# SAM 3 : SEGMENT ANYTHING WITH CONCEPTS

(Meta Superintelligence Labs)

**Paper Review**

2025.11.21

Changseon Yu

# Contents

1. Introduction

2. Purpose

3. Methodology

4. Result

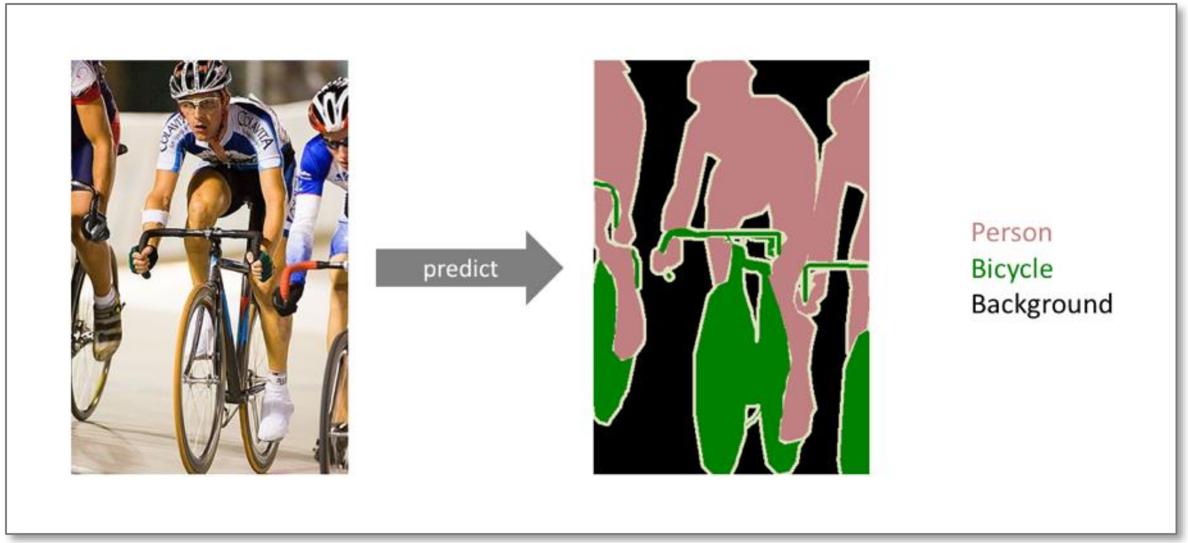
5. Conclusion

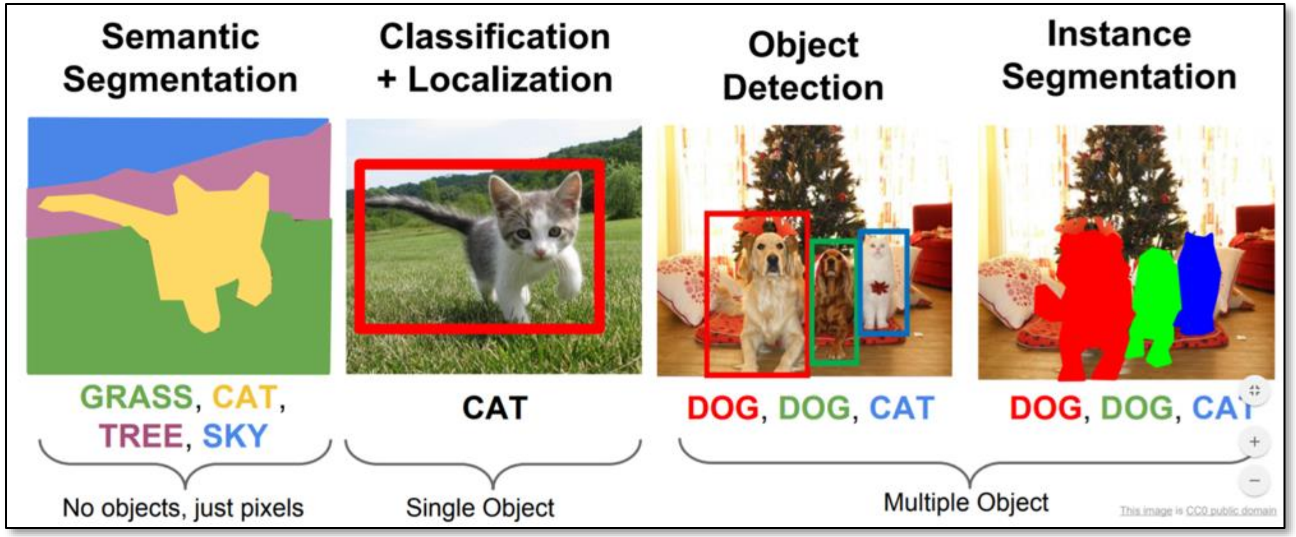
# Introduction

## Background

❖ **Traditional Semantic Segmentation**

- Semantic segmentation : 미리 정의된 클래스 레이블에 해당하는 데이터 (이미지, 3D point cloud, voxel 등)에 대해 객체를 분할
- 사전 정의된 클래스 레이블이 있는 훈련 데이터를 통해 학습하여, 각 픽셀이 특정 클래스에 속할 확률을 예측하는 것



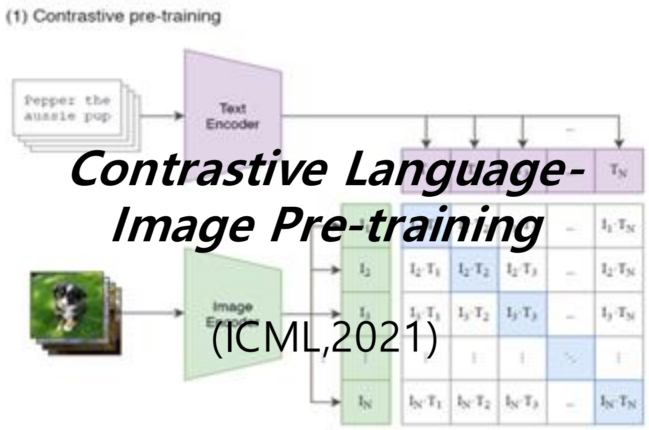


# Introduction

## Background

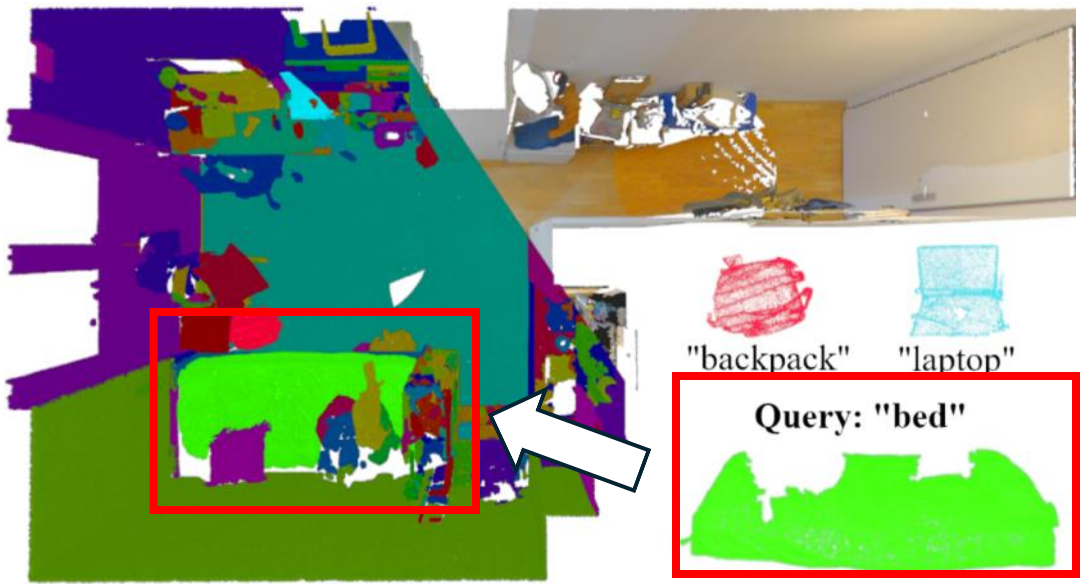
- ❖ **Open-vocabulary Segmentation**

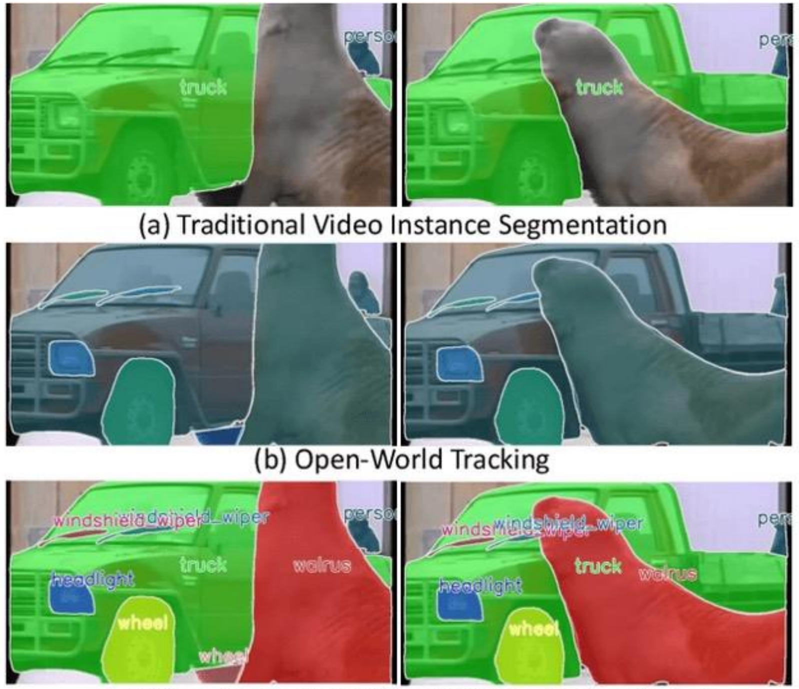
- 훈련 시 보지 못한 개념(텍스트로 설명)을 기반으로 객체를 분할
- 텍스트 설명(단어, 구문)을 이해하고 해당 설명에 맞는 객체를 찾아 분할
- 이미지 + 텍스트 쿼리 → 쿼리에 해당하는 객체 마스크



***Contrastive Language-Image Pre-training***

(ICML,2021)





(a) Traditional Video Instance Segmentation

(b) Open-World Tracking

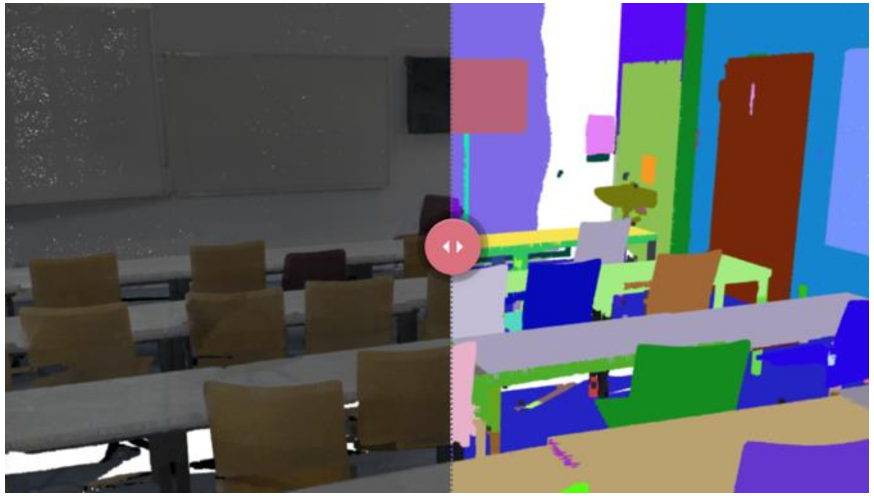
(c) Open-Vocabulary Video Instance Segmentation

