SAM3: Segment Anything with Concepts

DMQA Open Seminar (26. 05. 01)

Data Mining & Quality Analytics Lab.

김현이

발표자 소개

김현이(Hyeoni Kim)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석박사통합과정 (2026.03~)

Research Interest

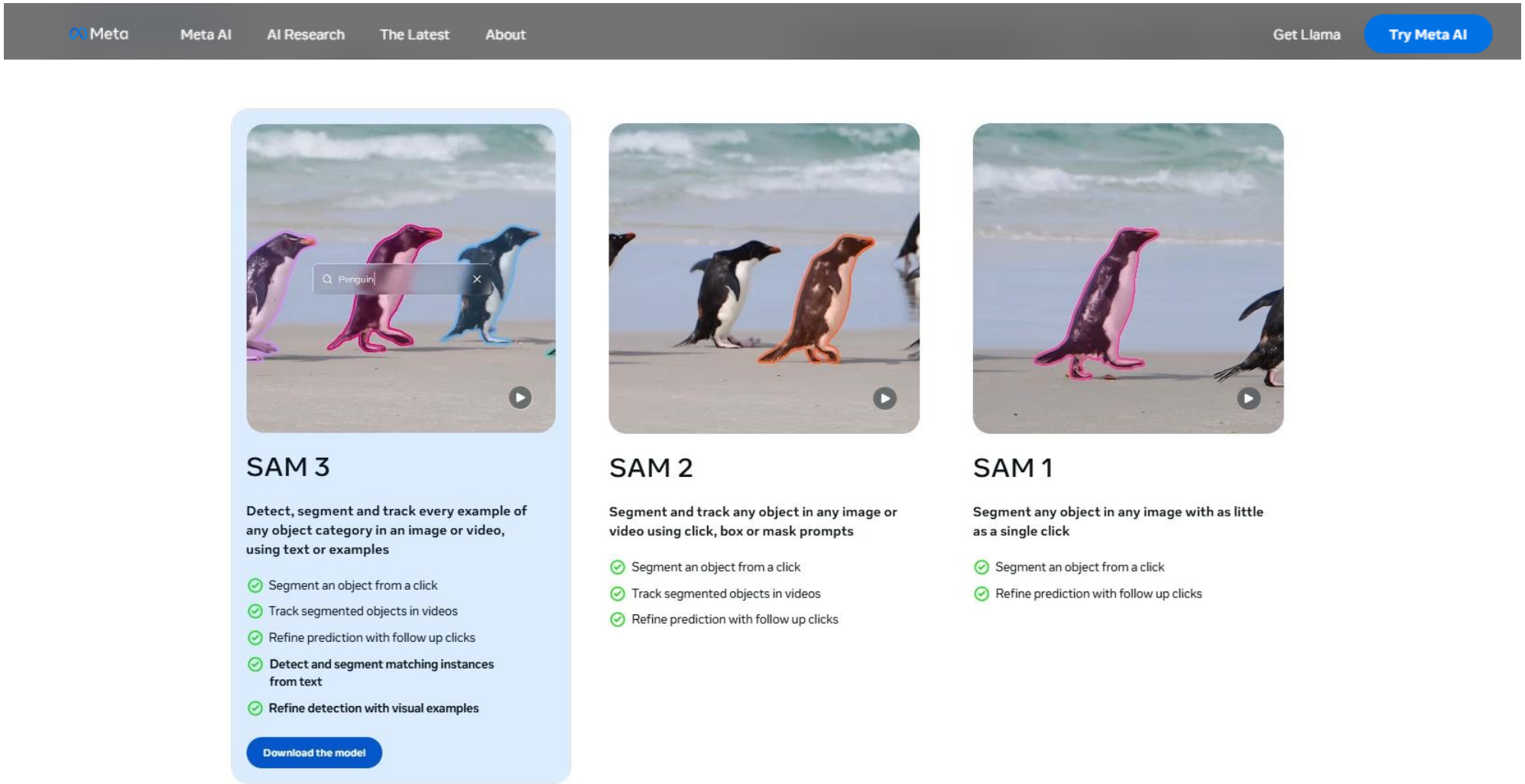
- Computer Vision
- Vision Language Model
- Test-time adaptation

Contact

- hyeon2k@korea.ac.kr

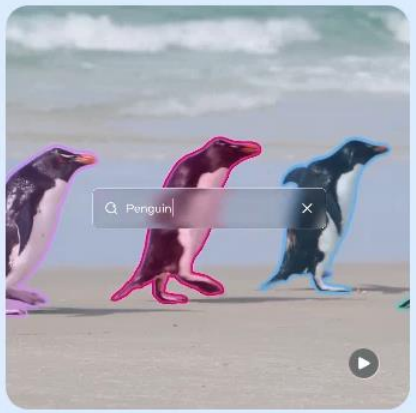
Introduction

SAM3: Segment Anything with Concepts



Introduction

SAM3: Segment Anything with Concepts



SAM 3

Detect, segment and track every example of any object category in an image or video, using text or examples

- Segment an object from a click
- Track segmented objects in videos
- Refine prediction with follow up clicks
- **Detect and segment matching instances from text**
- **Refine detection with visual examples**

Download the model

SAM 2

Segment and track any object in any image or video using click, box or mask prompts

- Segment an object from a click
- Track segmented objects in videos
- Refine prediction with follow up clicks

SAM 1

Segment any object in any image with as little as a single click

- Segment an object from a click
- Refine prediction with follow up clicks

종료

Segment Anything and its Adapter

2023. 12. 08.
조용원
Data Mining and Quality Analytics Lab

Segment Anything and its Adapter

발표자:  조용원

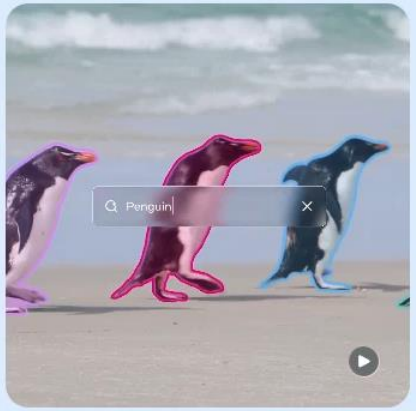
2023년 12월 8일
오전 12시 ~
고려대학교 신공학관 218호
온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Introduction

SAM3: Segment Anything with Concepts

Meta | Meta AI | AI Research | The Latest | About | Get Llama | Try Meta AI



SAM 3

Detect, segment and track every example of any object category in an image or video, using text or examples

- Segment an object from a click
- Track segmented objects in videos
- Refine prediction with follow up clicks
- **Detect and segment matching instances from text**
- **Refine detection with visual examples**

Download the model

SAM 2

Segment and track any object in any image or video using click, box or mask prompts

- Segment an object from a click
- Track segmented objects in videos
- Refine prediction with follow up clicks

SAM 1

Segment any object in any image with as little as a single click

- Segment an object from a click
- Refine prediction with follow up clicks

Introduction

What is segmentation?

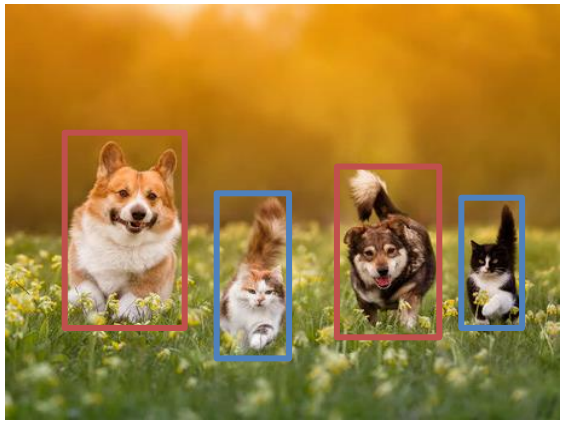
❖ Image segmentation

- 객체의 영역을 픽셀 단위로 나누는 Task

Classification
(분류)

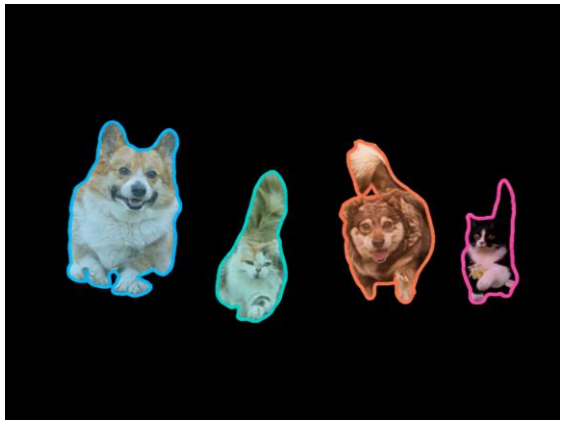
Dog

Object detection
(위치)



Dog, Cat

Segmentation
(영역)



Dog, Cat, Dog, Cat

Introduction

What is segmentation?

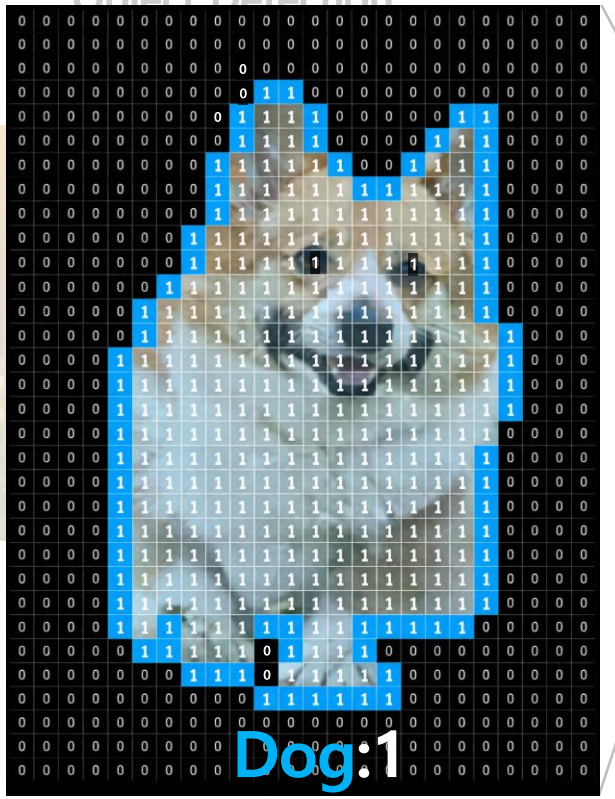
❖ Image segmentation

- 객체의 영역을 픽셀 단위로 나누는 Task

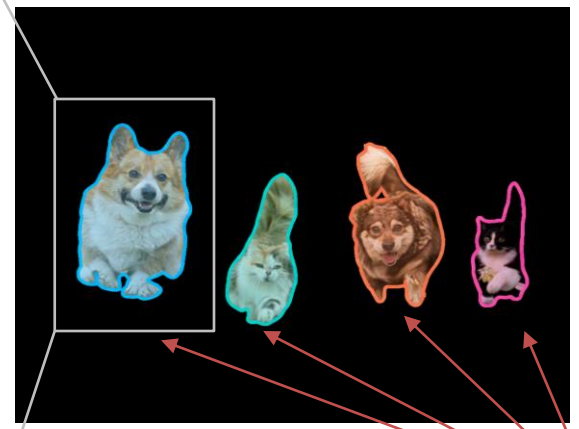
Classification
(분류)

Dog

Object Detection



Segmentation
(영역)



Dog, Cat, Dog, Cat

mask

Introduction

What is segmentation?

❖ Image segmentation의 종류

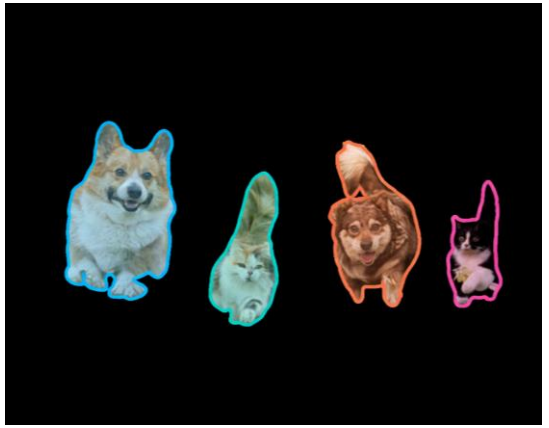
- Semantic segmentation: 픽셀 단위로 클래스만 구분
- Instance segmentation: 객체를 개별적으로 구분
- Panoptic segmentation: 객체 + 배경 전체를 모두 분할

Image

Semantic segmentation

Dog, Cat, Grass, Sky

Instance segmentation



Dog1, Cat1, Dog2, Cat2

Panoptic segmentation

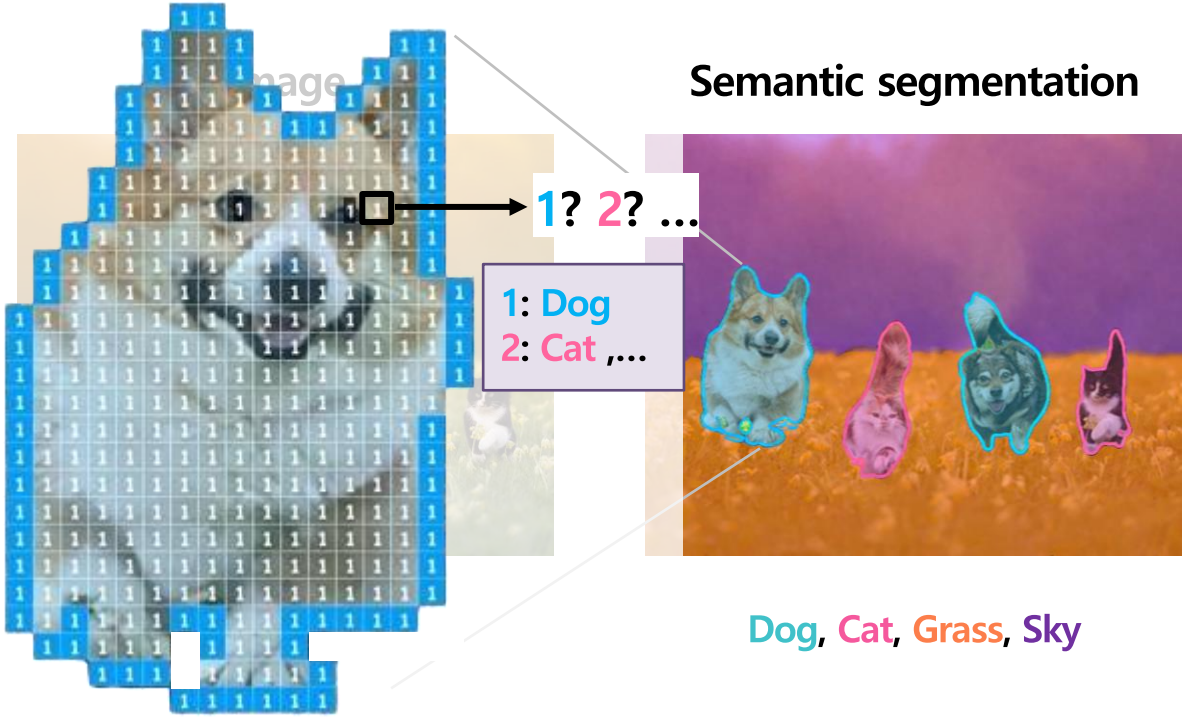
Dog1, Cat1, Dog2, Cat2, Grass , Sky

Introduction

What is segmentation?

