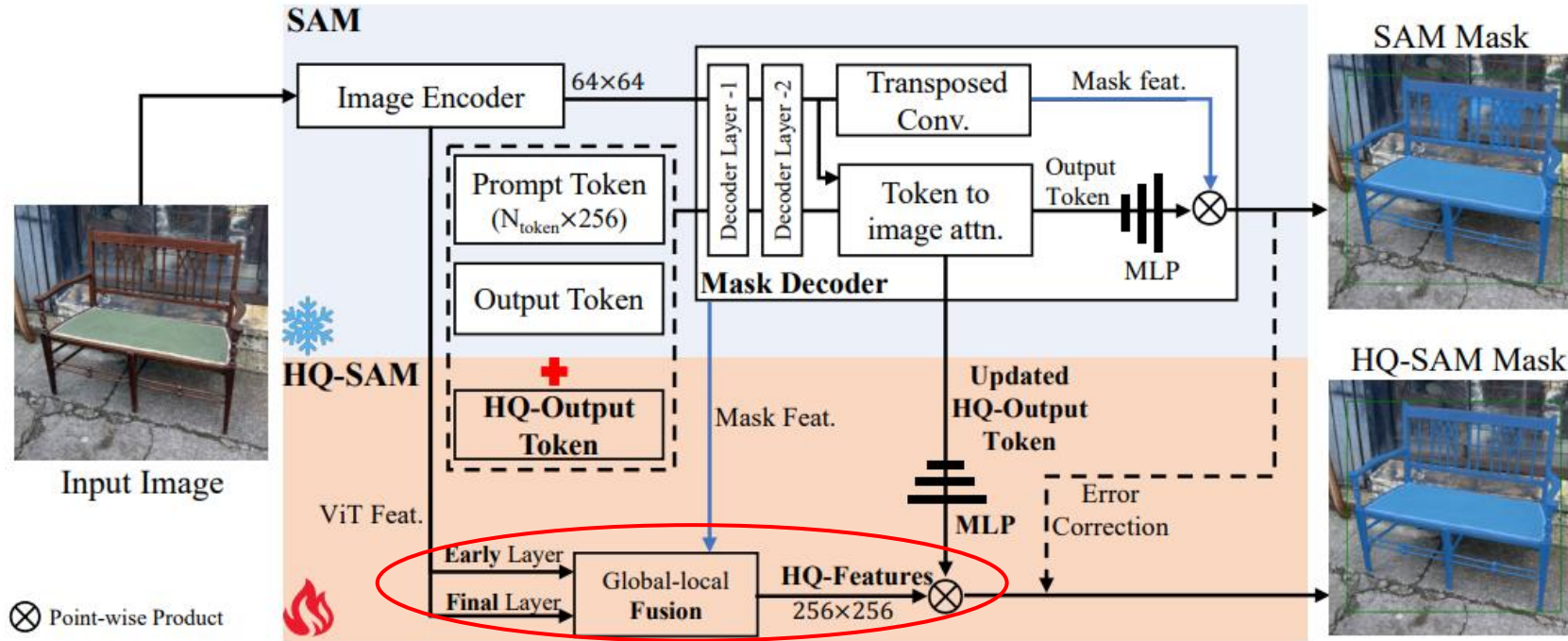
- HQ-Output Token

-SAM : prompt token( $N_{\text{prompts}} \times 256$ ) + output token( $4 \times 256$ ) -> decoder

-HQ-SAM : : prompt token( $N_{\text{prompts}} \times 256$ ) + output token( $4 \times 256$ ) + HQ-output token -> decoder



# 4. HQ-SAM



- Global-local Fusion for HQ-Features

To effectively learn both the local features such as the edges and boundary details of the image and the global context, the information from the **early layers** and **final layers** of SAM's image encoder is combined for use.



Model	$AP_B^{\text{strict}}$	$AP_{B75}^{\text{strict}}$	$AP_{B50}^{\text{strict}}$	$AP_B$	$AP_{B75}$	$AP_{B50}$	AP
SAM	8.6	3.7	25.6	17.3	14.4	37.7	29.7
HQ-SAM	<b>9.9</b>	<b>5.0</b>	<b>28.2</b>	<b>18.5</b>	<b>16.3</b>	<b>38.6</b>	<b>30.1</b>

1) A table comparing **zero-shot open-world instance segmentation** results in UVO.