# Introduction

## Background

❖ **Class-Agnostic Segmentation**

- 모든 객체 인스턴스를 종류와 무관하게 분할
- 분할된 객체가 무엇인지 명시적(explicit) 지정하지 않음
- 이미지 → 모든 객체 인스턴스 마스크 (label 없음)



**Any3DIS**
**(CVPR 2025)**

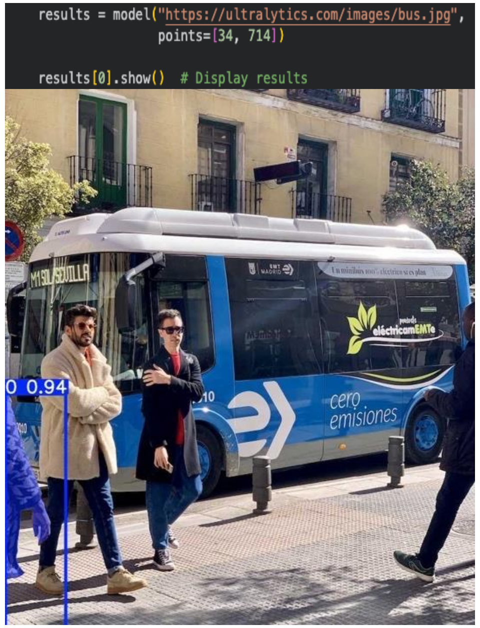
**SAM3D**
**(ECCV 2024)**

# Related works

## Segment Anything (2023)

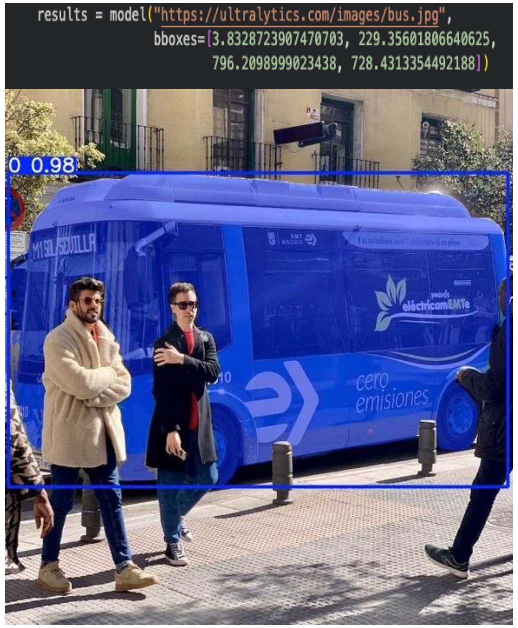
### ❖ Promptable Segmentation

- 1,100만 개의 큐레이팅 된 이미지, 10억 개 이상의 마스크가 포함된 광범위한 SA-1B 데이터 세트 기반 (Data engine 방식)
- 다양한 Prompt(예: point, box, text) 입력에 따라 유효한 Segmentation mask를 생성
- Prompt는 Foreground/Background Points, Rough Box, Mask, 자유 형식 Text 등 이미지에서 무엇을 분할할지 지정하는 모든 정보를 포함



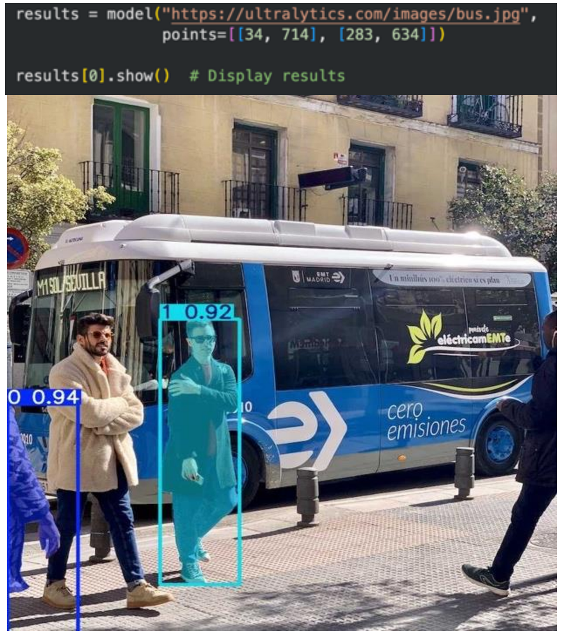
```
results = model("https://ultralytics.com/images/bus.jpg",
                points=[34, 714])

results[0].show()  # Display results
```

**Point prompt**



```
results = model("https://ultralytics.com/images/bus.jpg",
                bboxes=[3.8328723907470703, 229.3560180664062 5,
                        796.2098999023438, 728.4313354492188])
```

**Box prompt**



```
results = model("https://ultralytics.com/images/bus.jpg",
                points=[[34, 714], [283, 634]])

results[0].show()  # Display results
```

**Multiple points prompt**

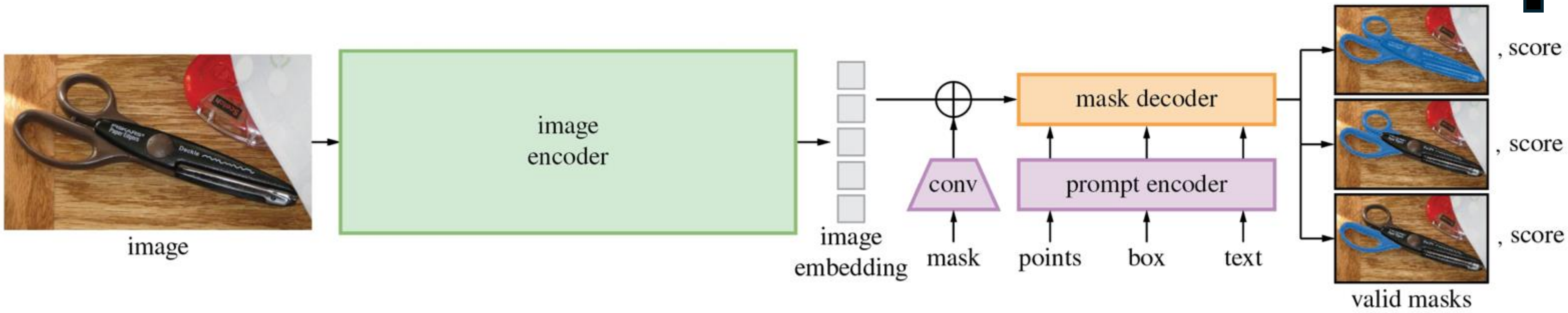
# Related works

## Segment Anything (2023)

### ❖ Architecture

- Image encoder : Masked Autoencoders 로 사전학습 된 Vision transformer 사용
- Conv layer : Dense한 Mask 프롬프트는 Convolution 레이어를 거쳐 이미지 임베딩과 element 별로 합쳐짐
- Prompt encoder : 모델이 이해할 수 있는 벡터의 형태 임베딩
- Mask decoder : 최종 분할 Mask 예측, Light weight한 특성으로 빠르게 동작

**Score : 마스크의 신뢰도, 예상 IoU**



# Related works

## Segment Anything (2023)

### ❖ Data Engine

- **Segmentation mask는 웹 상에 풍부하지 않음 → 자체 데이터 생산 & 모델 개선**

1. 사람이 SAM을 annotation 도구로 사용해서 Point 클릭으로 mask labeling (Annotation 시간 34s → 14s)
2. SAM이 신뢰도 높은 Mask 자동생성(bbox detector 기준) → 사람이 unlabeled 객체를 수동 annotation
3. 32x32 Grid의 point들이 prompt → 유효 객체 segment & NMS 중복제거

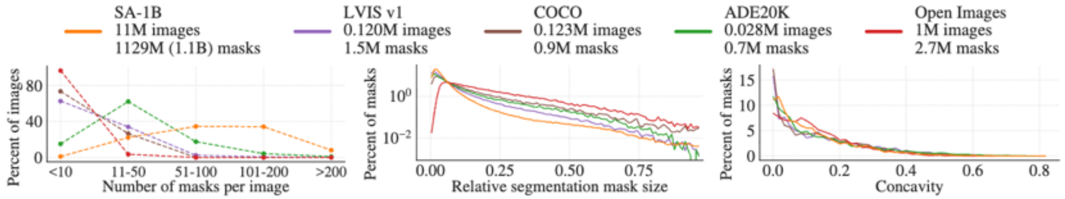


Figure 6: Dataset mask properties. The legend references the number of images and masks in each dataset. Note, that SA-1B has 11× more images and 400× more masks than the largest existing segmentation dataset Open Images [60].

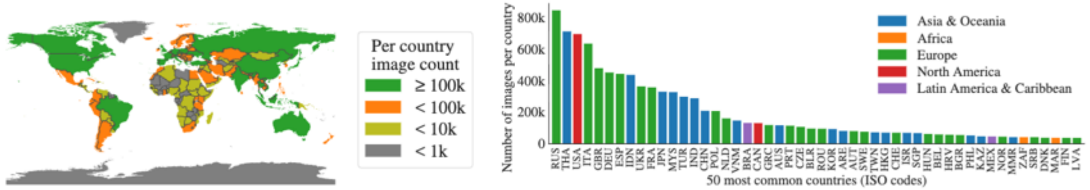
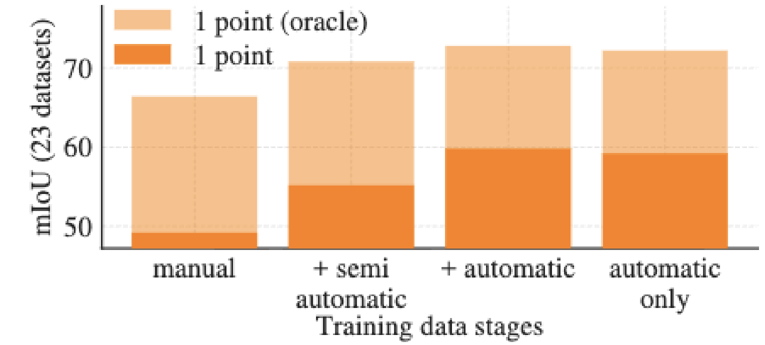


Figure 7: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.