❖ Image segmentation의 종류

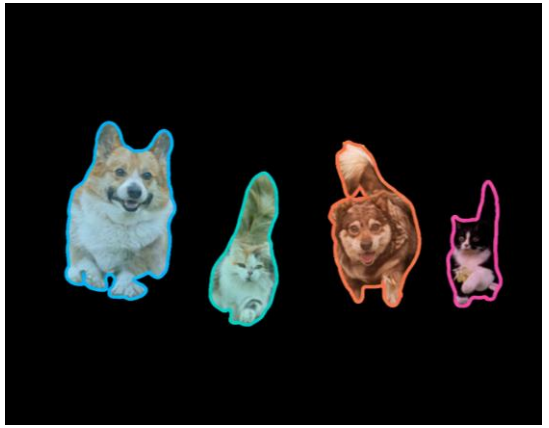
- **Semantic segmentation:** 픽셀 단위로 클래스만 구분
- **Instance segmentation:** 객체를 개별적으로 구분
- Panoptic segmentation: 객체 + 배경 전체를 모두 분할



Semantic segmentation

Dog, Cat, Grass, Sky

Instance segmentation



Dog1, Cat1, Dog2, Cat2

Panoptic segmentation

Dog1, Cat1, Dog2, Cat2, Grass , Sky

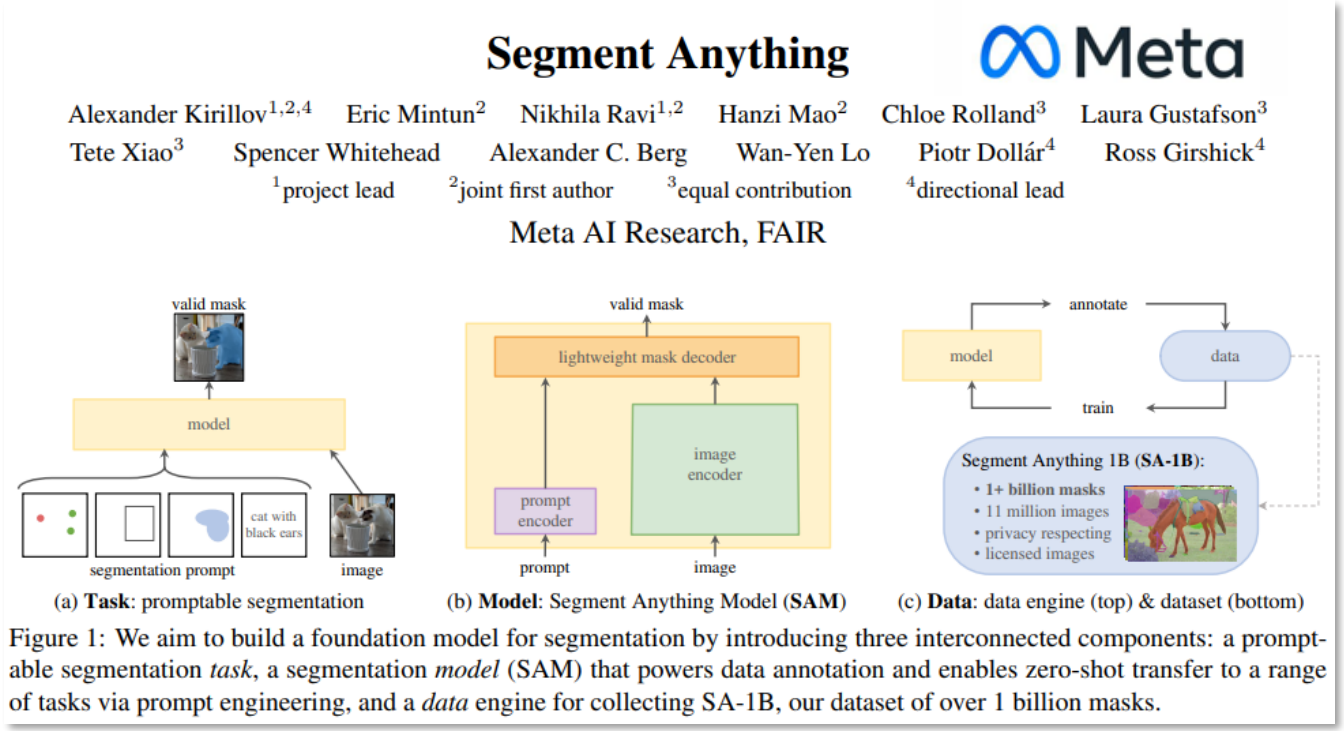
https://share.google/O0CiX1a7yr9HqLqLd

Introduction

SAM1

❖ SAM1: Segment Anything (ICCV 2023)

- **Meta**에서 개발한 segmentation foundation model
  - 대규모 데이터로 학습 → 다양한 이미지에 적용 가능
  - 학습하지 않은 이미지에서도 동작 (generalization)



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a prompt-able segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2023, pp. 4015–4026.

Introduction

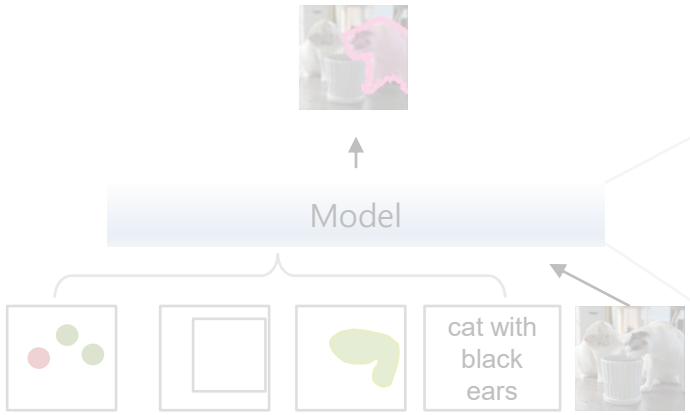
SAM1

❖ Foundation model

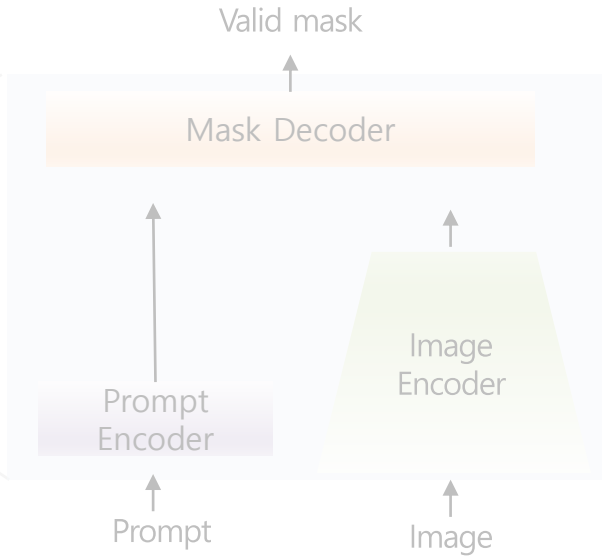
- **How is this possible?:** Data Engine
  - 약 **1,100만 장의** 이미지와 **11억 개 mask**

Segment Anything(**SA-1B**)

Data: data engine & dataset



Task: promptable segmentation



Model: Segment Anything Model

Introduction

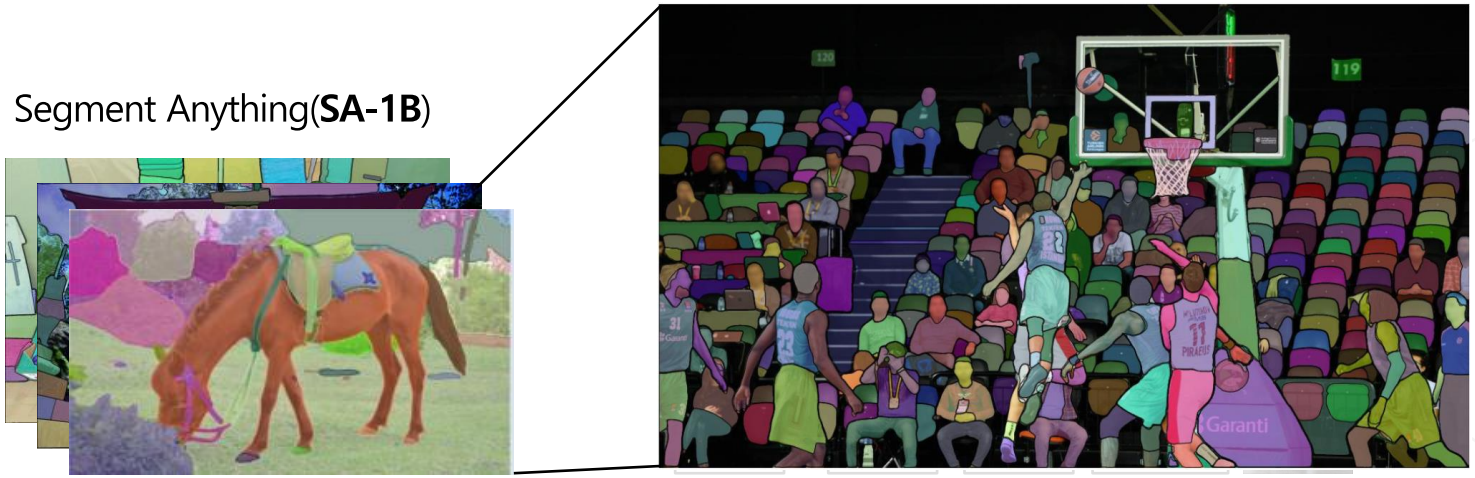
SAM1

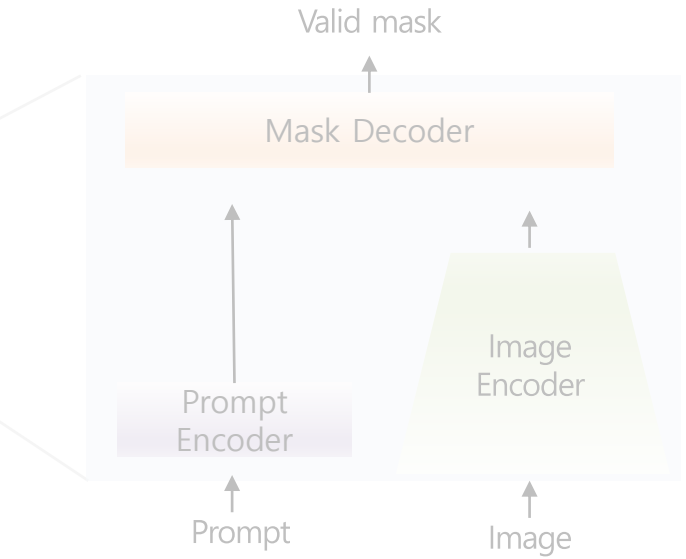
❖ Foundation model

- **How is this possible?:** Data Engine
  - 약 **1,100만 장의** 이미지와 **11억 개 mask**

어떻게 11억개의 mask를 만들었을까?

Segment Anything(**SA-1B**)





Data: data engine & dataset

Task: promptable segmentation

Model: Segment Anything Model

Introduction

SAM1

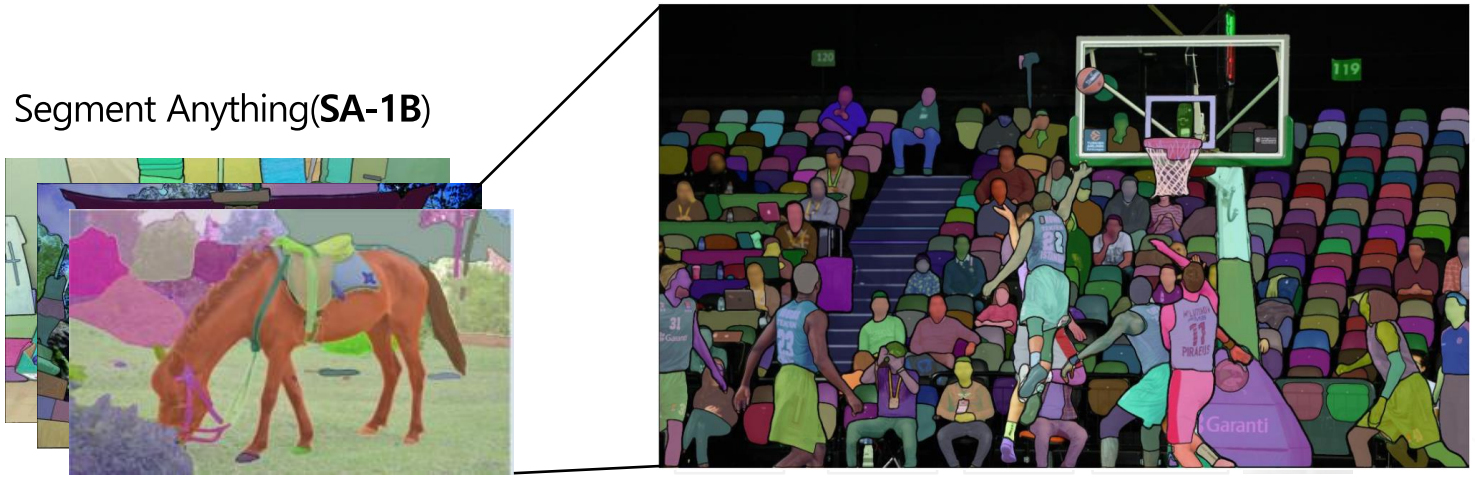
❖ Foundation model

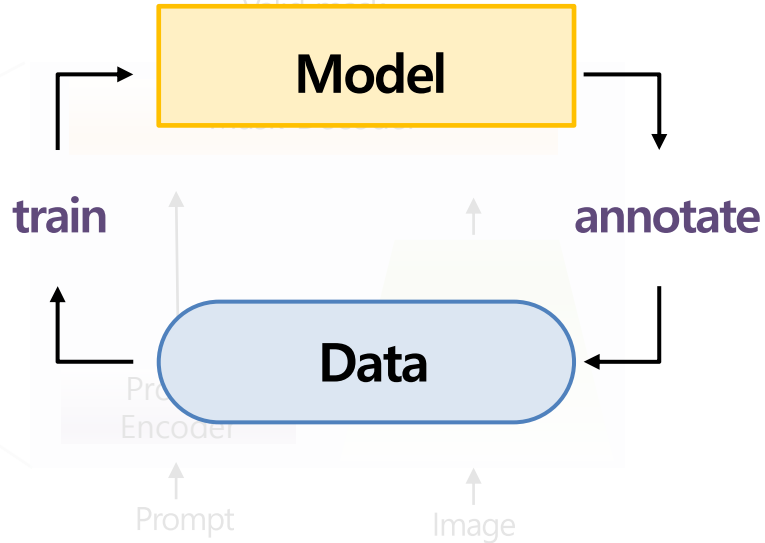
- **How is this possible?:** Data Engine
  - 약 **1,100만 장의** 이미지와 **11억 개 mask**
  - 고해상도 이미지: 라이선스 기반 확보
  - mask 생성: 모델이 생성 + 일부만 사람 검증 (99% 자동 생성)

어떻게 11억개의 mask를 만들었을까?