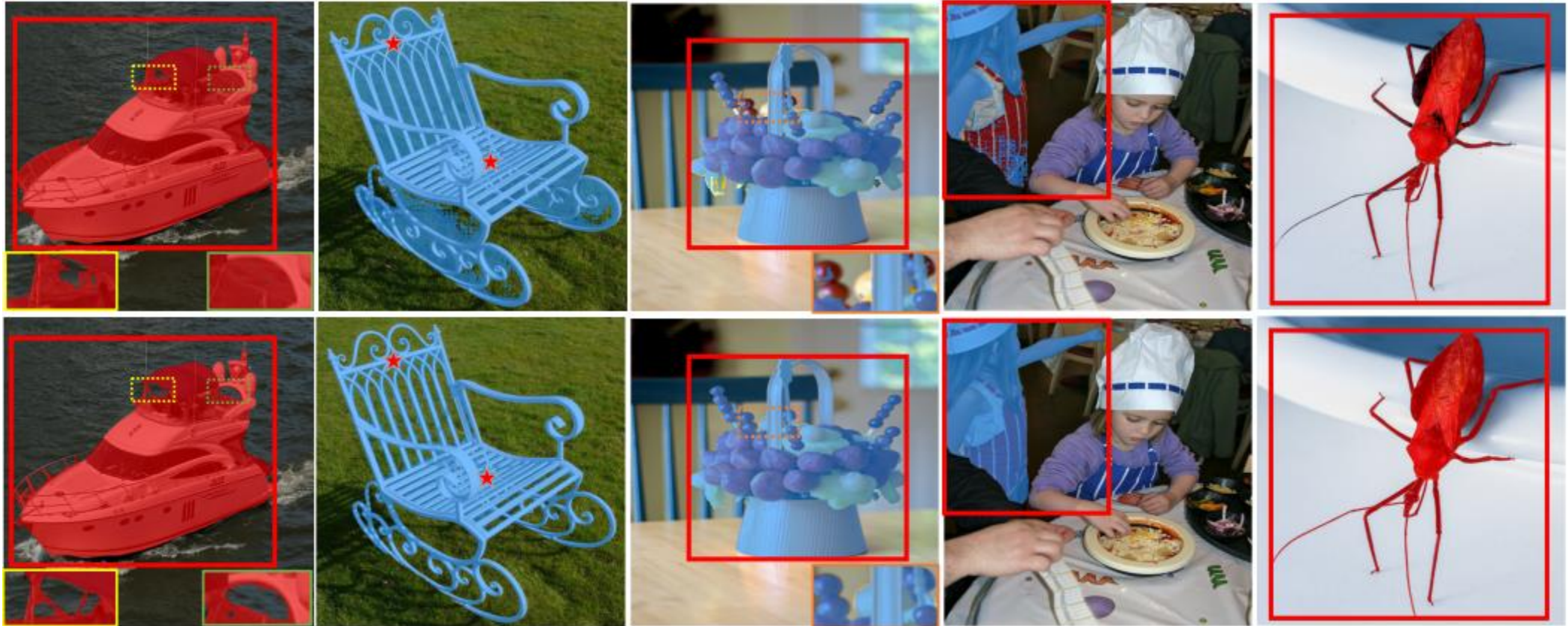
Model	GT Box Prompt		Mask Prompt	
	mIoU	mBIOU	mIoU	mBIOU
SAM	81.1	70.4	66.6	41.8
HQ-SAM	<b>86.0</b>	<b>75.3</b>	<b>86.9</b>	<b>75.1</b>

2) A table comparing **zero-shot segmentation** results on high-quality BIG benchmarks

Model	COCO		LVIS				
	$AP_B$	AP	$AP_B^{\text{strict}}$	$AP_{B75}^{\text{strict}}$	$AP_B$	$AP_{B75}$	AP
SAM	33.3	48.5	32.1	32.8	38.5	40.9	43.6
HQ-SAM	<b>34.4</b>	<b>49.5</b>	<b>32.5</b>	<b>33.5</b>	<b>38.8</b>	<b>41.2</b>	<b>43.9</b>

3) A table comparing **zero-shot instance segmentation** results on COCO and LVISv1 datasets.