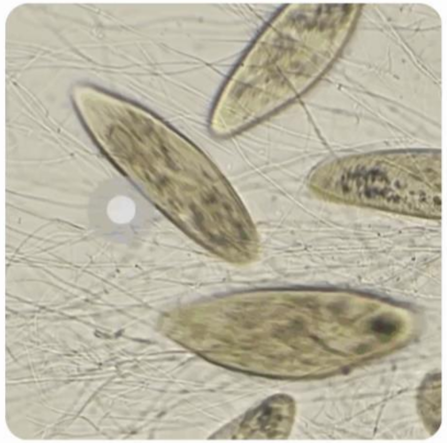
# Related works

## Segment Anything 2 (2024)

❖ **Promptable Visual Segmentation (PVS)**

- 기존 Image segmentation를 video domain으로 일반화
- 사용자는 어느 frame에서든 관심 객체를 point, box, mask 형태로 프롬프트 입력가능 (첫 frame 이외 가능)

Select objects and make adjustments across video frames



Robust segmentation, even in unfamiliar videos

Real-time interactivity and results

# Related works

## Segment Anything 2 (2024)

❖ **Promptable Visual Segmentation (PVS)**

- 기존 Image segmentation를 video domain으로 일반화
- 사용자는 어느 frame에서든 관심 객체를 point, box, mask 형태로 프롬프트 입력가능 (첫 frame 이외 가능)
- 기존 첫 프레임에만 prompt 부여하는 Video object detection을 포괄하는 개념

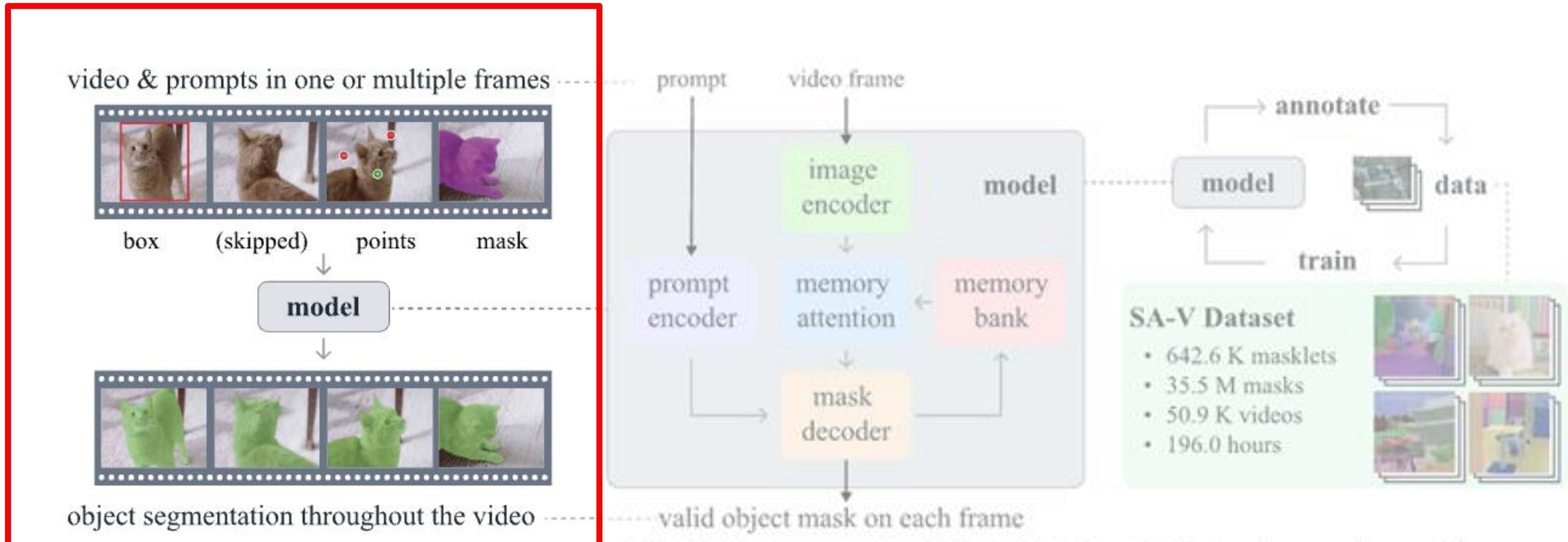
**SAM2**

# Related works

## Segment Anything 2 (2024)

### ❖ Promptable visual segmentation

1. 비디오의 특정 frame에 사용자가 관심 객체를 box, point 또는 mask 같은 prompt로 제공
2. SAM2는 프롬프트를 바탕으로 비디오 전체 프레임에 걸쳐 자동으로 segmentation 진행
3. 모델은 비디오의 프레임별로 유효한 객체 mask를 생성해서 비디오 내내 객체를 추적하고 분할



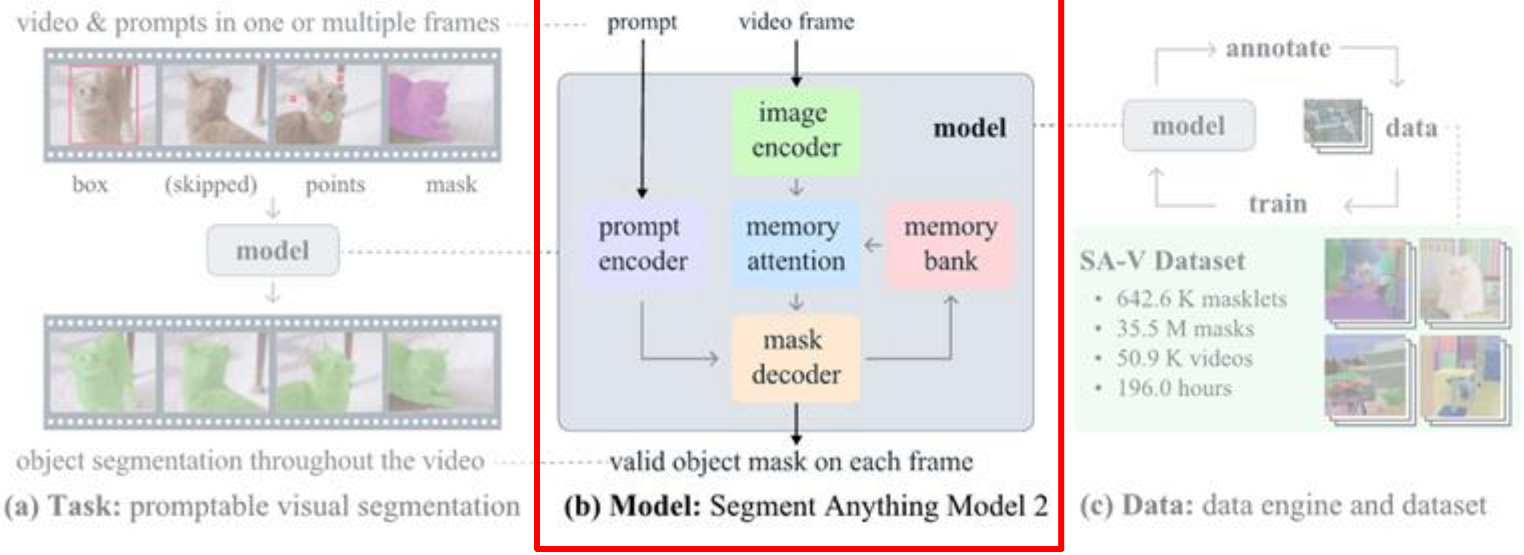
(a) **Task:** promptable visual segmentation (b) **Model:** Segment Anything Model 2 (c) **Data:** data engine and dataset

# Related works

## Segment Anything 2 (2024)

❖ **Architecture**

1. Image encoder : 입력된 비디오의 프레임을 처리해서 특징 추출
2. Prompt encoder : 사용자가 제공한 프롬프트를 임베딩 전환
3. Memory bank : 이전 프레임에서 처리된 객체 정보, 사용자 상호작용 및 예측 결과 저장
4. Memory attention : 현재 프레임 이미지 특징과 이전 memory 통합 → **객체 시공간적 일관성 (temporal consistency) 유지하며 분할**



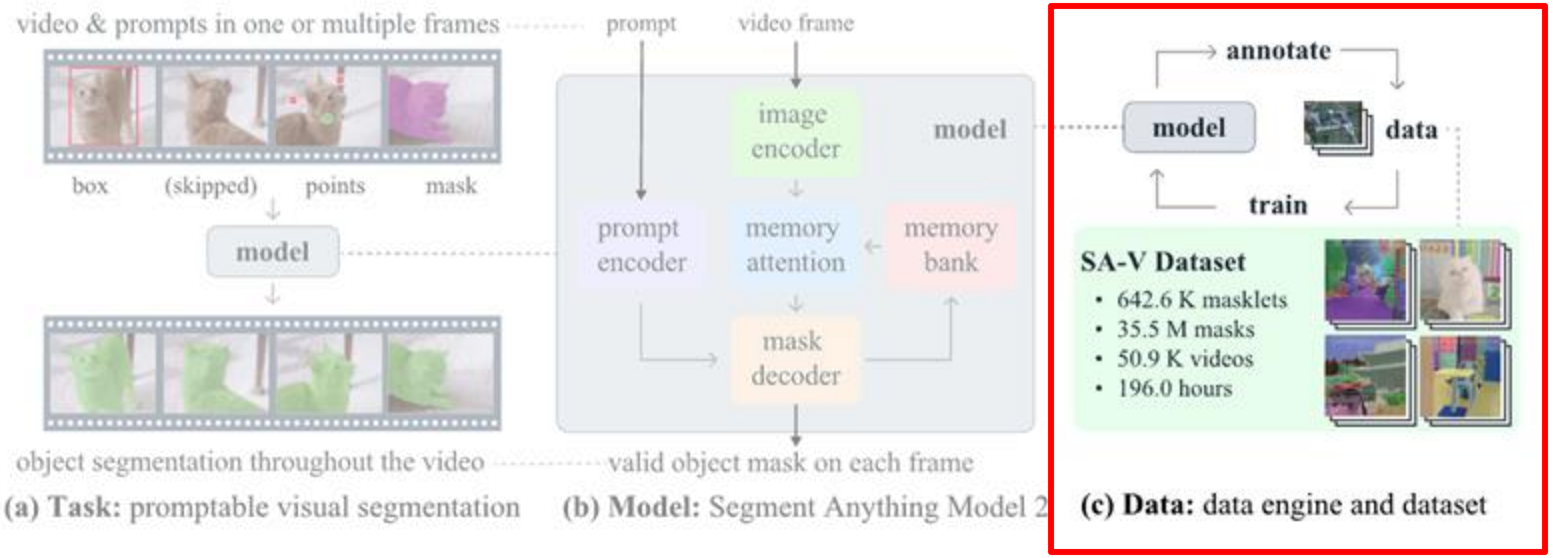
(a) **Task:** promptable visual segmentation (b) **Model:** Segment Anything Model 2 (c) **Data:** data engine and dataset

# Related works

## Segment Anything 2 (2024)

❖ **Data engine**

1. SAM per frame : 이미지 기반 SAM 모델로 비디오 프레임별 Annotation (1,400 videos, 16,000 masklets)
2. SAM + SAM 2 mask : 첫 frame에서 mask 생성 후 SAM 2 mask 이용해 다른 frame으로 확장 (63,500 masklets)
3. SAM 2 : point, mask 포함한 프롬프트 통해 temporal consistency 가진 SAM 2로 예측 (197,000 masklets)
4. Quality Verification : annotation masklet은 satisfactory /unsatisfactory 검증수행