⇒ 사람 + 모델 협업 기반 Data Engine

Segment Anything(**SA-1B**)





Data: data engine & dataset

Task: promptable segmentation

Model: Segment Anything Model

Introduction

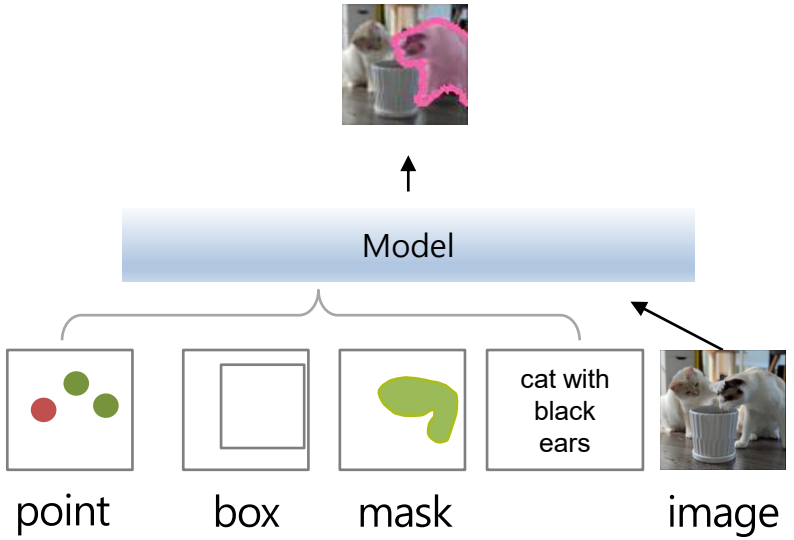
SAM1

❖ Promptable segmentation

- 사용자 입력 (point, box, mask)
  - 원하는 객체 지정 →해당 영역 segmentation

Segment Anything(**SA-1B**)

Data: data engine (top) & dataset(bottom)



Task: promptable segmentation

Model: Segment Anything Model

Introduction

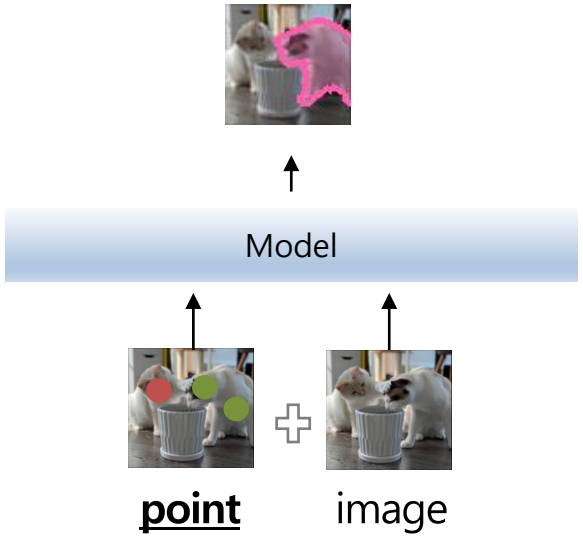
SAM1

❖ Promptable segmentation

- 사용자 입력 (point, box, mask)
  - 원하는 객체 지정 →해당 영역 segmentation

Segment Anything(**SA-1B**)

Data: data engine (top) & dataset(bottom)



Task: promptable segmentation

Model: Segment Anything Model

Introduction

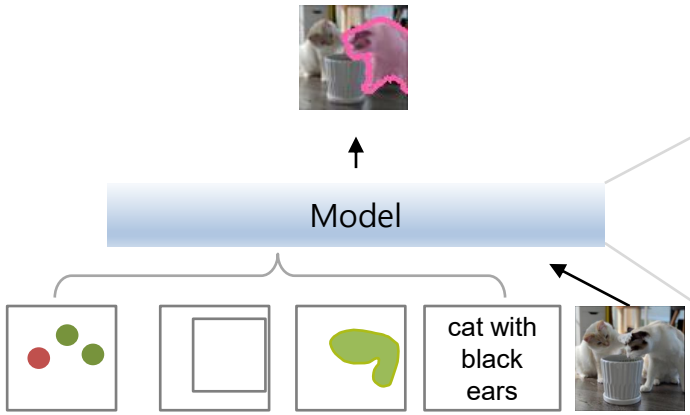
SAM1

❖ SAM1: Key components

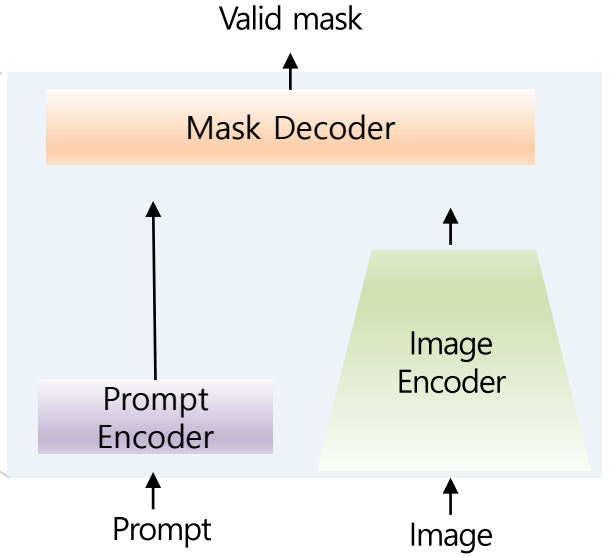
- **Data:** Large-scale dataset (data engine)
- **Task:** Promptable segmentation
- **Model:** Encoder–decoder architecture

Segment Anything(**SA-1B**)

Data: data engine & dataset



Task: promptable segmentation



Model: Segment Anything Model

Introduction

SAM1

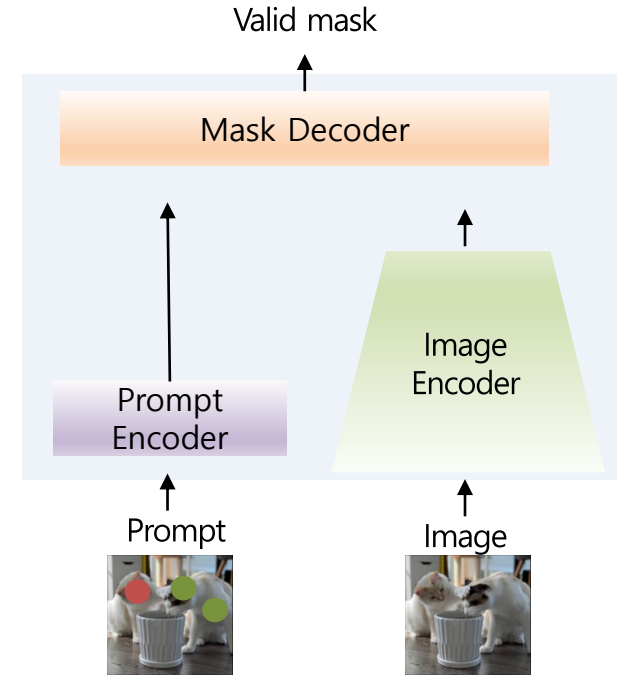
❖ SAM1: Key components

- **Data:** Large-scale dataset (data engine)
- **Task:** Promptable segmentation
- **Model:** Encoder–decoder architecture

Segment Anything(**SA-1B**)

Data: data engine

Task: promptable segmentation



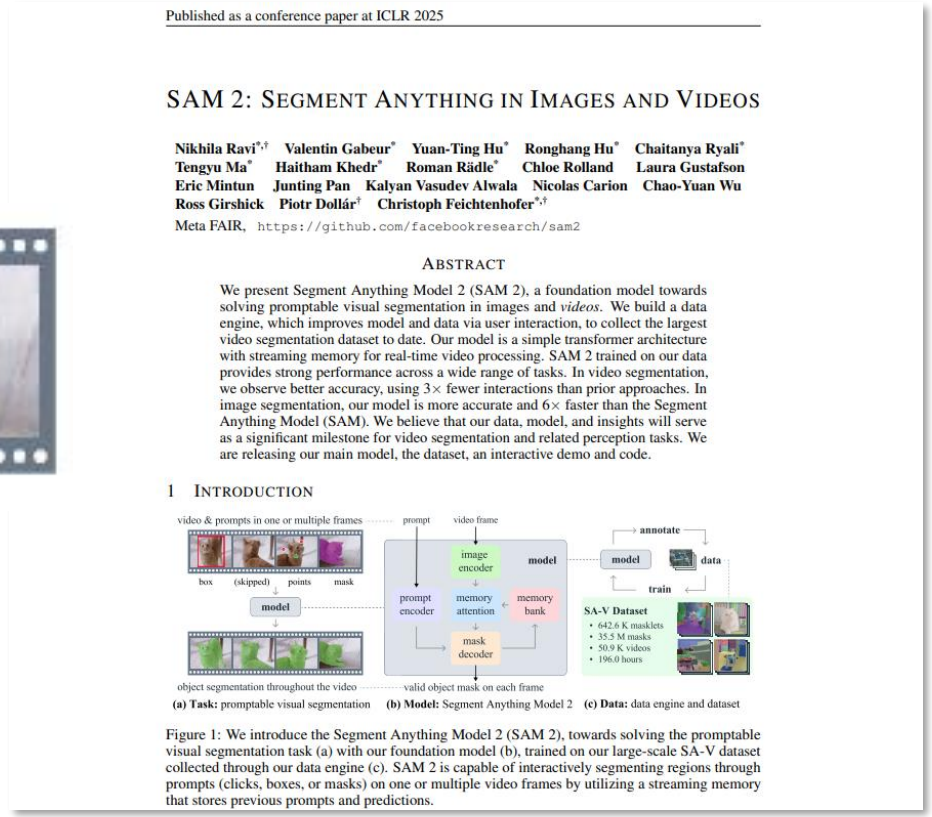
Model: Segment Anything Model

Introduction

SAM2

❖ SAM2: Segment Anything in Images and Videos(ICLR 2025)

- Image → **video** 확장
- 객체를 계속 추적하며 segmentation
- SAM1보다 6배 빠름



Published as a conference paper at ICLR 2025

SAM 2: SEGMENT ANYTHING IN IMAGES AND VIDEOS

Nikhila Ravi*,† Valentin Gabeur* Yuan-Ting Hu* Ronghang Hu* Chaitanya Ryali*
Tengyu Ma* Haitham Khedr* Roman Rädle* Chloe Rolland Laura Gustafson
Eric Mintun Junting Pan Kalyan Vasudev Alwala Nicolas Carion Chao-Yuan Wu
Ross Girshick Piotr Dollár† Christoph Feichtenhofer*,†
Meta FAIR, https://github.com/facebookresearch/sam2

ABSTRACT

We present Segment Anything Model 2 (SAM 2), a foundation model towards solving promptable visual segmentation in images and *videos*. We build a data engine, which improves model and data via user interaction, to collect the largest video segmentation dataset to date. Our model is a simple transformer architecture with streaming memory for real-time video processing. SAM 2 trained on our data provides strong performance across a wide range of tasks. In video segmentation, we observe better accuracy, using 3× fewer interactions than prior approaches. In image segmentation, our model is more accurate and 6× faster than the Segment Anything Model (SAM). We believe that our data, model, and insights will serve as a significant milestone for video segmentation and related perception tasks. We are releasing our main model, the dataset, an interactive demo and code.

1 INTRODUCTION

Figure 1: We introduce the Segment Anything Model 2 (SAM 2), towards solving the promptable visual segmentation task (a) with our foundation model (b), trained on our large-scale SA-V dataset collected through our data engine (c). SAM 2 is capable of interactively segmenting regions through prompts (clicks, boxes, or masks) on one or multiple video frames by utilizing a streaming memory that stores previous prompts and predictions.

[2] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment Anything in Images and Videos," in Proc. Int. Conf. Learn. Represent. (ICLR), 2025.

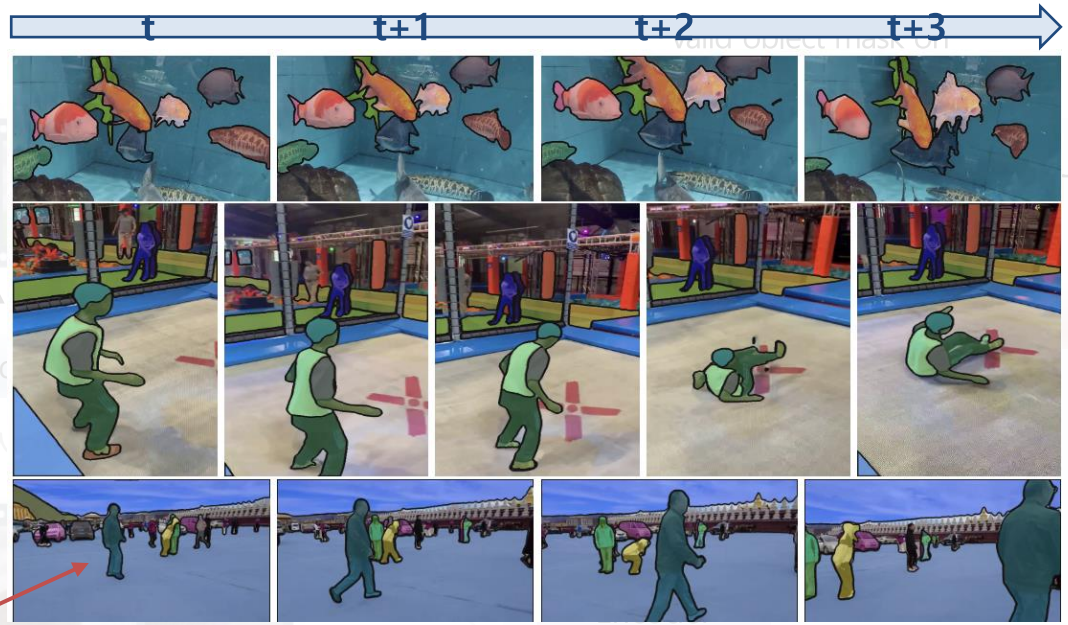
Introduction

SAM2

❖ SAM2: Key components

- **Data:** 이미지 + 비디오 데이터셋 (약 5만개의 비디오와 64만개의 masklet)
- **Task:** 프롬프트 기반 객체 추적
- **Model:** Memory attention + memory bank