# 4. HQ-SAM



# 5. Our proposed Method

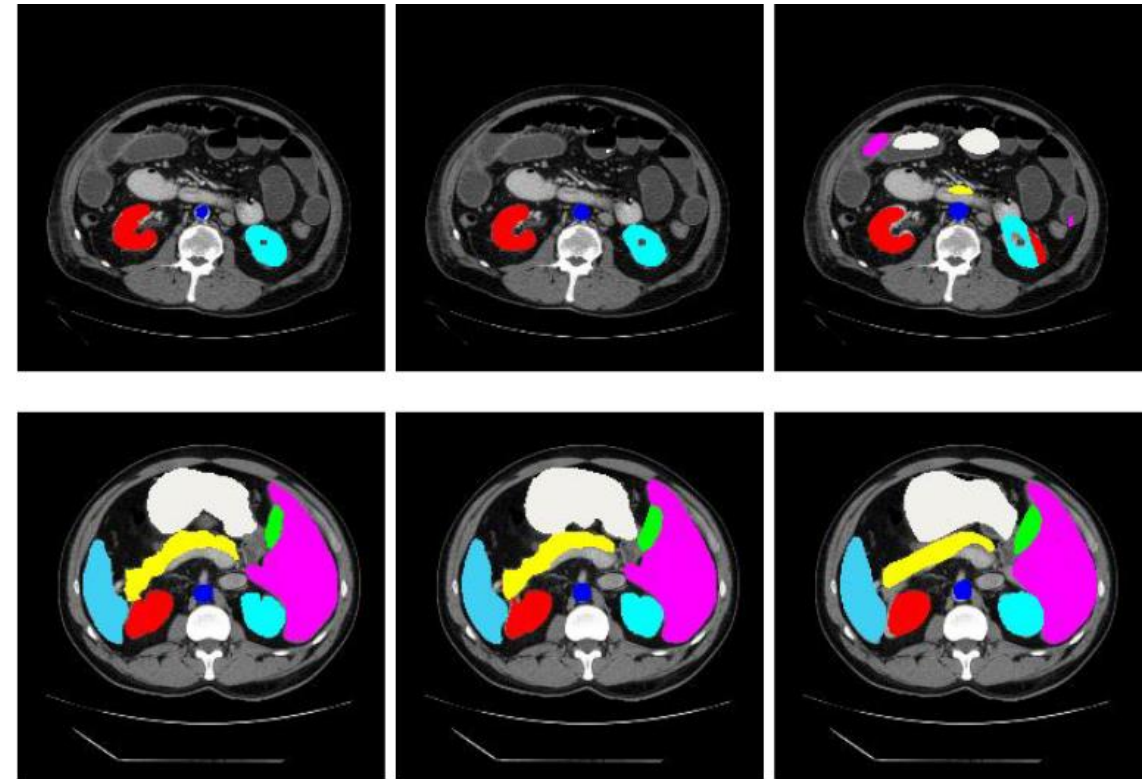
## Medical Image Segmentation

### [Definition]

- Medical Image Segmentation is the process of dividing medical images into anatomical structures, lesions, or other areas of interest.
- It primarily uses images from MRI, CT, ultrasound, and other modalities.

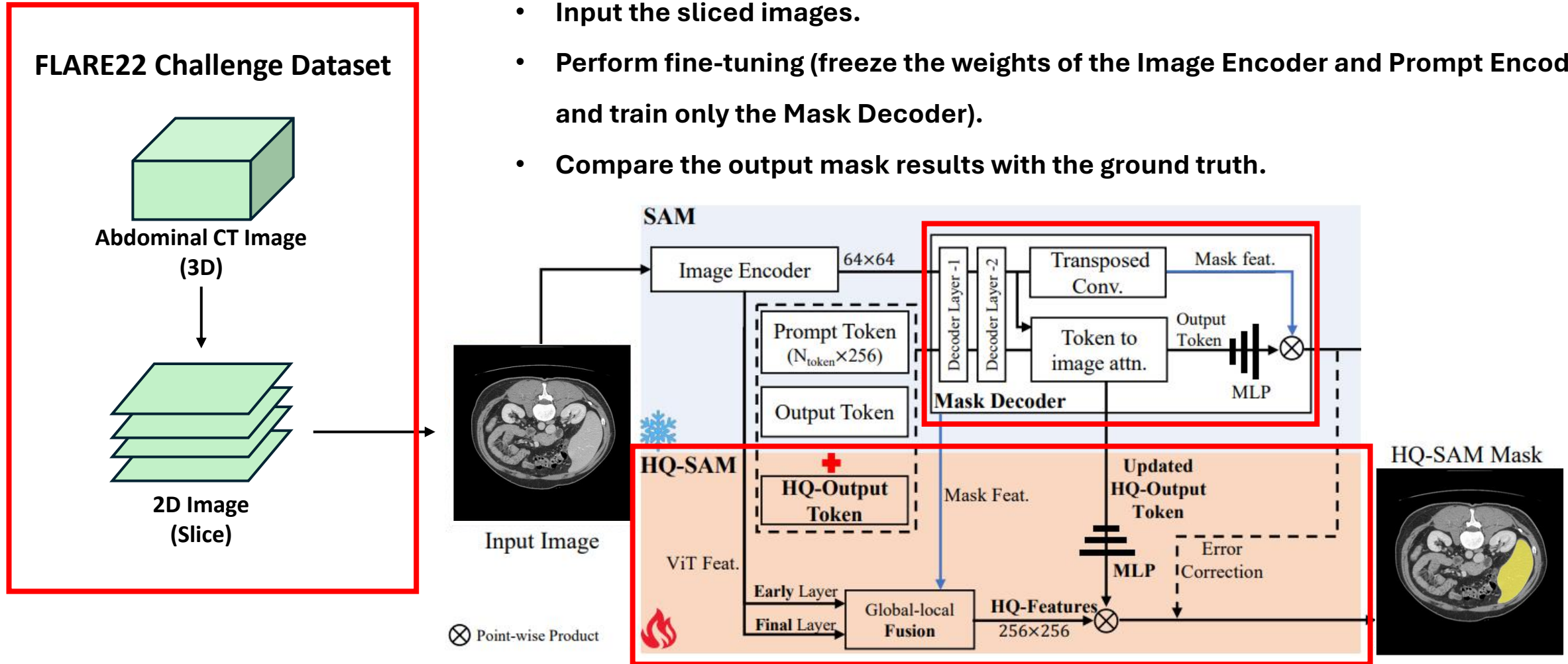
### [Importance]