(a) **Task:** promptable visual segmentation

(b) **Model:** Segment Anything Model 2

(c) **Data:** data engine and dataset

**SA-V Dataset**

**총 50.9K 비디오**

**642.6K 마스크릿**

**35.5M 마스크**

**총 196.0 시간 비디오**

# Introduction

## Background

### ❖ Promptable Concept Segmentation (PCS)

- Concept : simple noun phrases (NPs) , Image exemplars
- All prompts must be consistent in their category definition
- Concept based Open vocabulary segmentation

Positive example

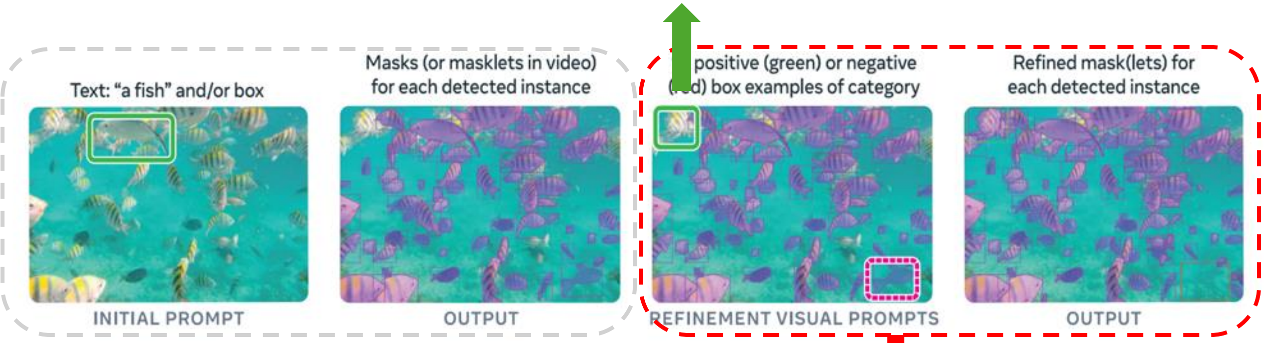

Text: "a fish" and/or box
Masks (or masklets in video) for each detected instance
positive (green) or negative (red) box examples of category
Refined mask(lets) for each detected instance
INITIAL PROMPT
OUTPUT
REFINEMENT VISUAL PROMPTS
OUTPUT

Figure 2: Illustration of supported initial and optional interactive refinement prompts in the PCS task.

Negative example

SAM & SAM2

SAM 3

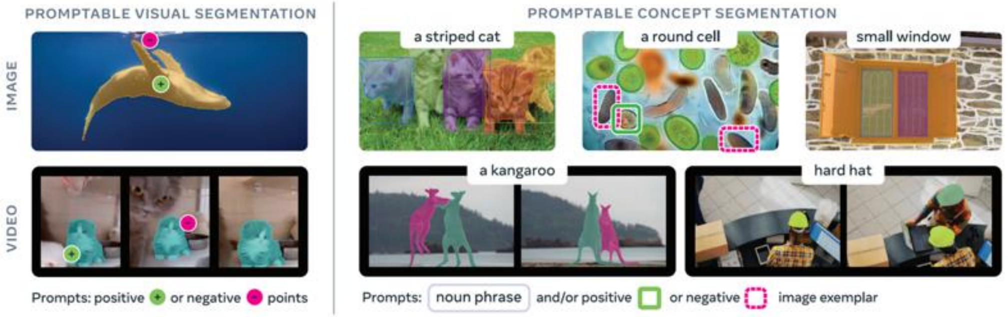

PROMPTABLE VISUAL SEGMENTATION
PROMPTABLE CONCEPT SEGMENTATION
IMAGE
VIDEO
a striped cat
a round cell
small window
a kangaroo
hard hat
Prompts: positive ● or negative ● points
Prompts: noun phrase and/or positive □ or negative ⬚ image exemplar

Figure 1: SAM 3 improves over SAM 2 on promptable *visual* segmentation with clicks (left) while advancing promptable *concept* segmentation (right) where users can segment all instances of a visual concept specified by a short noun phrase, image exemplars, or a combination of both.

# Proposal

1. Returns segmentation masks and unique identities for **all matching object instances**
2. Produces a high-quality dataset with 4M unique concept labels(hard negatives, images, videos)
3. **Presence head boosts detection accuracy** in recognition & localization

# Methodology

## Overview

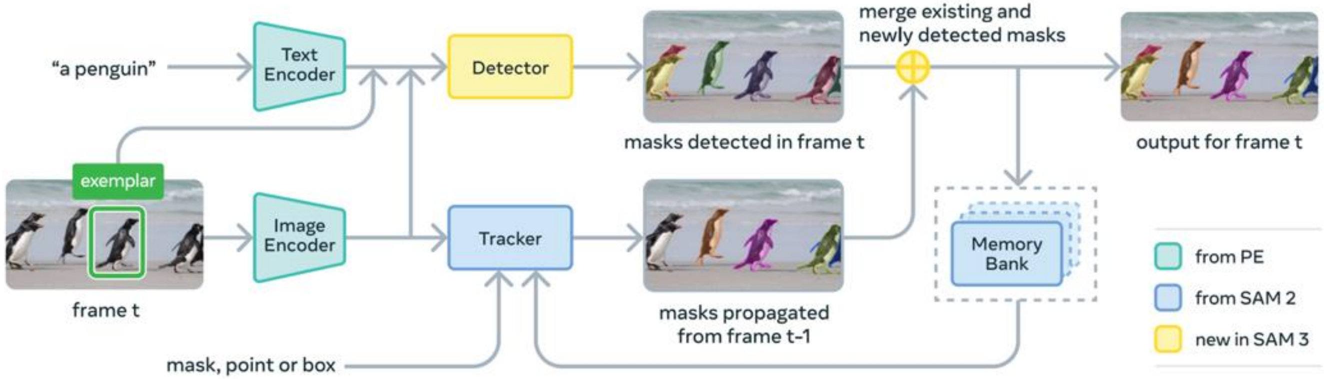


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.

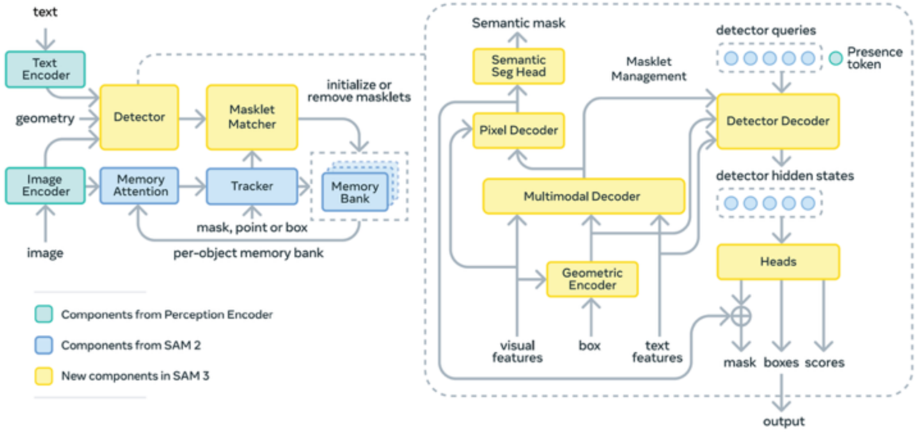


Figure 8: SAM 3 architecture. We highlight new components in yellow, SAM 2 (Ravi et al., 2024) in blue and PE (Bolya et al., 2025) in cyan.

# Methodology

## Architecture

### ❖ Inputs

- Input text query (“a penguin”)
- Green bounding box : Image exemplar (positive / negative)
- Visual prompt : mask, point or box (refinement, SAM 2 PVS)

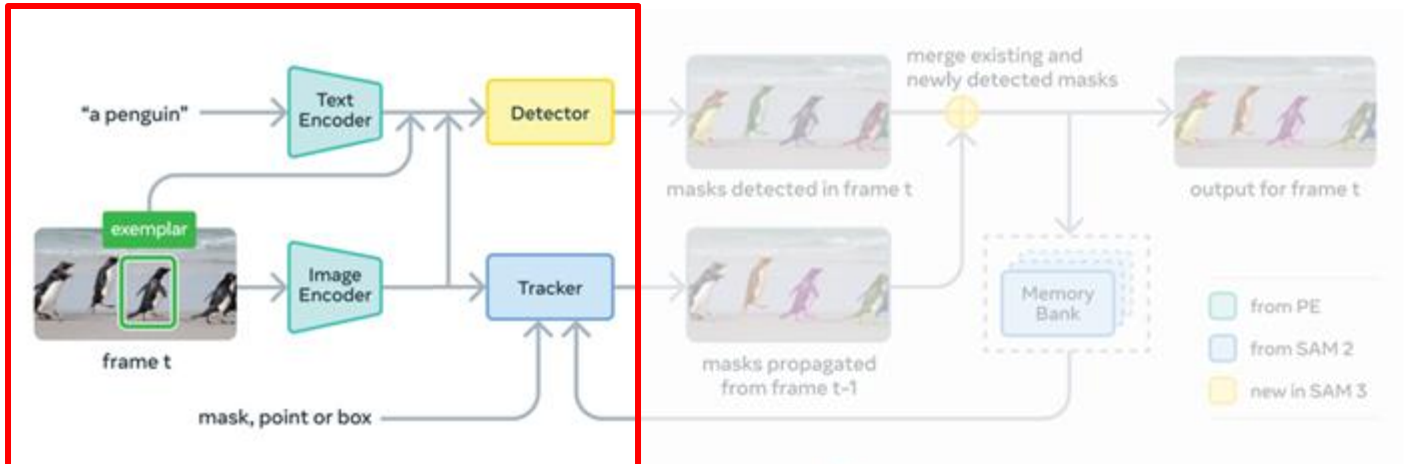


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.

# Methodology

## Architecture

### ❖ Detector

- Detects all object instances that match the concept prompt in the current frame and generate masks
- *“masks detected in frame t”*

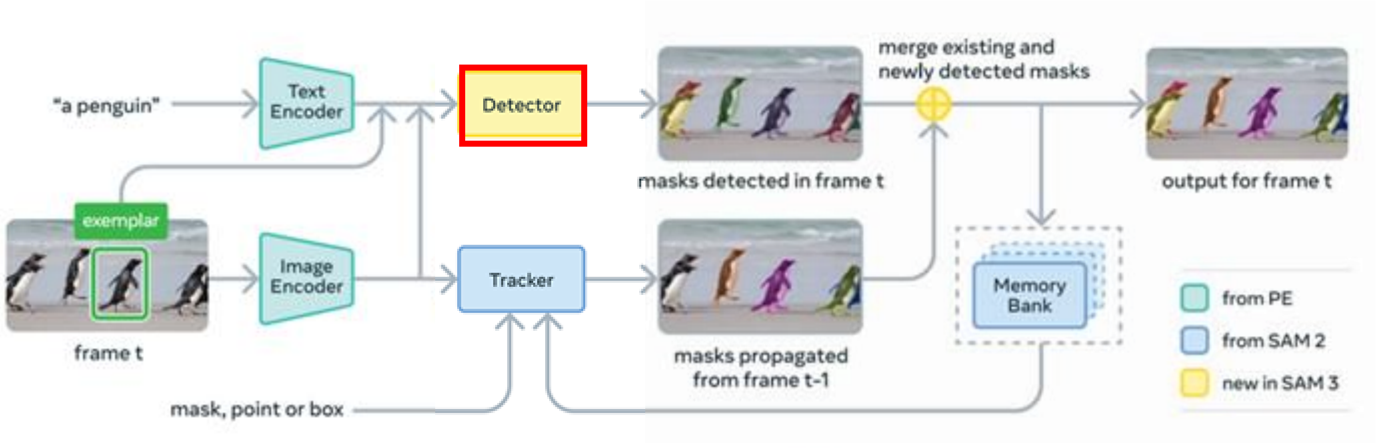


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.