SA-V Dataset



Masklet

: 시공간적인 mask

Data: data engine & dataset

Task: promptable segmentation

Model: Segment Anything Model

Introduction

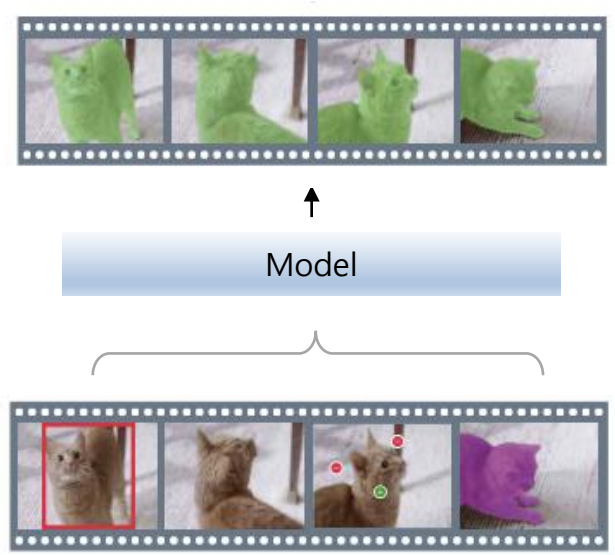
SAM2

❖ SAM2: Key components

- **Data:** 이미지 + 비디오 데이터셋
- **Task:** 프롬프트 기반 객체 추적
- **Model:** Memory attention + memory bank

SA-V Dataset

Data: data engine & dataset



Task: promptable segmentation

Model: Segment Anything Model

Introduction

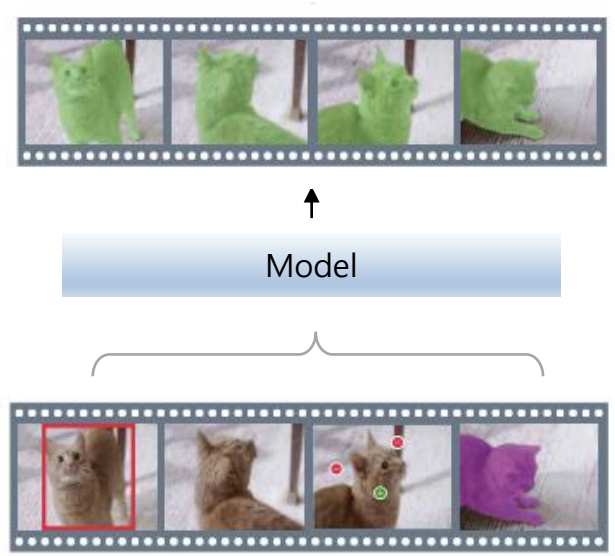
SAM2

❖ SAM2: Key components

- **Data:** 이미지 + 비디오 데이터셋
- **Task:** 프롬프트 기반 객체 추적
- **Model:** Memory attention + memory bank

어떻게 video를 segmentation할까?

Data: data engine & dataset



Task: promptable segmentation

Model: Segment Anything Model

Introduction

SAM2

❖ SAM2: Key components

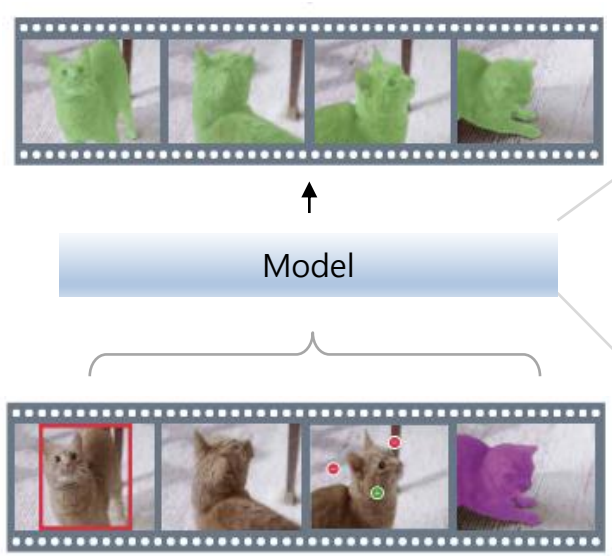
- **Data:** 이미지 + 비디오 데이터셋
- **Task:** 프롬프트 기반 객체 추적
- **Model:** Memory attention + memory bank

어떻게 video를 segmentation할까?

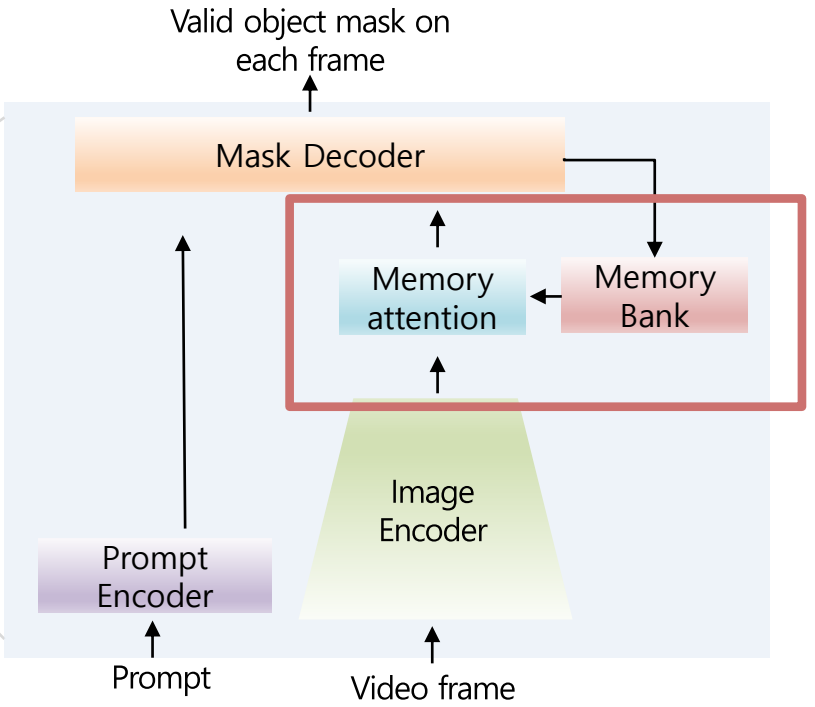
⇒ 이전 프레임의 정보를 기억하고 활용

SA-V Dataset

Data: data engine & dataset



Task: promptable segmentation



Model: Segment Anything Model 2

Introduction

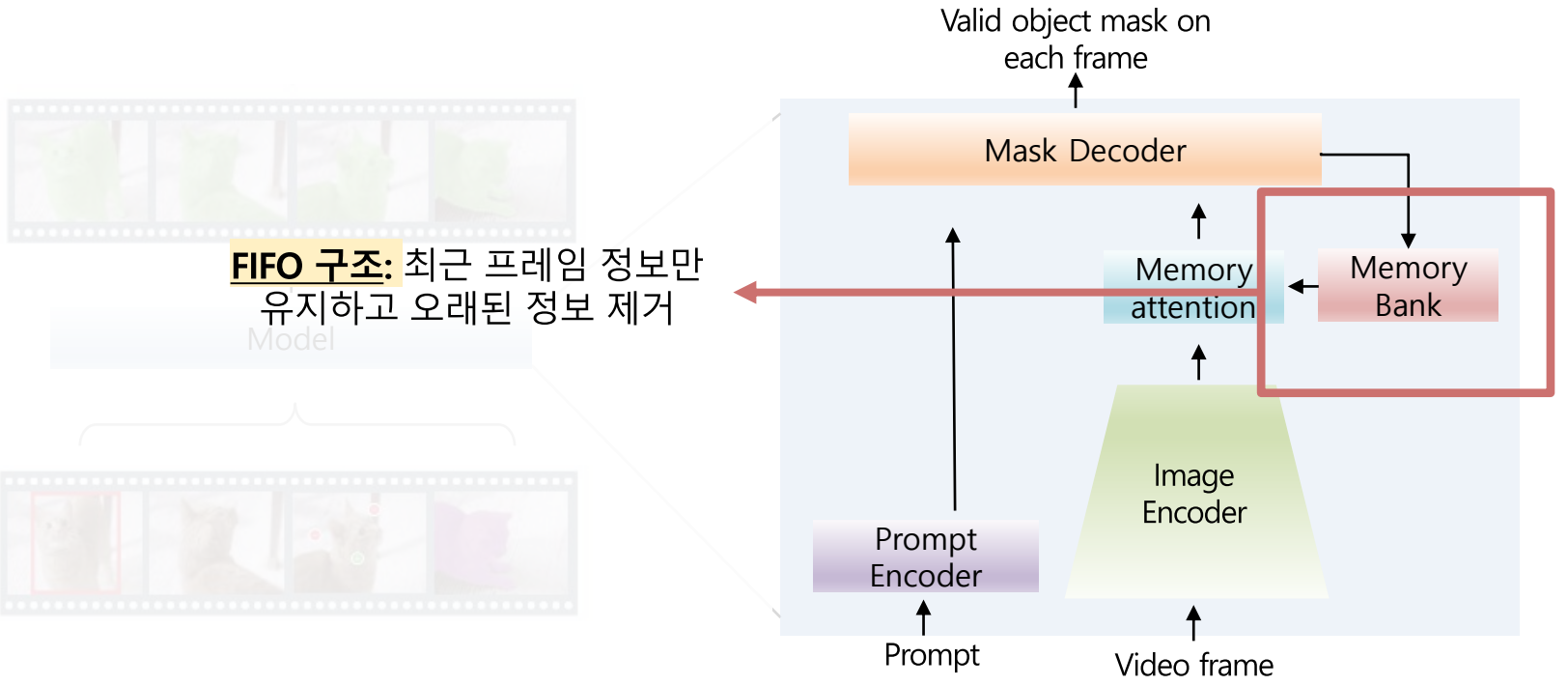
SAM2

❖ SAM2: Key components

- **Data:** 이미지 + 비디오 데이터셋
- **Task:** 프롬프트 기반 객체 추적
- **Model:** Memory attention + memory bank

어떻게 video를 segmentation할까?

⇒ 이전 프레임의 정보를 기억하고 활용



FIFO 구조: 최근 프레임 정보만 유지하고 오래된 정보 제거

Model: Segment Anything Model 2

Introduction

SAM2

❖ SAM2: Key components

- **Data:** 이미지 + 비디오 데이터셋
- **Task:** 프롬프트 기반 객체 추적
- **Model:** Memory attention + memory bank

어떻게 video를 segmentation할까?

⇒ 이전 프레임의 정보를 기억하고 활용

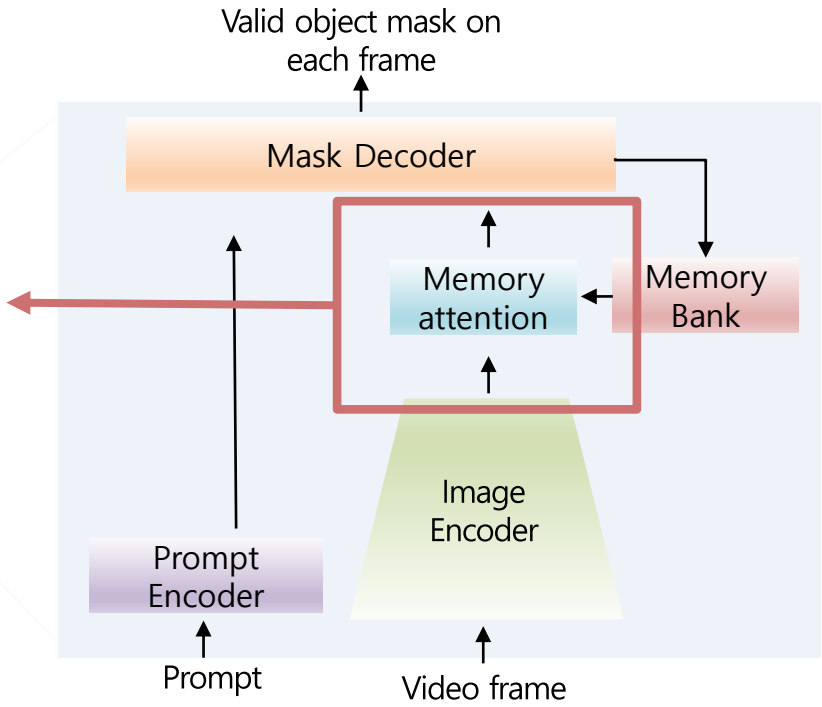
SA-V Dataset

Data: data engine & dataset

현재 프레임의 특징과 **이전 프레임의 정보**를 비교면서 동일한 객체를 찾음

①**Self-attention:** 어떤 픽셀들이 서로 관련이 있을까?
② **Cross-attention:** 과거 정보들과 비교

Task: promptable segmentation



Model: Segment Anything Model 2

Introduction

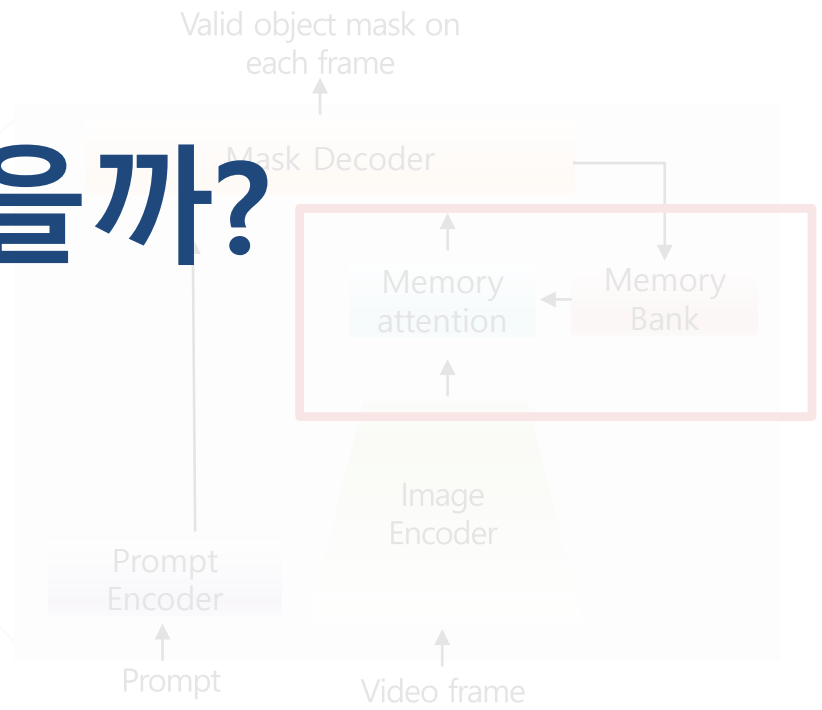
SAM2

❖ **SAM2: Key components**

- **Data:** 이미지 + 비디오 데이터셋
- **Task:** 프롬프트 기반 객체 추적
- **Model:** Memory attention + memory bank

왜 SAM3가 등장했을까?

SA-V Dataset



Data: data engine & dataset

Task: promptable segmentation

Model: Segment Anything Model

Introduction

Why SAM3?