- Essential for accurate diagnosis and treatment planning.
- Allows precise identification of lesion size, shape, and location.
- Plays a crucial role in surgical planning, radiation therapy, and disease monitoring.



# 5. Our proposed Method

- Slice the 3D abdominal images into 2D images.
- Input the sliced images.
- Perform fine-tuning (freeze the weights of the Image Encoder and Prompt Encoder, and train only the Mask Decoder).
- Compare the output mask results with the ground truth.



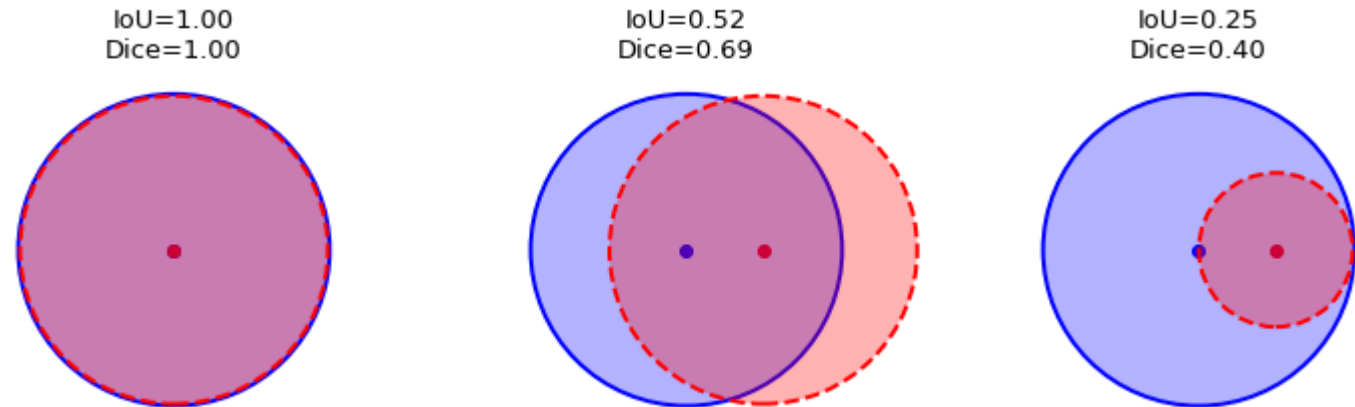


# 5. Our proposed Method

## Dice Score

- It is a measure of the similarity between two sets of data, usually represented as binary arrays.
- The Dice Score is Metric, which values overlapping areas more than the IoU Score, It evaluates the model more reasonably

$$Dice = \frac{2\|A \cap B\|}{\|A\| + \|B\|} = \frac{2TP}{2TP + FP + FN}$$



$$Dice\ loss = 1 - Dice\ Score = 1 - \frac{2 * sum(pred * mask)}{2 * sum(pred * mask) + sum(pred) + sum(mask)}$$



# 6. Experiment

- Compare the output images with the ground truth images (Dice Score).
- Fine-tune SAM, Unet, and DeepLabV3+ in the same manner.
- Evaluate the four methods using the validation dataset. (FLARE22 Challenge Dataset 20%)

## Result

적용모델 \ 평가방식	Dice Score
SAM	0.9169
U-NET	0.8420
DeepLabV3+	0.8994
제안한 방식(HQ-SAM)	0.9359