# Methodology

## Architecture

### ❖ Tracker

- Receive visual features of the current frame(t) from the image encoder
- Use object information(t-1) and refining information through mask and point or box in the memory bank
- Propagate a mask of objects tracked in the previous frame t-1 to a new location in the current frame t

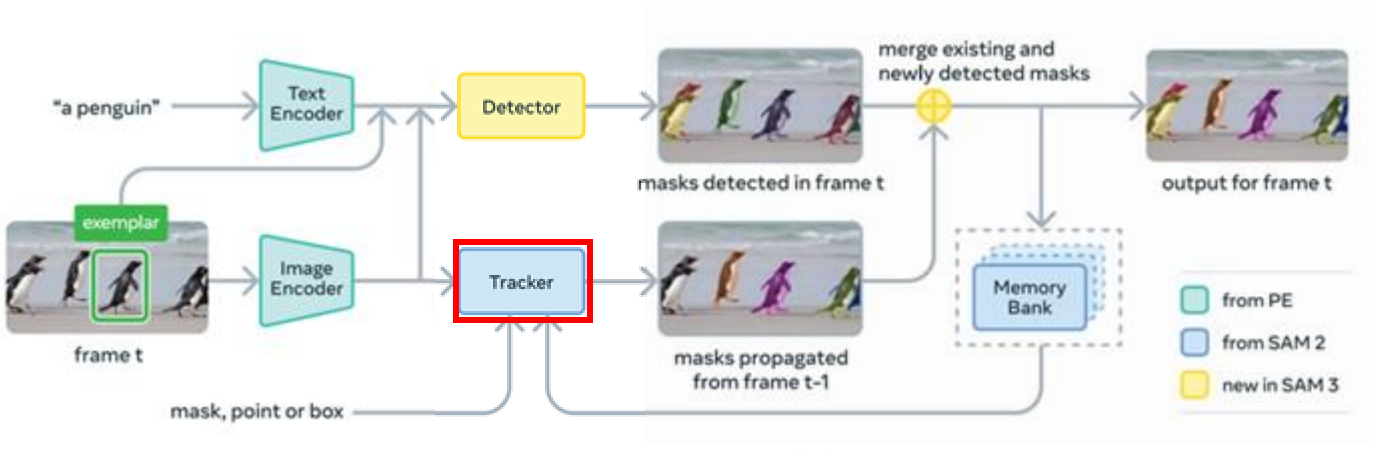


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.

# Methodology

## Architecture

### ❖ Memory bank

- Store and manage the visual features and identity information (ID) of the objects being tracked over time
- Provides the information for the tracker to maintain and track the identity of the objects, and is updated from the final output mask

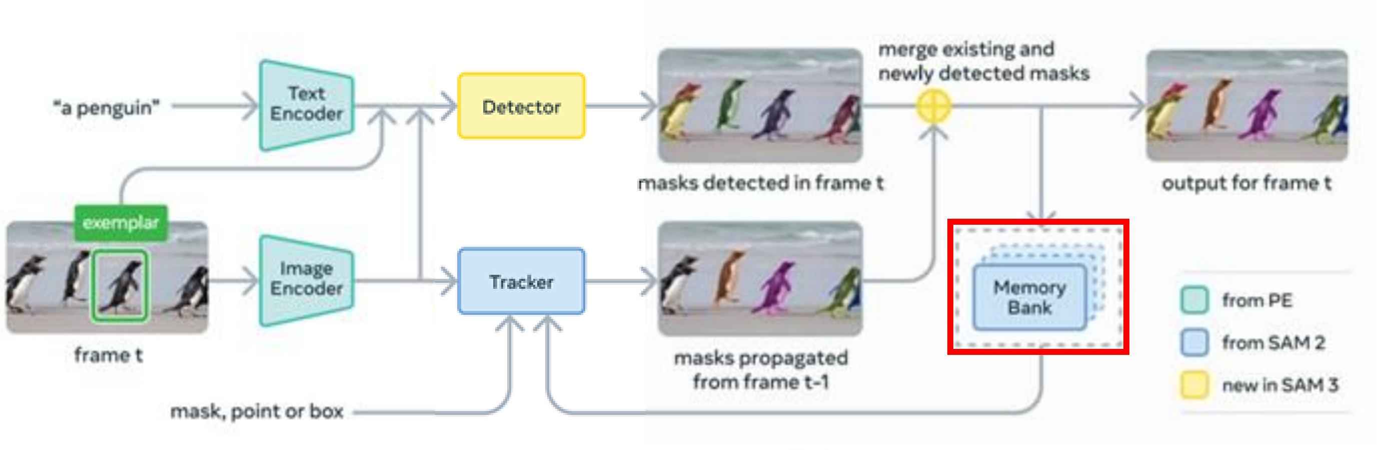


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.

# Methodology

## Architecture

❖ **Merge existing and newly detected masks**

- Provides the information for the tracker to maintain and track the identity of the objects, and is updated from the final output mask
- Combine newly detected masks from the detector with existing masks propagated from the tracker
- Determine matching score between objects, add new objects, or process missing objects

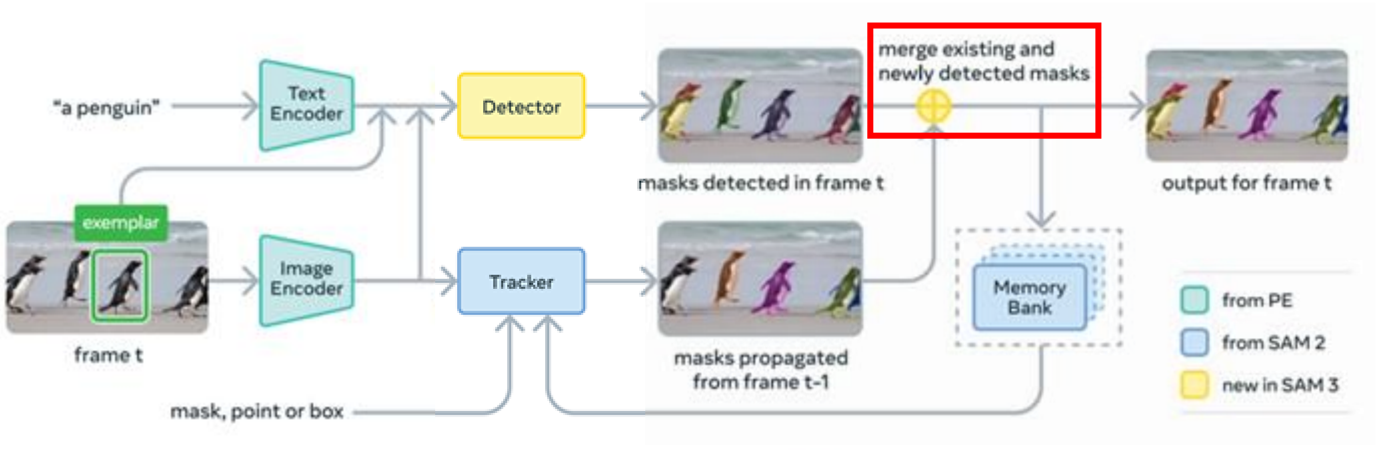


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.

- **segmentation mask**
- **Each IDs**

# Methodology

## Detail Architecture

### ❖ Inputs

- Visual features: Image features extracted from the image encoder
- Box: Box prompts are processed through the Geometric Encoder
- Text features: Noun phrases are embedded via the text encoder

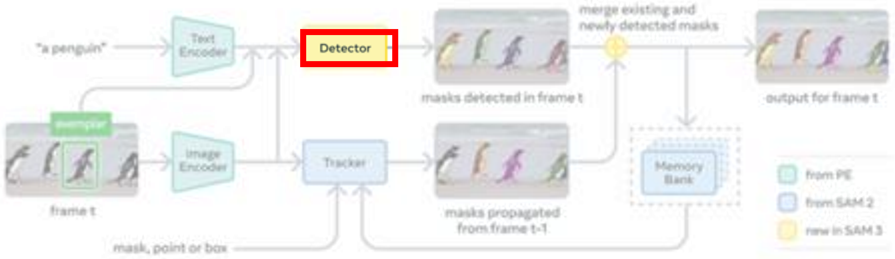
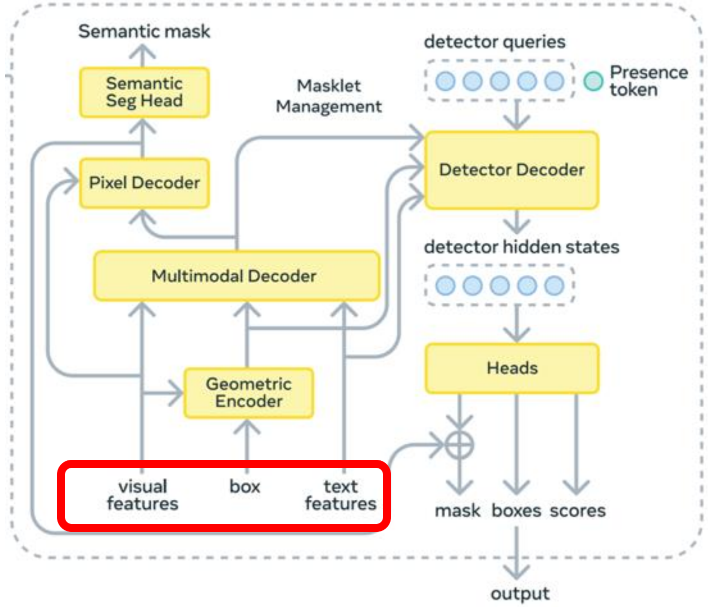


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.