❖ Limitation of SAM1 & SAM2

- 정확한 segmentation 성능 but **위치 힌트**(점, 박스 등)에 의존
  - "**무엇(개념)**"을 찾아야 하는지는 모름

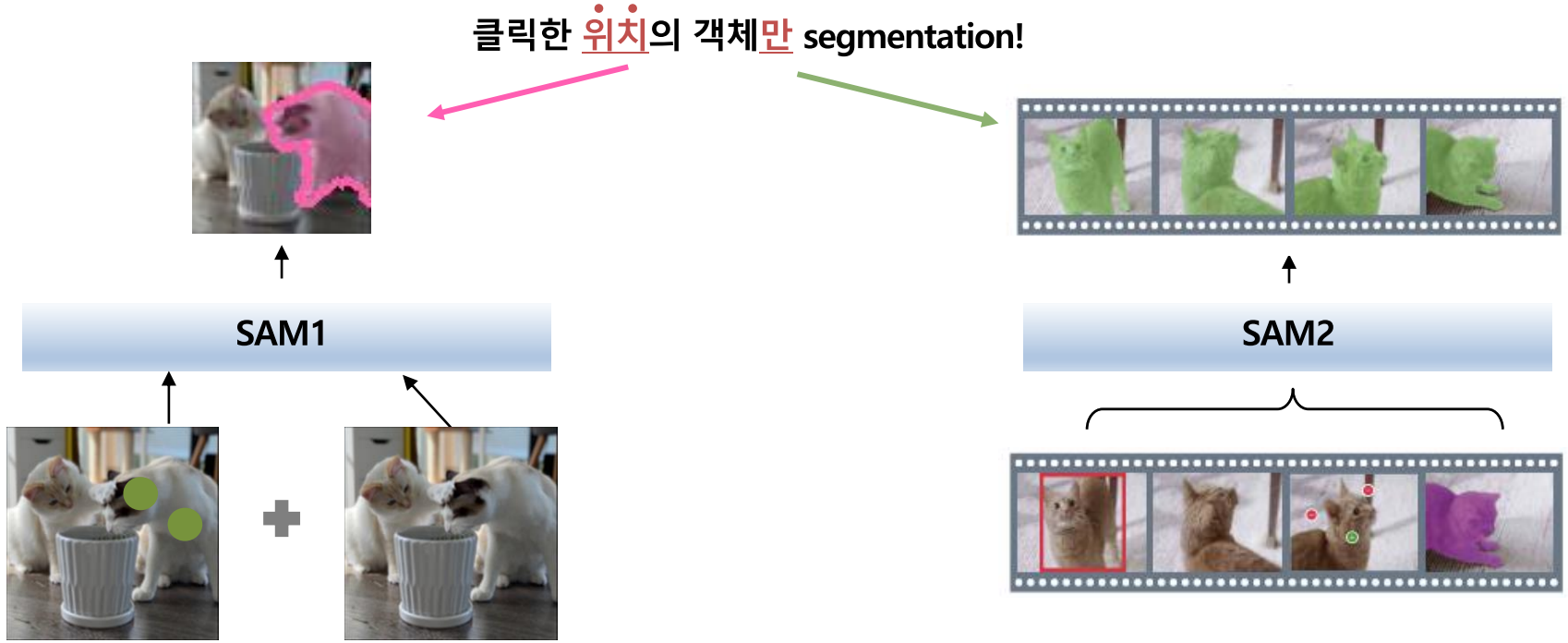
Introduction

Why SAM3?

❖ Limitation of SAM1 & SAM2

- 정확한 segmentation 성능 but 위치 한정(점, 박스 형)에 의존
  - "무엇(개념)을 찾아야 하는지는 모름"

위치를 지정하지 않고, 개념만으로 객체를 찾을 수는 없을까?

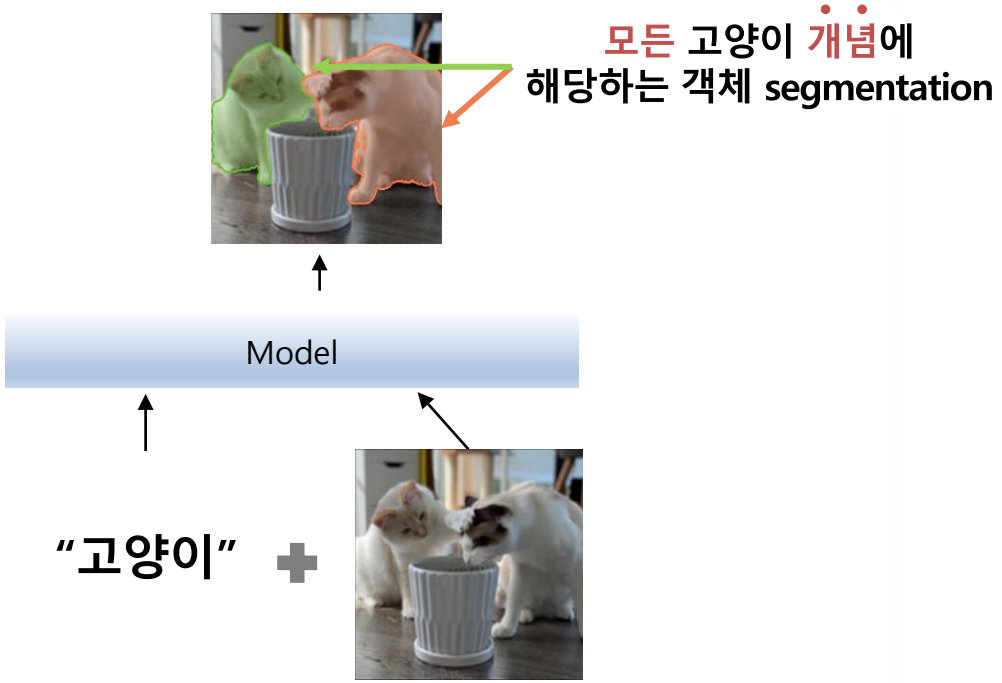


Introduction

SAM3

❖ SAM 3: Segment Anything with Concepts (ICLR 2026)

- **텍스트, 이미지 exemplar** 기반 개념 입력 (“고양이”, 펭귄 예시사진)
- 장면 전체에서 해당 개념에 해당하는 모든 객체 탐색 ⇒ **PCS**



Meta

SAM 3: Segment Anything with *Concepts*

Promptable

Concept

Segmentation

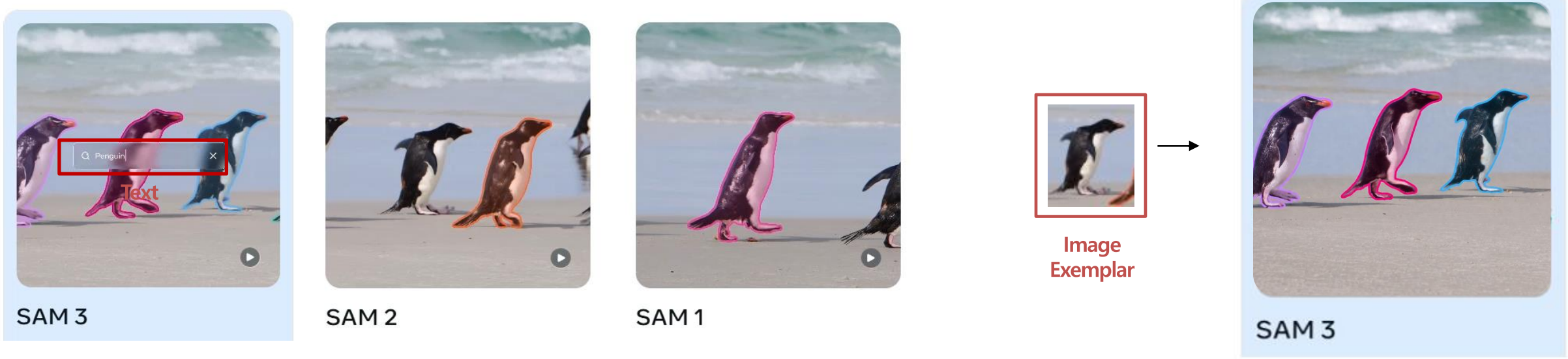
[3] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer, “SAM 3: Segment Anything with Concepts,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2026.

Introduction

SAM3

❖ SAM 3: Segment Anything with Concepts (ICLR 2026)

- **텍스트, 이미지 exemplar** 기반 개념 입력 (“고양이”, 펭귄 예시사진)
- 장면 전체에서 해당 개념에 해당하는 모든 객체 탐색 ⇒ **PCS**

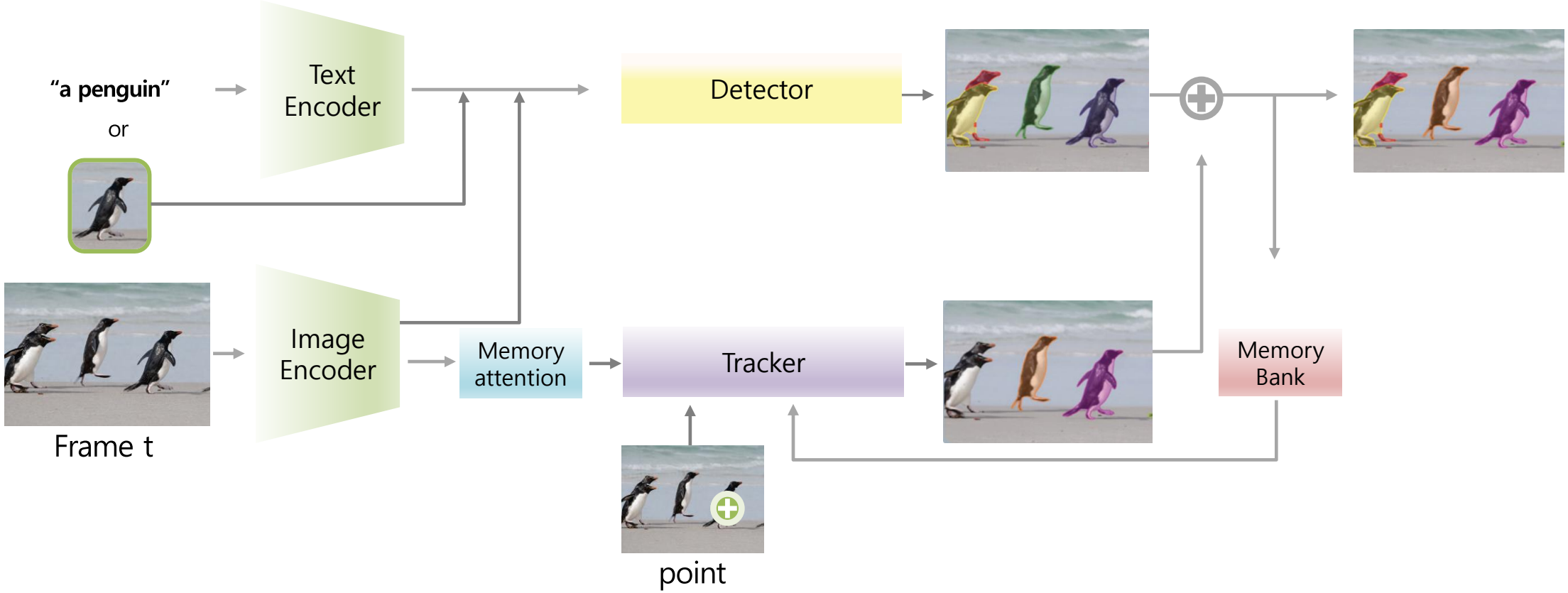


Method

SAM3

❖ Promptable Concept Segmentation이 어떻게 가능할까?

- **찾기(find) → 분할하기(segment) → 추적하기(track)**
  - Detector: 개념 기반 객체 위치 탐색
  - Tracker: 시간 축에서 객체 지속 추적

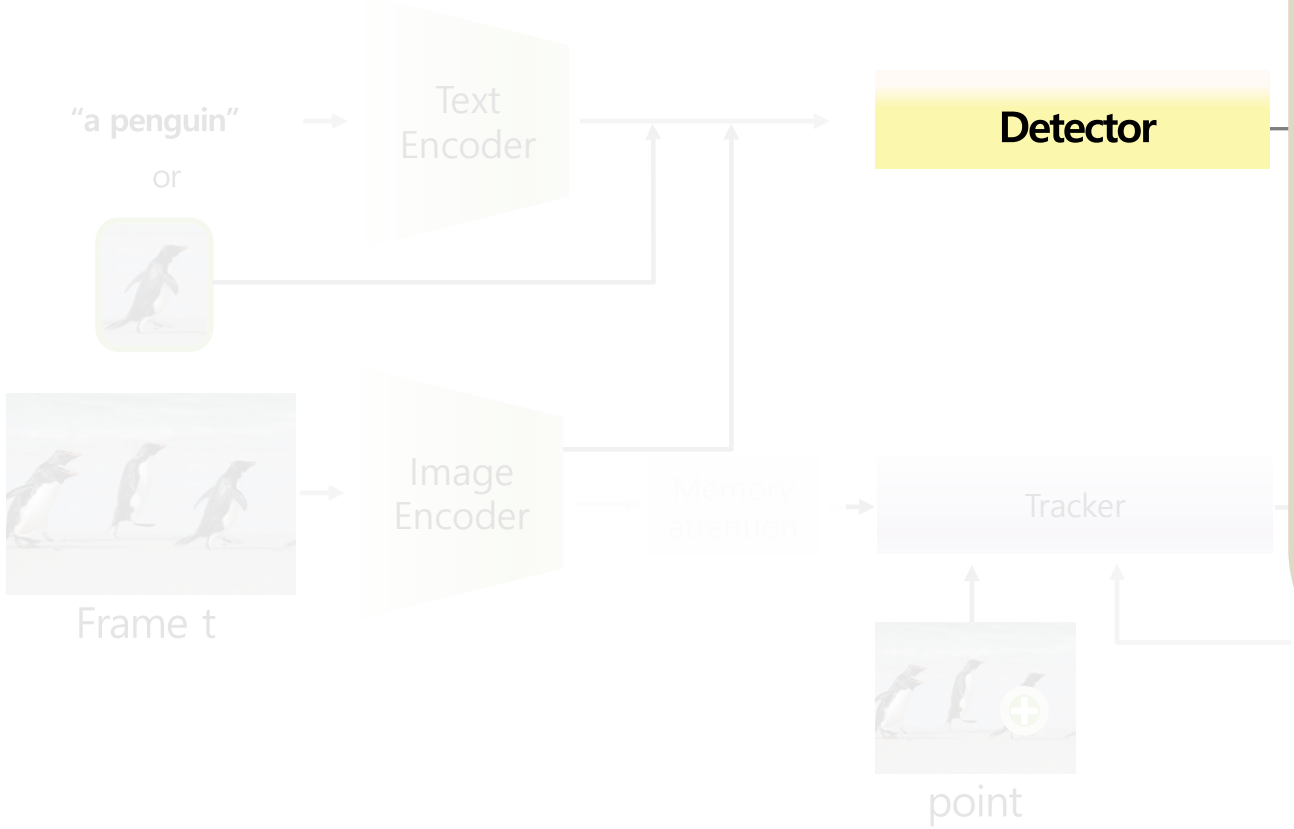


Method

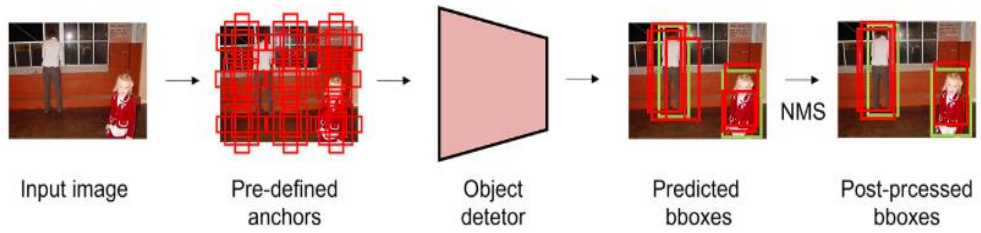
SAM3

❖ Promptable Concept Segmentation이 어떻게 가능할까?