# Methodology

## Detail Architecture

### ❖ Multimodal Decoder (Fusion Encoder)

- Combines the Visual features with the prompt tokens (Text features and Geometric Encoder outputs)
- "Conditions" the image features based on the provided prompts

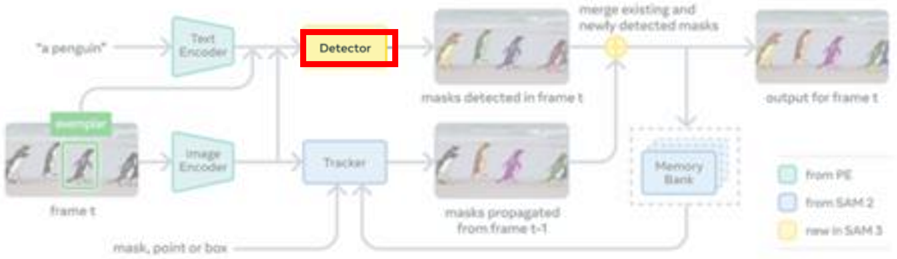
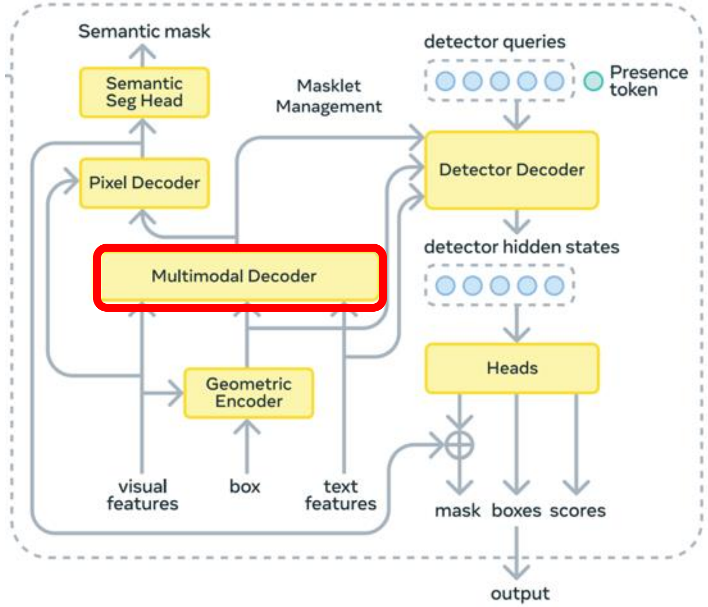


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.



# Methodology

## Detail Architecture

### ❖ Semantic Segmentation

- The conditioned features are passed to the Pixel Decoder
- The Semantic Seg Head then generates a Semantic mask for the whole image

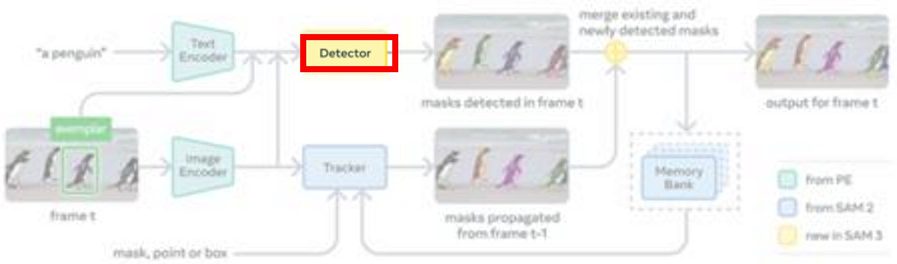
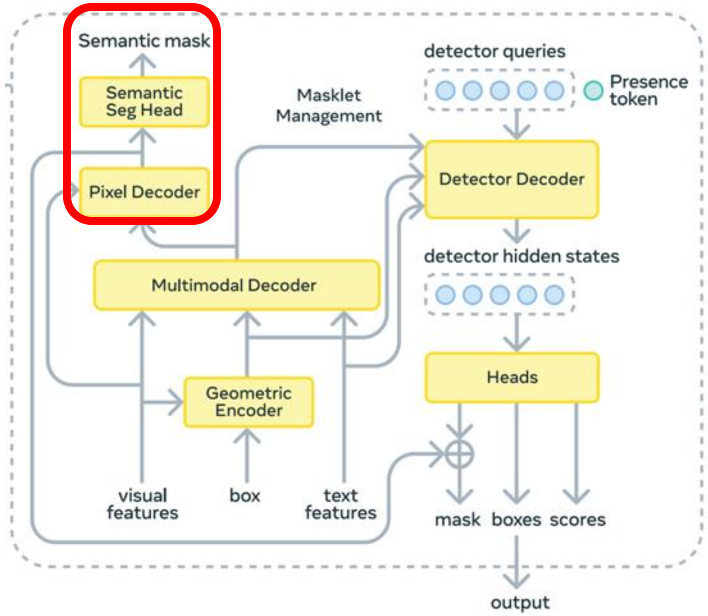


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.



# Methodology

## Detail Architecture

### ❖ Detector Decoder

- Detector Queries: Learnable object queries used to find objects
- Presence token : solely predicts the Presence score whether the concept exists in the image
- Detector Decoder: Queries perform cross-attention with the conditioned image features to extract object information

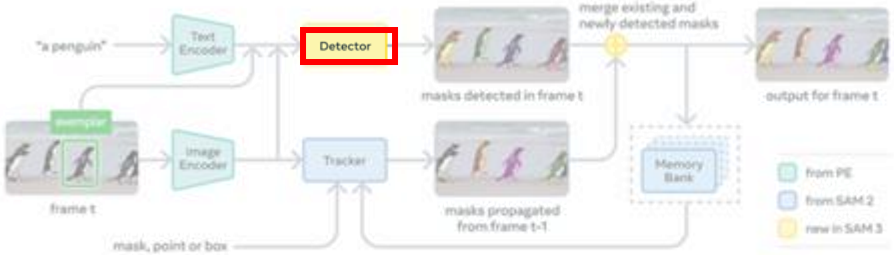
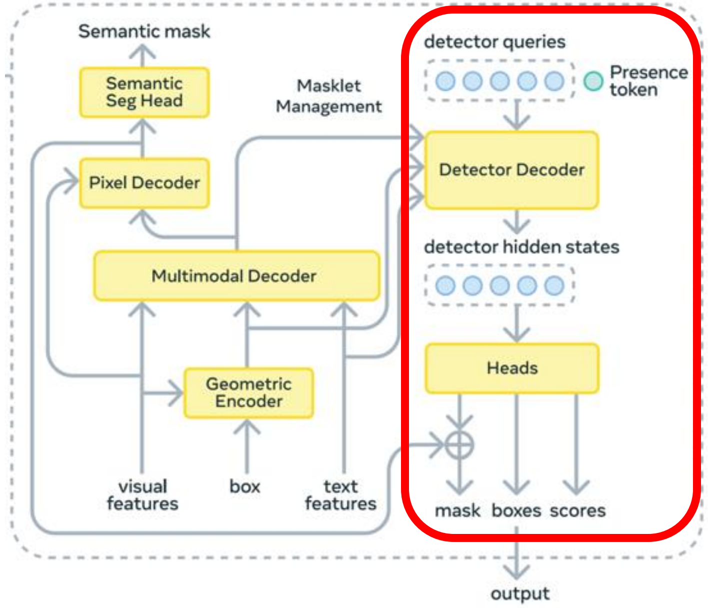


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.



# Methodology

## Detail Architecture

### ❖ Heads & Output

- The final score is individual query score and the Presence token score (separating recognition from localization)
- Scores: Predicts the probability of an object
- Boxes: Predicts bounding boxes.
- Mask: Generates the final instance Mask by combining outputs with the Pixel Decoder

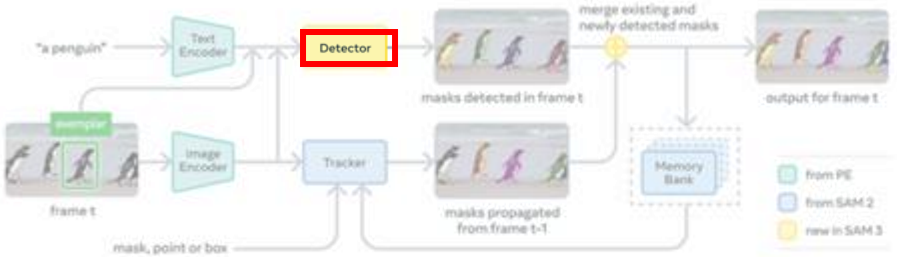
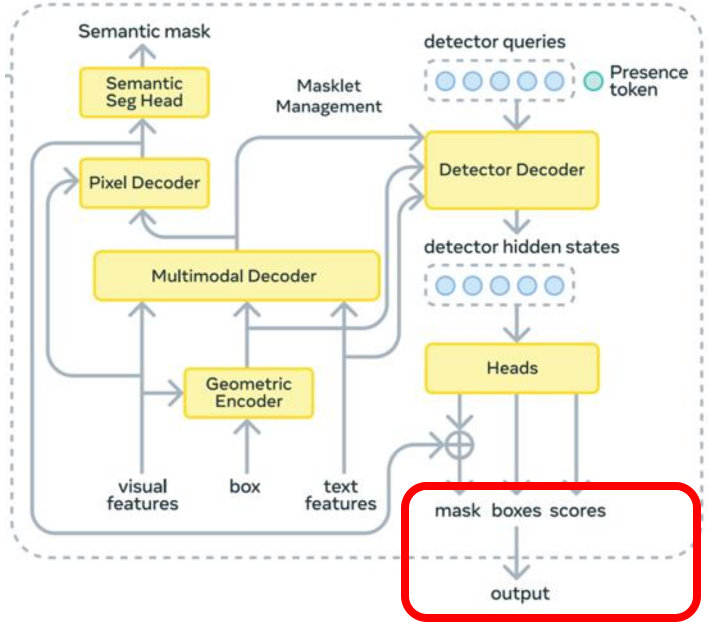


Figure 3: SAM 3 architecture overview. See Fig. 8 for a more detailed diagram.



# Experiments

## Data Engine

### ❖ Phase 1 : Human Verification

- Mine data : Randomly sampling
- Propose NPs : Utilize simple captioner
- Propose mask : SAM 2
- Verifier : Human annotator
- Result : SA-Co/HQ 4.3m Image-NP pairs → **SAM 3 Input**

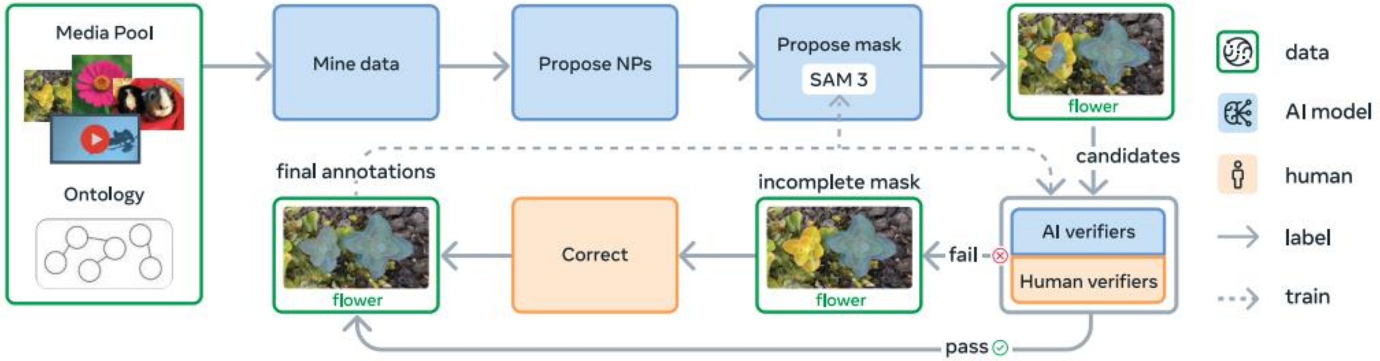


Figure 4: Overview of the final SAM 3 data engine. See §F.1 for details of collected data.