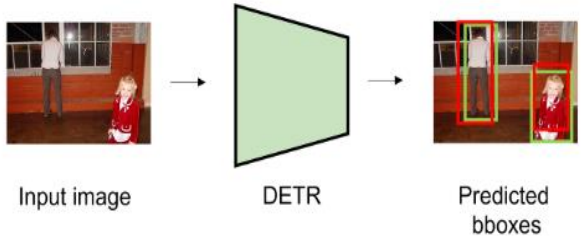
- **찾기(find)** → 분할하기(segment) → 추적하기(track)
  - Detector: **개념 기반 객체 위치 탐색**
  - Tracker: 시간 축에서 객체 지속 추적



DETR



[기존 Object detector의 detection 과정]



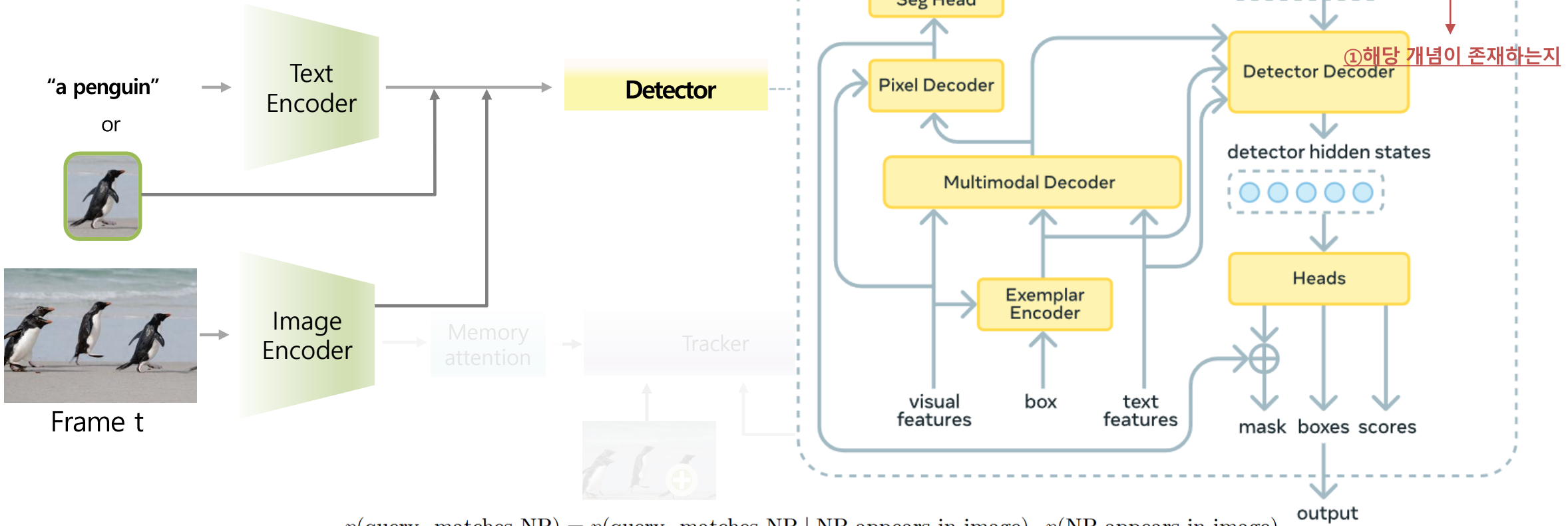
[DETR의 detection 과정]

Method

SAM3

❖ Promptable Concept Segmentation이 어떻게 가능할까?

- **찾기(find)** → 분할하기(segment) → 추적하기(track)
  - Detector: **개념 기반 객체 위치 탐색**
  - Tracker: 시간 축에서 객체 지속 추적



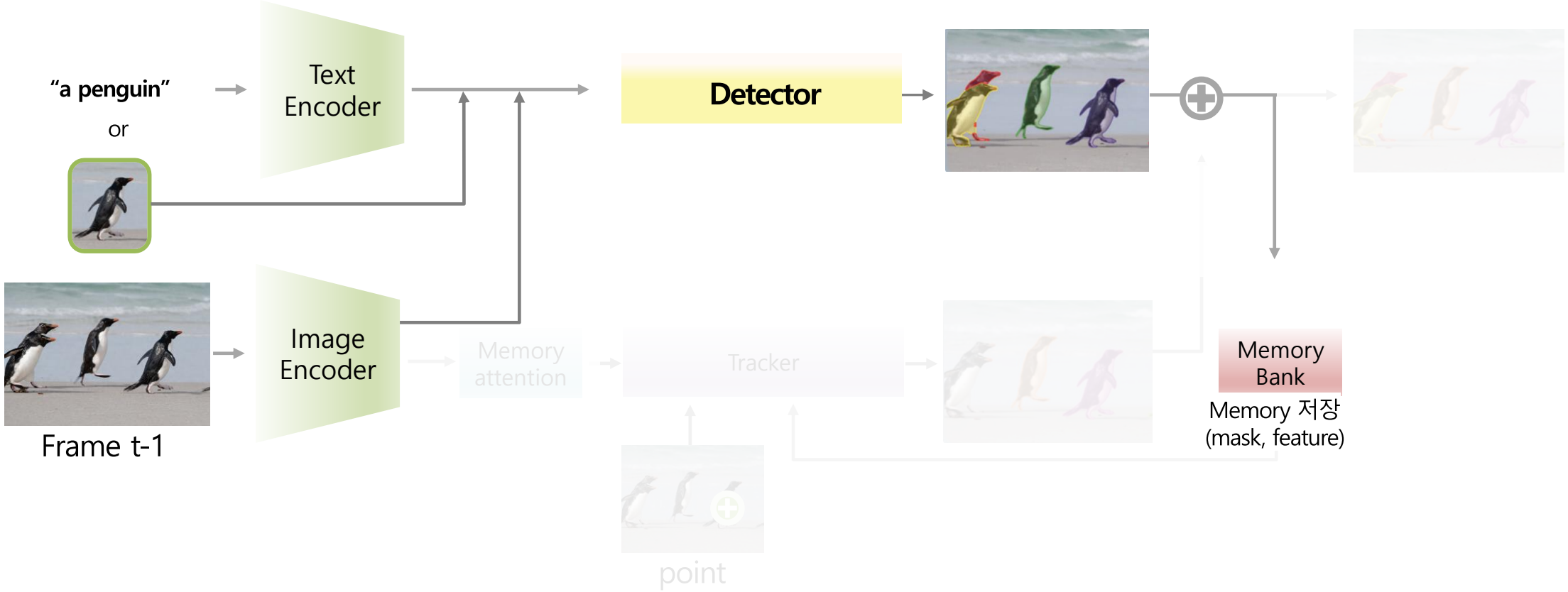
$$p(\text{query}_i \text{ matches NP}) = p(\text{query}_i \text{ matches NP} \mid \text{NP appears in image}) \cdot p(\text{NP appears in image}).$$

Method

SAM3

❖ Promptable Concept Segmentation이 어떻게 가능할까?

- **찾기(find)** → 분할하기(segment) → 추적하기(track)
  - Detector: **개념 기반 객체 위치 탐색**
  - Tracker: 시간 축에서 객체 지속 추적

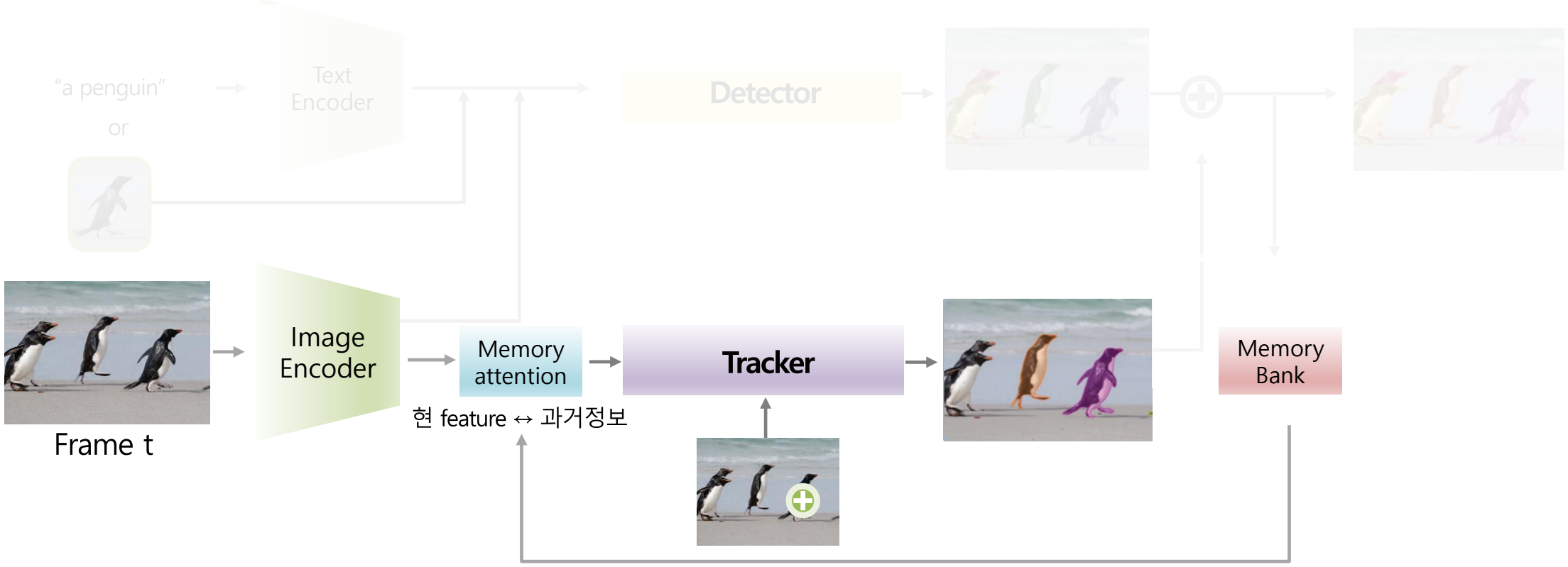


Method

SAM3

❖ Promptable Concept Segmentation이 어떻게 가능할까?

- 찾기(find) → 분할하기(segment) → **추적하기(track)**
  - Detector: 개념 기반 객체 위치 탐색
  - **Tracker: 시간 축에서 객체 지속 추적**

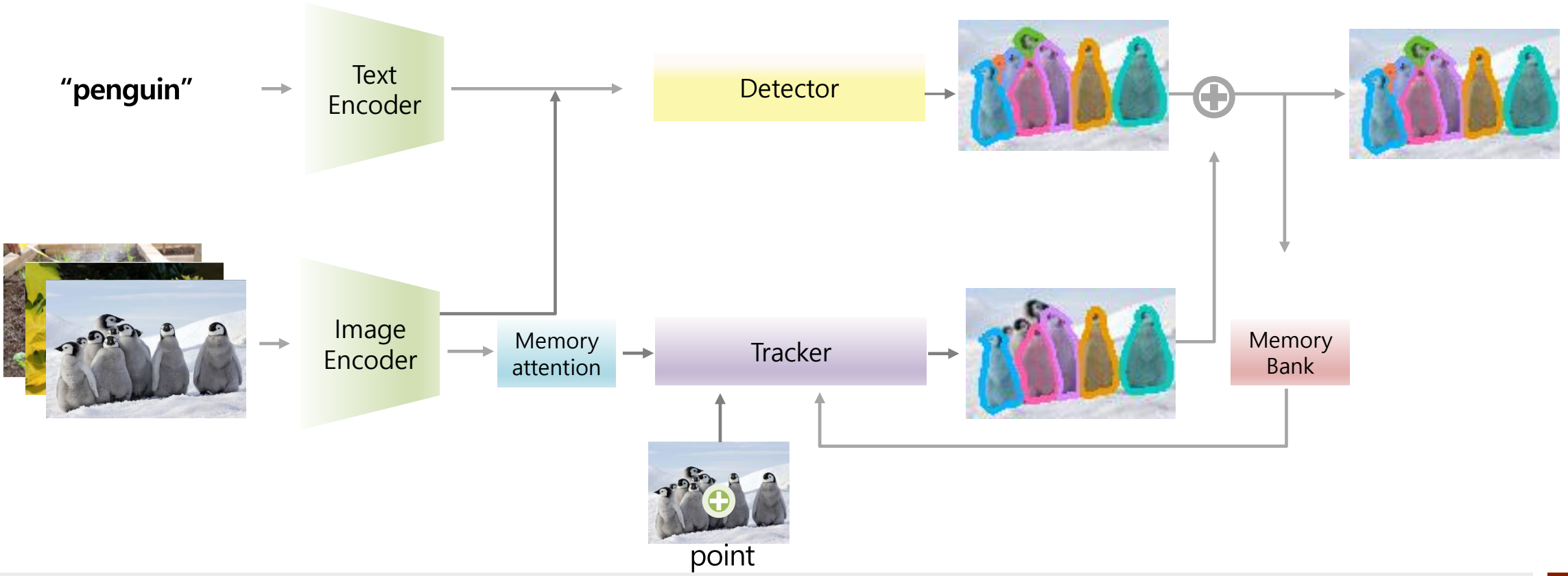


Appendix

SAM3

❖ SAM3는 어떻게 학습될까?

- 기능을 단계적으로 학습하는 구조
  - ① Concept 이해 → ② Object Localization → ③ Segmentation → ④ Tracking (Video)

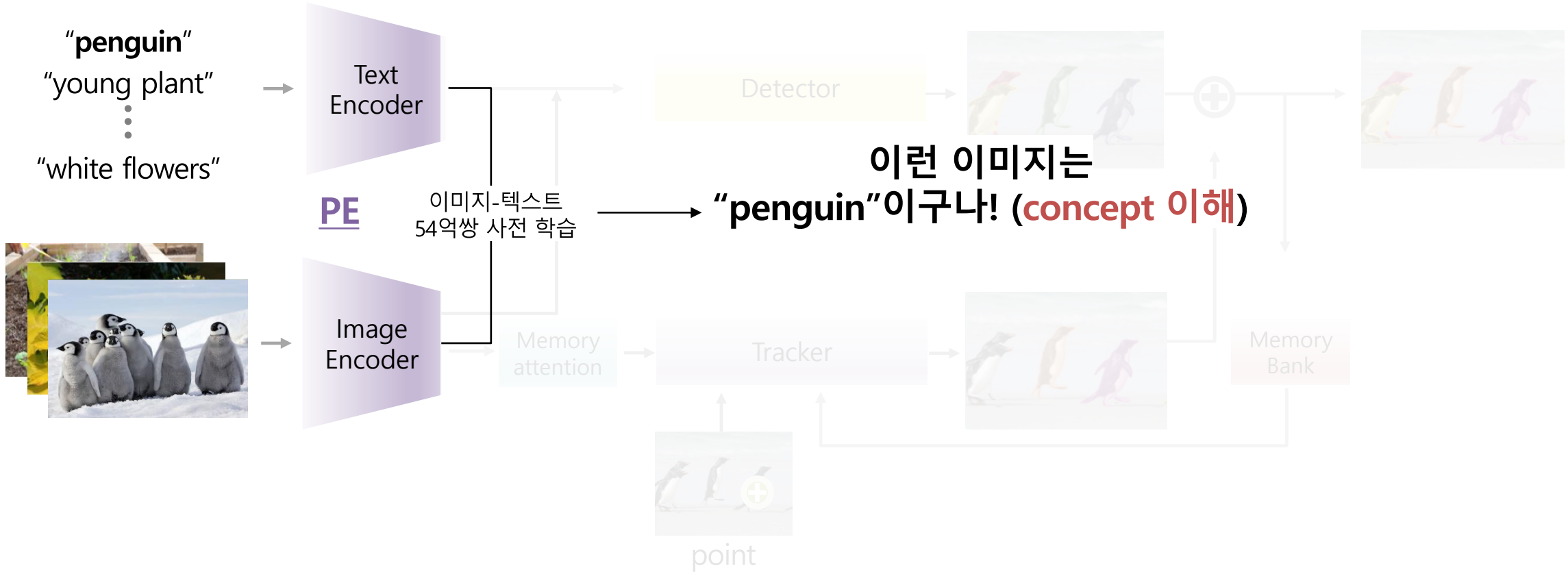


Appendix

SAM3

❖ SAM3는 어떻게 학습될까?

- 기능을 단계적으로 학습하는 구조
  - ① Concept 이해(Perception encoder pre-training) → ② Object Localization → ③ Segmentation → ④ Tracking (Video)



Appendix

SAM3

❖ **Perception encoder**(NeurIPS 2025)

- Meta에서 2025년에 발표한 large-scale vision encoder models (SOTA)
- 대규모 이미지/비디오 데이터를 입력 받아 contrastive vision-language learning을 수행
- 모든 downstream task에 대한 strong, general embedding 생성 가능

Perception Encoder: The best visual embeddings are not at the output of the network

Daniel Bolya[1,*] Po-Yao Huang[1,*] Peize Sun[1,*] Jang Hyun Cho[1,2,*,†] Andrea Madotto[1,*] Chen Wei[1] Tengyu Ma[1] Jiale Zhi[1] Jathushan Rajasegaran[1] Hanoona Rasheed[3,†] Junke Wang[4,†] Marco Monteiro[1] Hu Xu[1] Shiyu Dong[1] Nikhila Ravi[1] Daniel Li[1] Piotr Dollár[1] Christoph Feichtenhofer[1]
[1]Meta [2]UT Austin [3]MBZUAI [4]Fudan University
[*]Joint first author [†]Work done during internships at Meta

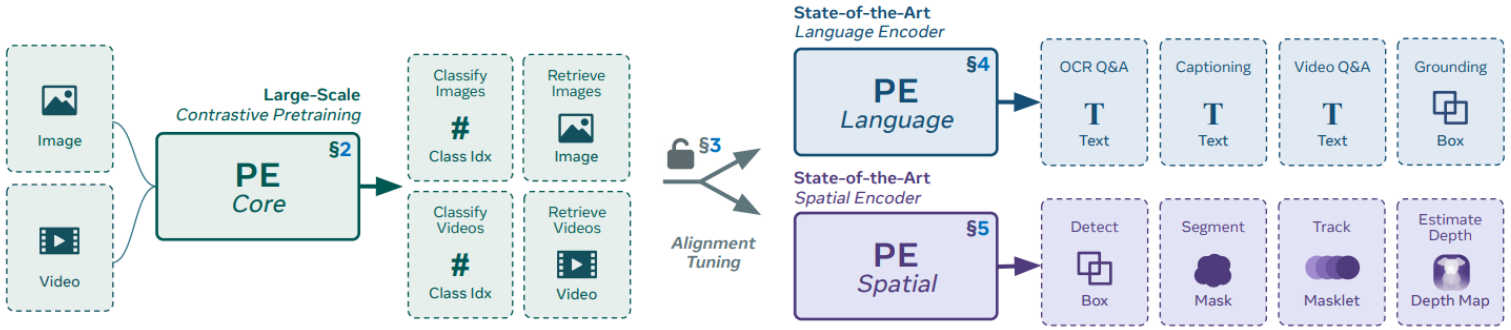


Figure 1 Perception Encoder (PE) is a family of large-scale vision encoder models with state-of-the-art performance on a large variety of vision tasks. By using a robust contrastive pretraining recipe and finetuning on synthetically aligned videos, PE not only outperforms all existing models on classification and retrieval (§2), but it also internally produces strong, general features that *scale* for downstream tasks (§3). PE unlocks the ability for large-scale contrastive pretraining to transfer to downstream tasks with alignment tuning to capitalize on those general features (§4, §5).

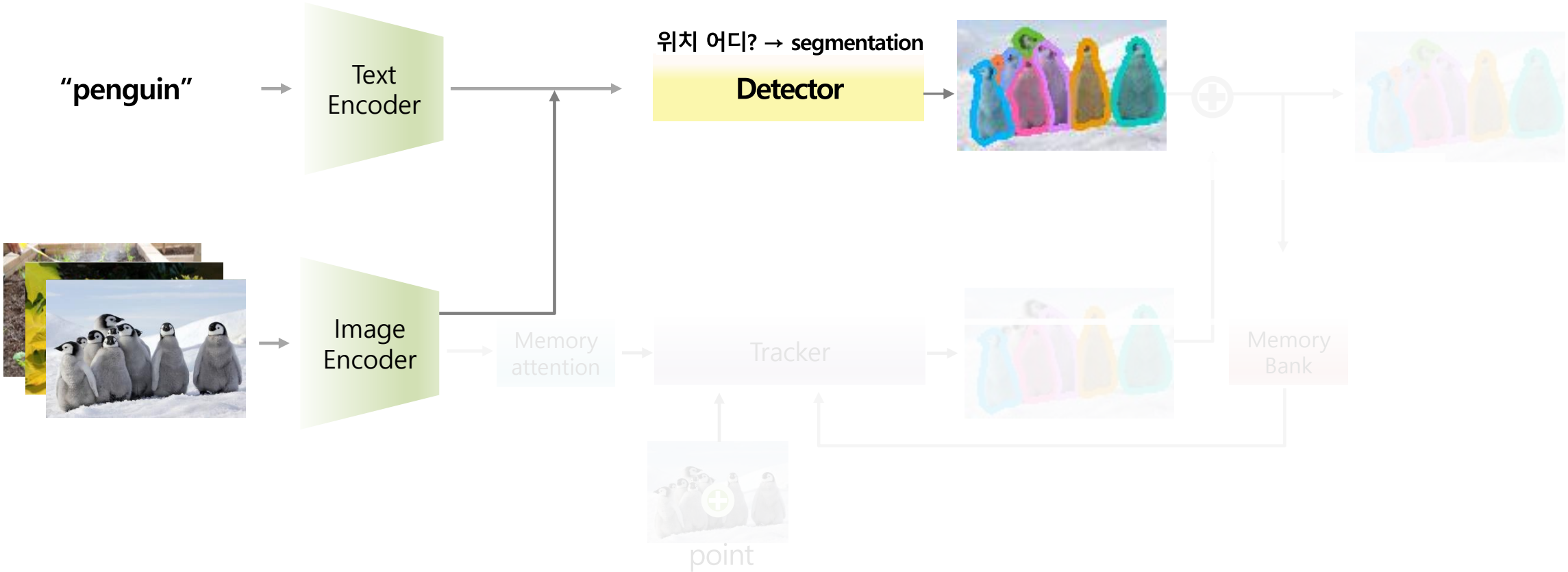
D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. A. Rasheed, J. Wang, M. Monteiro, H. Xu, S. Dong, N. Ravi, S.-W. Li, P. Dollár, and C. Feichtenhofer, "Perception Encoder: The best visual embeddings are not at the output of the network," in *Proc. 39th Annu. Conf. Neural Information Processing Systems (NeurIPS)*, 2025. [Online]. Available: https://openreview.net/forum?id=INqBOmwIpG

Appendix

SAM3

❖ SAM3는 어떻게 학습될까?

- 기능을 단계적으로 학습하는 구조
  - ① Concept 이해 → ② Concept-based Object Localization & Segmentation → ③ Refinement & Interaction → ④ Tracking (Video)

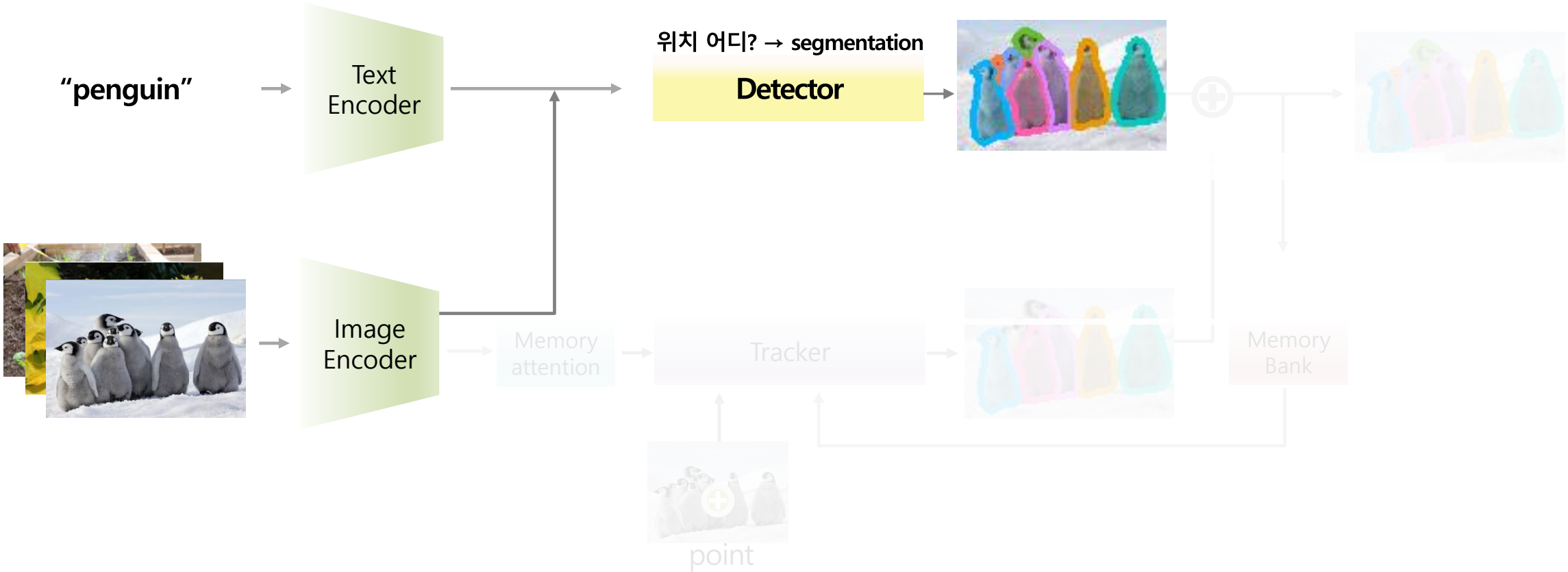


Appendix

SAM3

❖ SAM3는 어떻게 학습될까?

- 기능을 단계적으로 학습하는 구조
  - ① Concept 이해 → ② Concept-based Object Localization & Segmentation→ **③ Refinement & Interaction**→ ④ Tracking (Video)

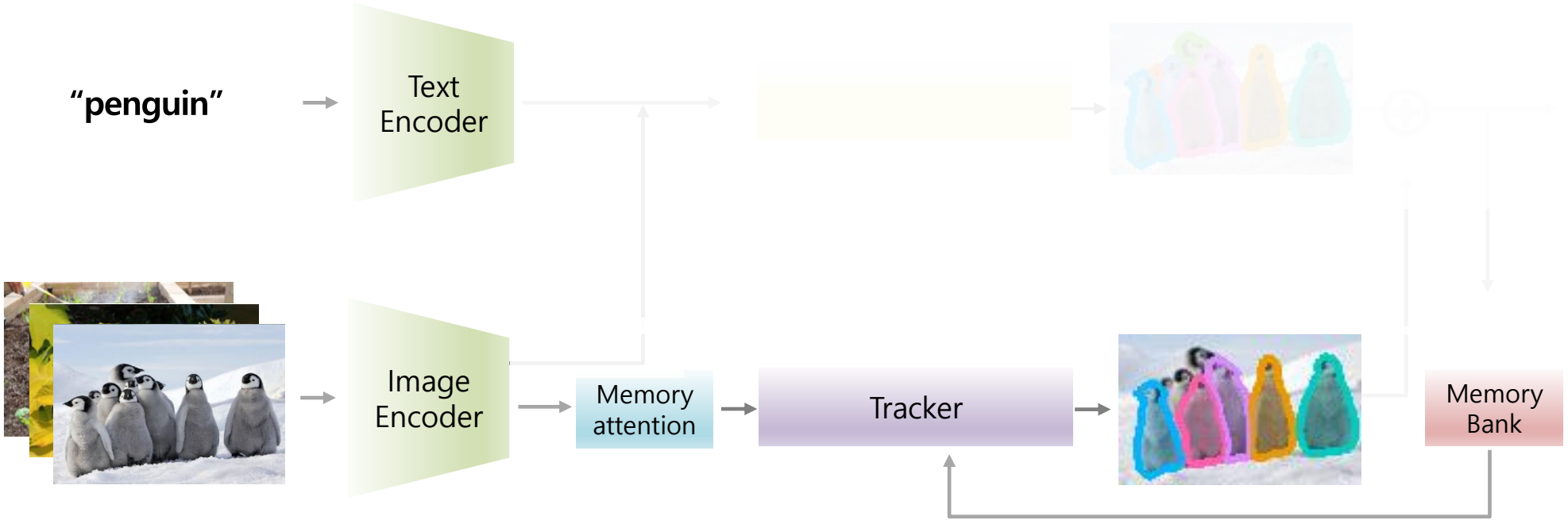


Appendix

SAM3

❖ SAM3는 어떻게 학습될까?