# Experiments

## Data Engine

### ❖ Phase 2 : Human + AI Verification

- AI verifier : Fine tuned Llama 3.2
- Automatic verify process : (Image+query+mask) → **mask quality & perfection**
- NP proposal update : propose hard negative NPs
- Result : **+ SA-Co/HQ 1.2B Image-NP pairs**

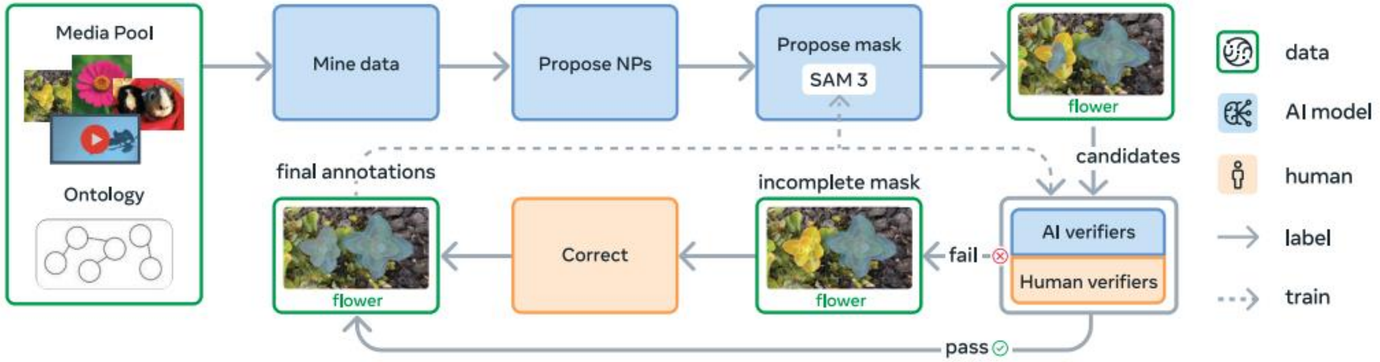


Figure 4: Overview of the final SAM 3 data engine. See §F.1 for details of collected data.

# Experiments

## Data Engine

### ❖ Phase 3 : Scaling and Domain Expansion

- Expand domain (15 datasets)
- **Improve zero shot performance**
- Result : 19m Image-NP pair

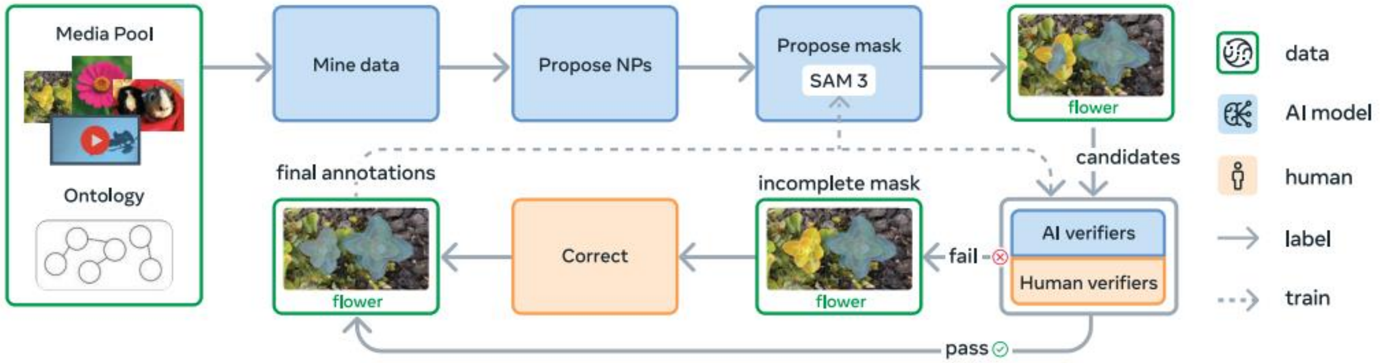


Figure 4: Overview of the final SAM 3 data engine. See §F.1 for details of collected data.

# Experiments

❖ **Dataset**

### SA-Co/HQ (High Quality)

- Data Engine results
- Human verified
- 5.2M images + 4M unique NPs

### SA-Co/SYN (Synthetic)

- Mature version of the data engine without any human involvement

### SA-Co/EXT

- fifteen external datasets

### SA-Co/VIDEO

- 52.5K videos
- 24.8K unique NPs

# Result

## ❖ Metric

- Measure the usefulness of the model in downstream applications
- Average Precision (AP) has limitations because how reliable the model's predictions are calibrated
- Only evaluate predictions with confidence above 0.5

- **Localization** : positive macro F1 (pmF1)
- **Classification :** image-level Matthews Correlation Coefficient (IL MCC) [-1,1]
- **Primary metric** : classification-gated F1 (CGF1)
- $\boldsymbol{CGF1 = 100 * pmF1 * ILMCCC}$

$$F_1^{\tau} = \frac{2 \times TP}{2 \times TP + FP + FN} (\tau = 0.5)$$

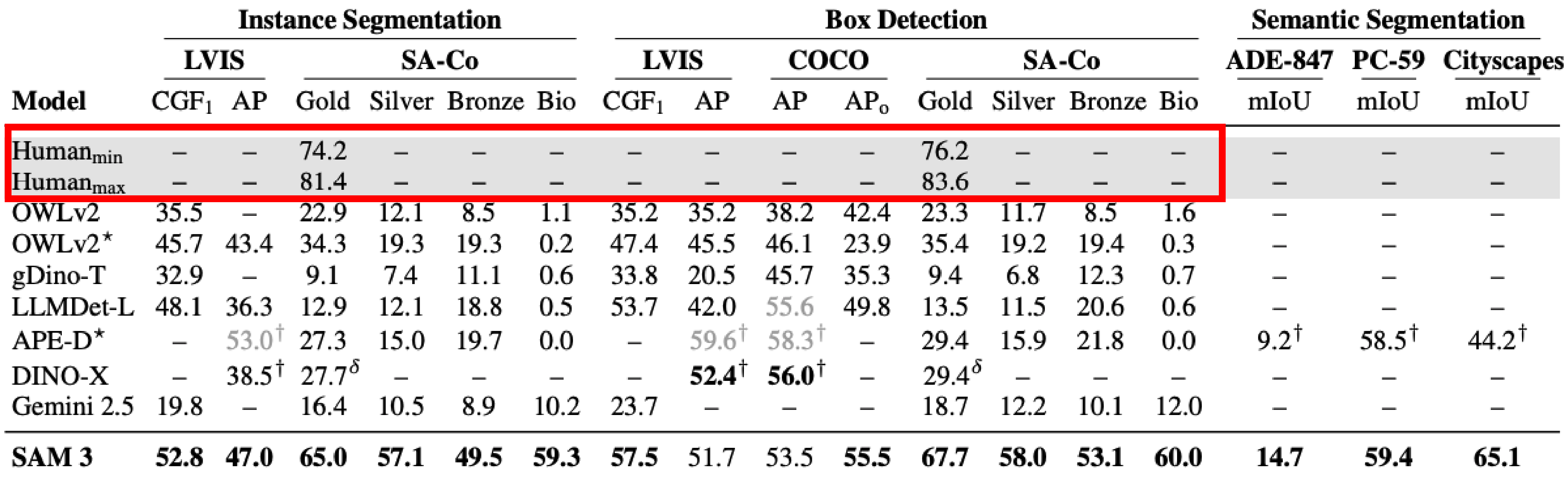
$$pmF_1 = \frac{1}{|T|} \sum_{\tau \in \boldsymbol{T}} F_1^{\tau} \ , \{T = 0.50, 0.55, \ldots, 0.95\}$$

$$\text{IL } MCC = \frac{IL\,TP \times IL\,TN \ - IL\,FP \ \times \ IL\,FN}{\sqrt{(IL\,TP + IL\,FP)(IL\,TP + IL\,FN)(IL\,TN + IL\,FP)(IL\,TN + IL\,FN)}}$$

# Result

❖ **Evaluation on image concept segmentation with text**

- Measure the usefulness of the model in downstream applications



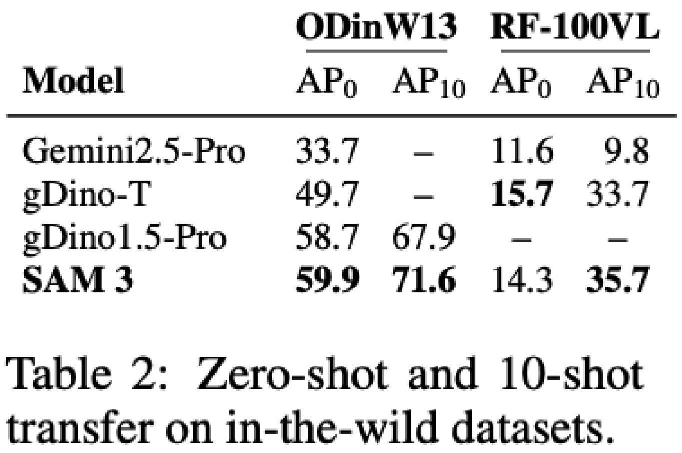
| | Instance Segmentation | | | | | | Box Detection | | | | | | | | Semantic Segmentation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LVIS | | SA-Co | | | | LVIS | | COCO | | SA-Co | | | | ADE-847 | PC-59 | Cityscapes |
| **Model** | CGF$_1$ | AP | Gold | Silver | Bronze | Bio | CGF$_1$ | AP | AP | AP$_o$ | Gold | Silver | Bronze | Bio | mIoU | mIoU | mIoU |
| Human$_{min}$ | – | – | 74.2 | – | – | – | – | – | – | – | 76.2 | – | – | – | – | – | – |
| Human$_{max}$ | – | – | 81.4 | – | – | – | – | – | – | – | 83.6 | – | – | – | – | – | – |
| OWLv2 | 35.5 | – | 22.9 | 12.1 | 8.5 | 1.1 | 35.2 | 35.2 | 38.2 | 42.4 | 23.3 | 11.7 | 8.5 | 1.6 | – | – | – |
| OWLv2$^\star$ | 45.7 | 43.4 | 34.3 | 19.3 | 19.3 | 0.2 | 47.4 | 45.5 | 46.1 | 23.9 | 35.4 | 19.2 | 19.4 | 0.3 | – | – | – |
| gDino-T | 32.9 | – | 9.1 | 7.4 | 11.1 | 0.6 | 33.8 | 20.5 | 45.7 | 35.3 | 9.4 | 6.8 | 12.3 | 0.7 | – | – | – |
| LLMDet-L | 48.1 | 36.3 | 12.9 | 12.1 | 18.8 | 0.5 | 53.7 | 42.0 | 55.6 | 49.8 | 13.5 | 11.5 | 20.6 | 0.6 | – | – | – |
| APE-D$^\star$ | – | $53.0^\dagger$ | 27.3 | 15.0 | 19.7 | 0.0 | – | $59.6^\dagger$ | $58.3^\dagger$ | – | 29.4 | 15.9 | 21.8 | 0.0 | $9.2^\dagger$ | $58.5^\dagger$ | $44.2^\dagger$ |
| DINO-X | – | $38.5^\dagger$ | $27.7^\delta$ | – | – | – | – | $\mathbf{52.4}^\dagger$ | $\mathbf{56.0}^\dagger$ | – | $29.4^\delta$ | – | – | – | – | – | – |
| Gemini 2.5 | 19.8 | – | 16.4 | 10.5 | 8.9 | 10.2 | 23.7 | – | – | – | 18.7 | 12.2 | 10.1 | 12.0 | – | – | – |
| **SAM 3** | **52.8** | **47.0** | **65.0** | **57.1** | **49.5** | **59.3** | **57.5** | 51.7 | 53.5 | **55.5** | **67.7** | **58.0** | **53.1** | **60.0** | **14.7** | **59.4** | **65.1** |

Table 1: Evaluation on image concept segmentation with text. AP$_o$ corresponds to COCO-O accuracy, $^\star$partially trained on LVIS. $^\dagger$from original papers, $^\delta$from DINO-X API. Gray numbers indicate usage of respective closed set training data (LVIS/COCO). Upper and lower bound for human performance given, see §F.4 for details.

# Result

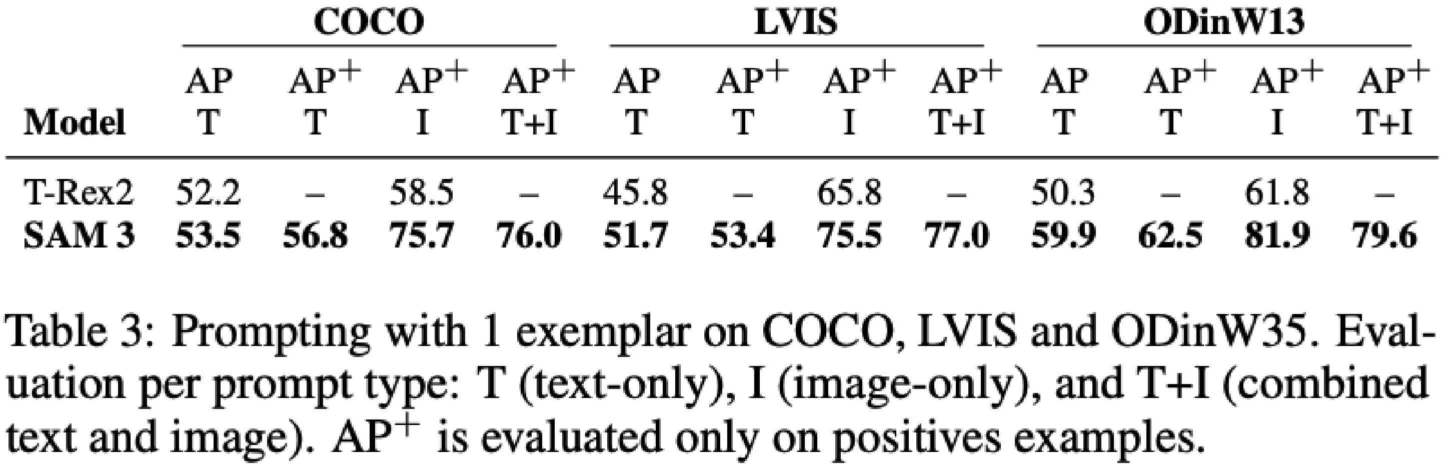
❖ **Few shot & exemplar performance**

- Measure the usefulness of the model in downstream applications



| Model | ODinW13 | | RF-100VL | |
|---|---|---|---|---|
| | $AP_0$ | $AP_{10}$ | $AP_0$ | $AP_{10}$ |
| Gemini2.5-Pro | 33.7 | – | 11.6 | 9.8 |
| gDino-T | 49.7 | – | **15.7** | 33.7 |
| gDino1.5-Pro | 58.7 | 67.9 | – | – |
| **SAM 3** | **59.9** | **71.6** | 14.3 | **35.7** |

Table 2: Zero-shot and 10-shot transfer on in-the-wild datasets.



| Model | COCO | | | | LVIS | | | | ODinW13 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AP<br>T | $AP^+$<br>T | $AP^+$<br>I | $AP^+$<br>T+I | AP<br>T | $AP^+$<br>T | $AP^+$<br>I | $AP^+$<br>T+I | AP<br>T | $AP^+$<br>T | $AP^+$<br>I | $AP^+$<br>T+I |
| T-Rex2 | 52.2 | – | 58.5 | – | 45.8 | – | 65.8 | – | 50.3 | – | 61.8 | – |
| **SAM 3** | **53.5** | **56.8** | **75.7** | **76.0** | **51.7** | **53.4** | **75.5** | **77.0** | **59.9** | **62.5** | **81.9** | **79.6** |

Table 3: Prompting with 1 exemplar on COCO, LVIS and ODinW35. Evaluation per prompt type: T (text-only), I (image-only), and T+I (combined text and image). $AP^+$ is evaluated only on positives examples.

$AP_0$ : $Zero\ shot\ performance$

$AP_{10}$: $few\ shot\ performance$

Single exemplar

# Result

❖ **SAM 3’s interactive exemplar prompts vs the ideal PVS baseline on SA-Co.**

- Orange : Ideal PVS performance(prompt, mask, point) → **Modify individual instance**
- Blue : 2 prompts PCS → PVS (hybrid)
- Red : 4 prompts PCS → PVS
- **Green : Only PCS**

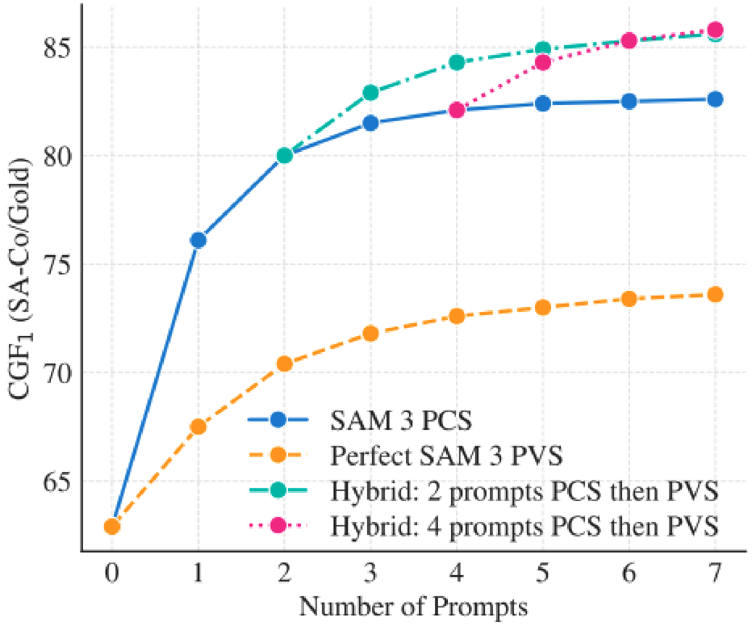
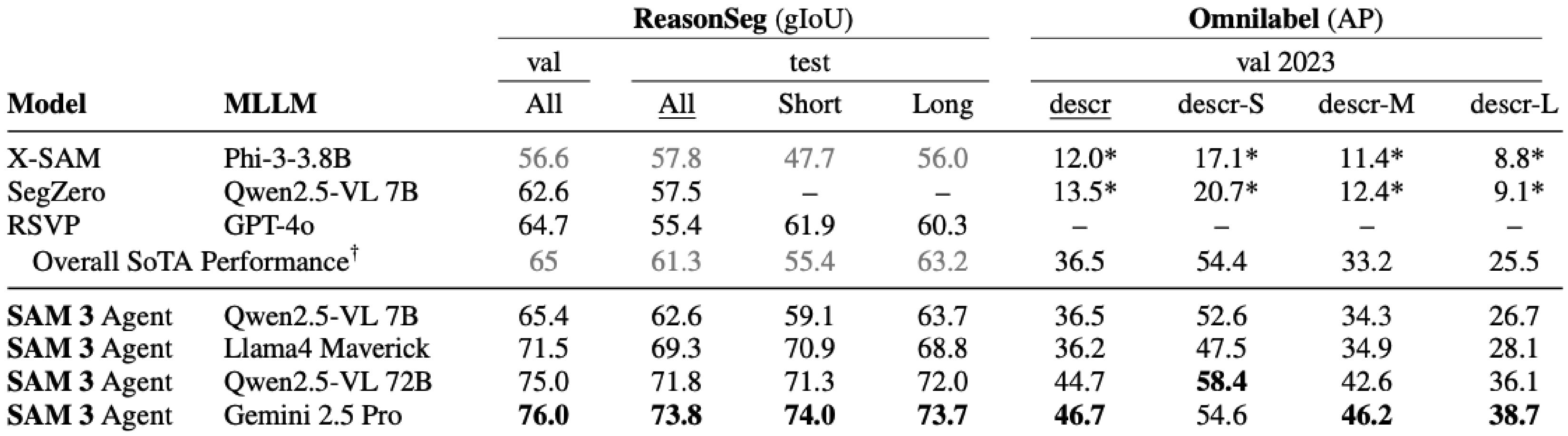


Figure 6: SAM 3’s interactive exemplar prompts vs the ideal PVS baseline on SA-Co. We report $CGF_1$ vs # of box prompts, averaged over all SA-Co/Gold phrases.

# Result

## ❖ SAM 3 Agent result

- Complex text queries process ability
- SAM 3 Agent refers to a multimodal large-scale language model (MLLM) that utilizes the SAM 3 model as a tool



| | | **ReasonSeg** (gIoU) | | | | **Omnilabel** (AP) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | val | test | | | val 2023 | | | |
| **Model** | **MLLM** | All | All | Short | Long | descr | descr-S | descr-M | descr-L |
| X-SAM | Phi-3-3.8B | 56.6 | 57.8 | 47.7 | 56.0 | 12.0* | 17.1* | 11.4* | 8.8* |
| SegZero | Qwen2.5-VL 7B | 62.6 | 57.5 | – | – | 13.5* | 20.7* | 12.4* | 9.1* |
| RSVP | GPT-4o | 64.7 | 55.4 | 61.9 | 60.3 | – | – | – | – |
| Overall SoTA Performance† | | 65 | 61.3 | 55.4 | 63.2 | 36.5 | 54.4 | 33.2 | 25.5 |
| **SAM 3** Agent | Qwen2.5-VL 7B | 65.4 | 62.6 | 59.1 | 63.7 | 36.5 | 52.6 | 34.3 | 26.7 |
| **SAM 3** Agent | Llama4 Maverick | 71.5 | 69.3 | 70.9 | 68.8 | 36.2 | 47.5 | 34.9 | 28.1 |
| **SAM 3** Agent | Qwen2.5-VL 72B | 75.0 | 71.8 | 71.3 | 72.0 | 44.7 | **58.4** | 42.6 | 36.1 |
| **SAM 3** Agent | Gemini 2.5 Pro | **76.0** | **73.8** | **74.0** | **73.7** | **46.7** | 54.6 | **46.2** | **38.7** |

Table 8: SAM 3 Agent results. Gray indicates fine-tuned results on ReasonSeg (train), * indicates reproduced results, underline indicates the main metric. †: LISA-13B-LLaVA1.5 for ReasonSeg; REAL for OmniLabel.

# Conclusion

### ❖ Concept based interactive segmentation

- Introducing the PCS task and SA-Co benchmark
- Doubling performance over prior systems for PCS on SA-Co in images and videos

### ❖ Limitations

- Detailed classification and special domains
- Language Complexity (Beyond Simple Noun Phrases)
- Scalability in Video : linearly increase computation cost / Real-time (30 FPS) processing requires huge resources
- 8 x H200 GPUs are required to track 64 objects simultaneously
- Interaction Mode Switching