- 기능을 단계적으로 학습하는 구조
  - ① Concept 이해 → ② Object Localization → ③ Segmentation → ④ **Tracking (tracker training with a frozen backbone)**

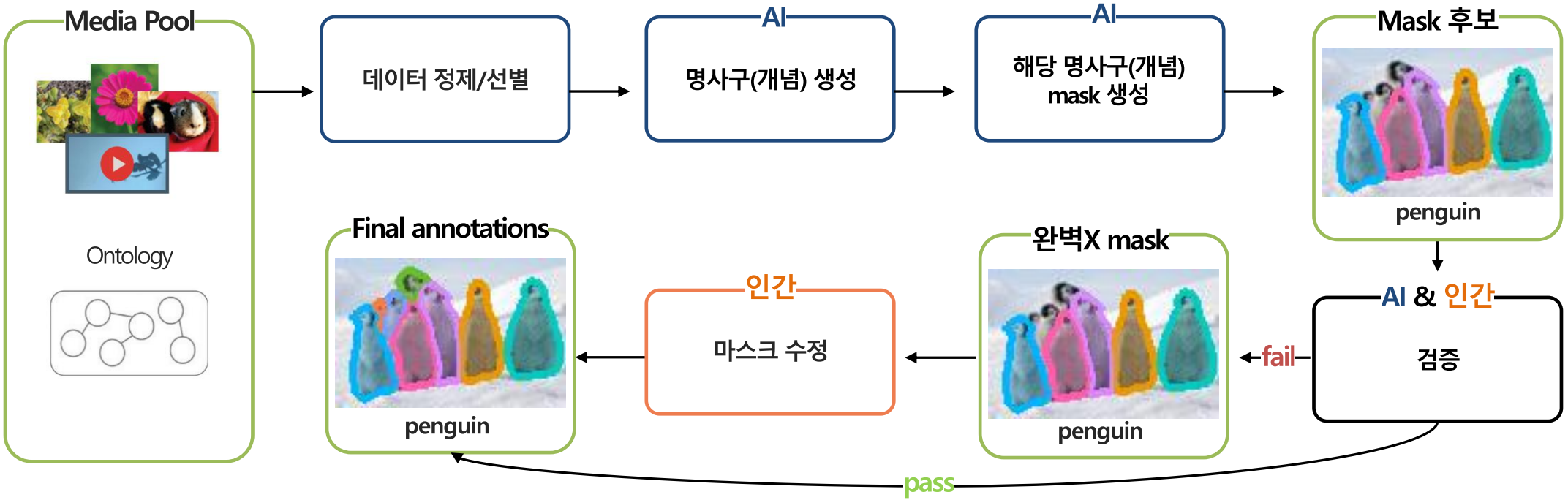


Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- 12만개 이미지 + 1.7천 개의 비디오 20만 7천개의 고유 개념 + 완전한 마스크
- 기존 벤치마크보다 50배 이상 많은 개념
- Data engine (총 4단계)



Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- **Data engine 1단계 – 인간 검증**
  - 방법: 이미지-NP 자동 생성 (Captioner+Parser)
  - SAM2 + Open-vocabulary detector로 초기 마스크 생성
  - 인간 검증을 통한 데이터 필터링 (Accept / Reject) ⇒ (결과) 430만 쌍 이미지-NP 확보 (해당 데이터로 초기 SAM3 학습)

— penguins —

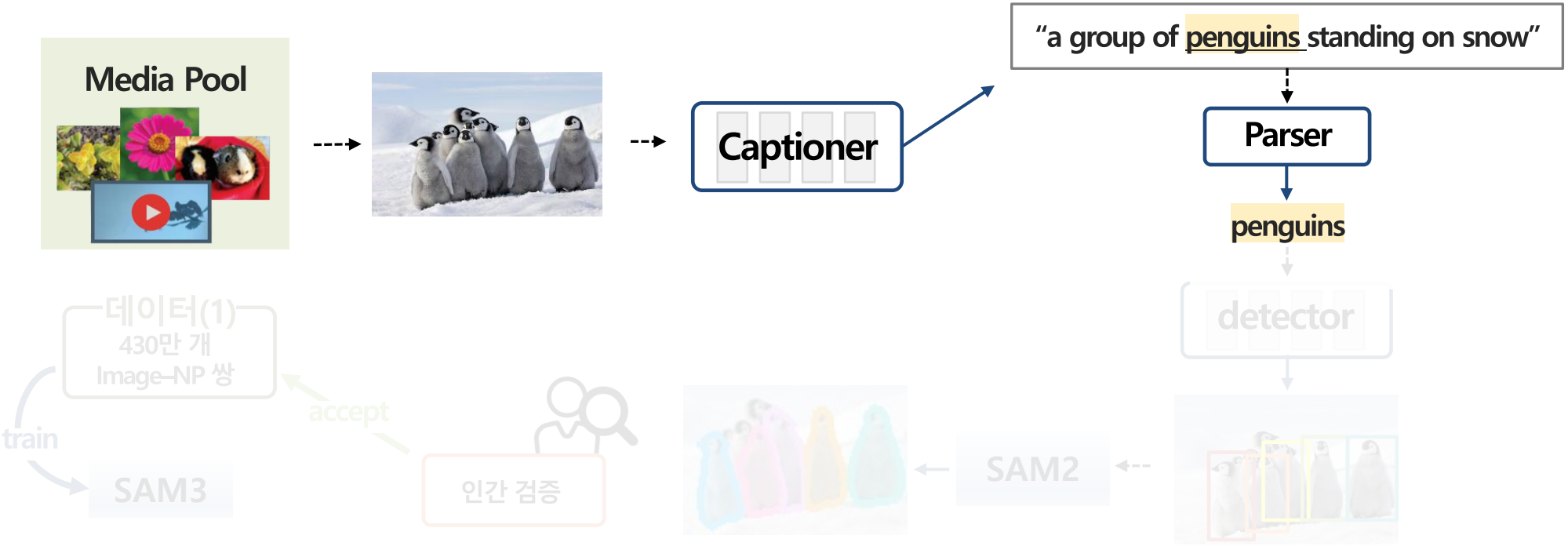
https://share.google/qiG3EgumwlEkbdbzK

Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- **Data engine 1단계 – 인간 검증**
  - 방법: 이미지-NP 자동 생성 (Captioner+Parser)
  - SAM2 + Open-vocabulary detector로 초기 마스크 생성
  - 인간 검증을 통한 데이터 필터링 (Accept / Reject) ⇒ (결과) 430만 쌍 이미지-NP 확보 (해당 데이터로 초기 SAM3 학습)

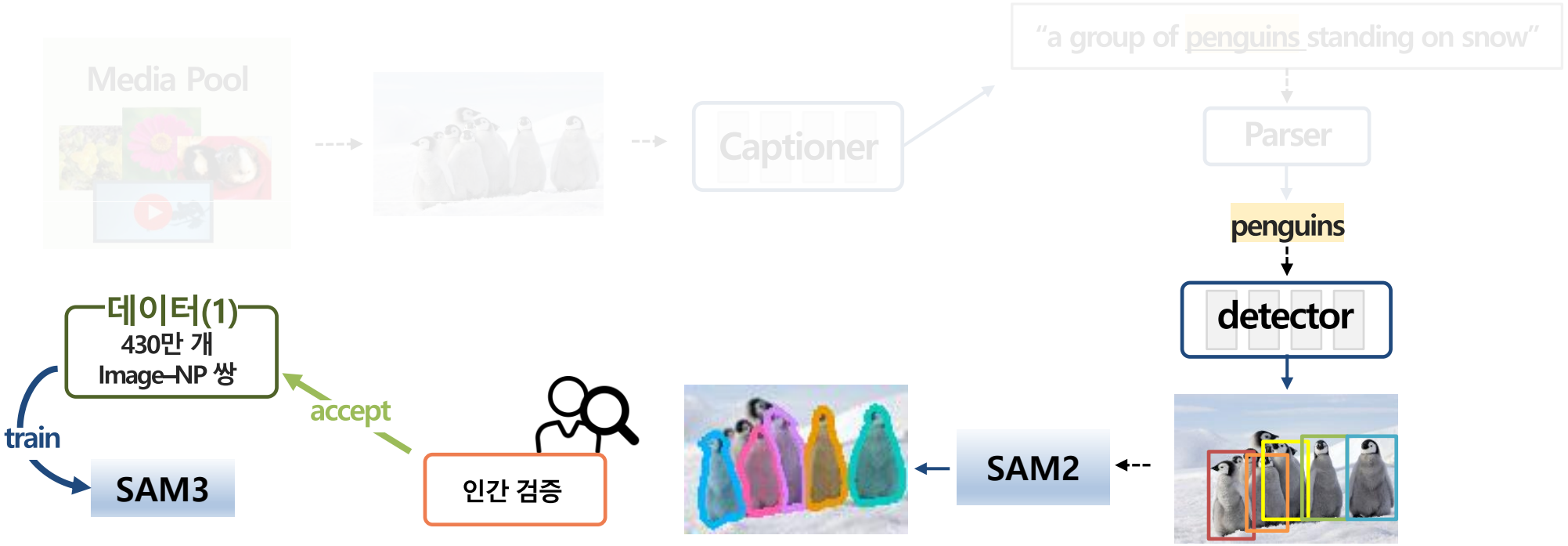


Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- **Data engine 1단계 – 인간 검증**
  - 방법: 이미지-NP 자동 생성 (Captioner+Parser)
  - SAM2 + Open-vocabulary detector로 초기 마스크 생성
  - 인간 검증을 통한 데이터 필터링 (Accept / Reject) ⇒ (결과) 430만 쌍 이미지-NP 확보 (해당 데이터로 초기 SAM3 학습)

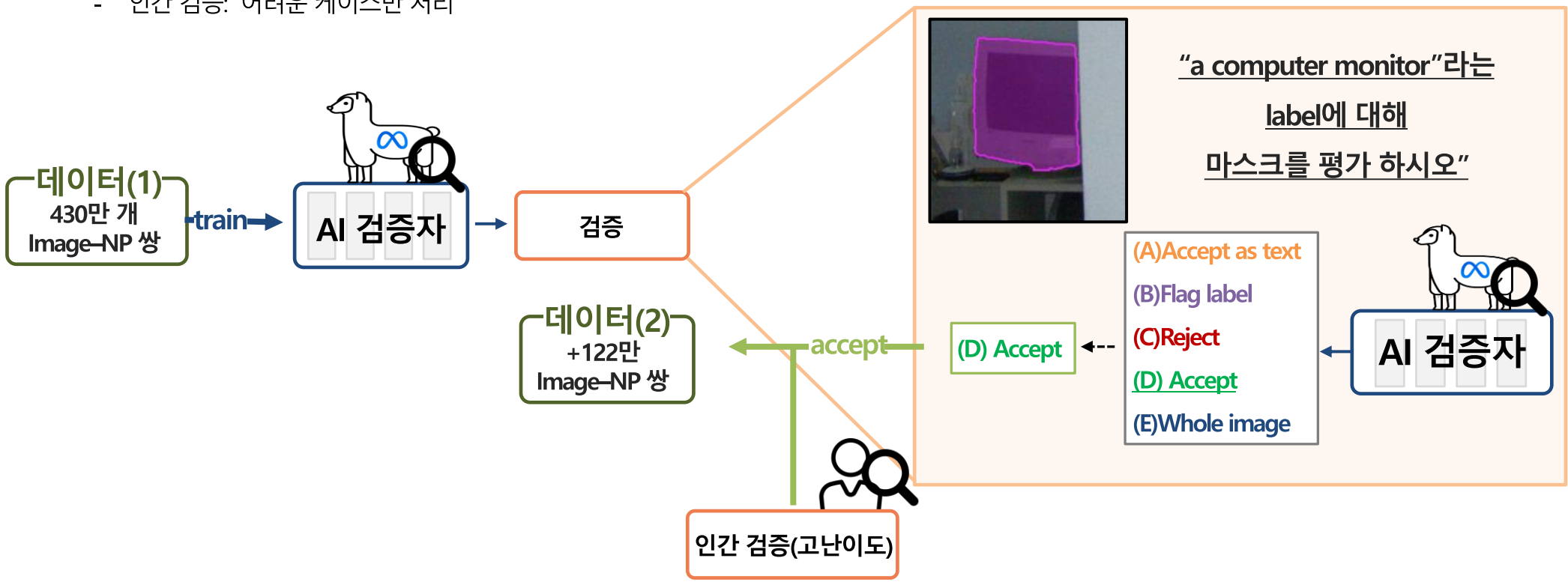


Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- **Data engine 2단계 – 인간+AI 검증 (반자동)**
  - **1단계에서 수집한 데이터로 Llmama3.2 기반 AI 검증자 학습** ⇒ mask quality / coverage 평가 (객관식)
  - SAM3 + AI verifier 반복 개선 (6회 업데이트) (결과): 122M 이미지–NP 쌍 추가
  - 인간 검증: 어려운 케이스만 처리



Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- **Data engine 2단계 – 인간+AI 검증 (반자동)**
  - **AI 검증자**
    - "당신은 객체 segmentation 마스크를 평가하는 전문가 주석자입니다. 이미지와 사전에 정의된 라벨이 주어지고, 하나의 마스크가 제공됩니다. 이 마스크의 품질을 평가해야 합니다."

[1번 규칙]
마스크가 라벨과 정확히 일치하고, 객체를 잘 덮고 있으며 경계도 적절하다면 "(D)Accept"로 평가하시오.
완벽한 픽셀 단위 정확도는 필요 없지만, 객체의 중요한 부분은 모두 포함해야 합니다

[2번 규칙]
마스크가 텍스트를 덮고 있고, 라벨이 그 텍스트와 정확히 일치하면 (A)"Accept as text"로 평가하시오. 글자의 모든 중요한 부분이 포함되어야 합니다.

[3번 규칙]
라벨이 민감한 정보(인종, 종교, 성적 지향 등)에 해당하면 (B)"Flag label"로 평가하시오.

[4번 규칙]
라벨이 이미지 전체를 의미하는 경우 (예: 장소, 환경, 장면 등) → (E)"Whole image"로 평가하시오.

[5번 규칙]
위에 해당하지 않으면 (C)"Reject"로 평가하시오.

"a computer monitor"라는 라벨에 대해 마스크를 평가하시오"

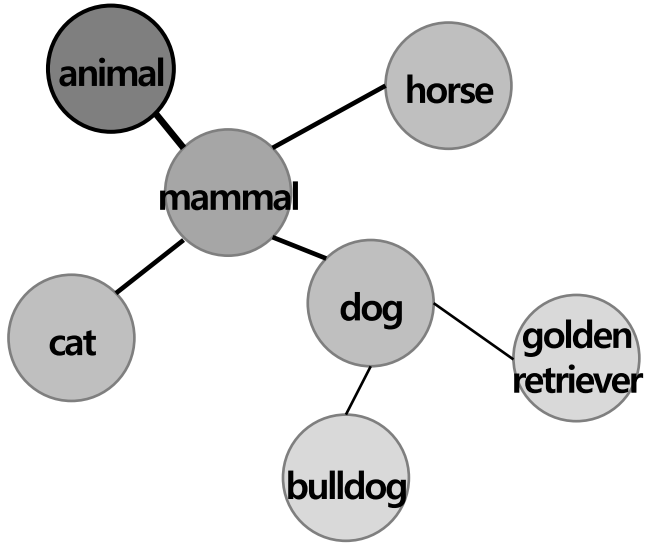
Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- **Data engine 3단계 – 더 어려운 개념 + 더 많은 도메인으로 확장**
  - **15개 데이터셋으로 도메인 확장**(+고난이도 데이터셋)
  - **Alt-text와 Wikidata 기반 명사구(NP) 확장**, 2,240만 노드 ontology 활용
  - **MV** verifier(개별 마스크 품질 검사) → zero-shot 가능 / **EV**(누락 객체 검증) verifier → 도메인별 인간 검증 필요

[Alt-text]



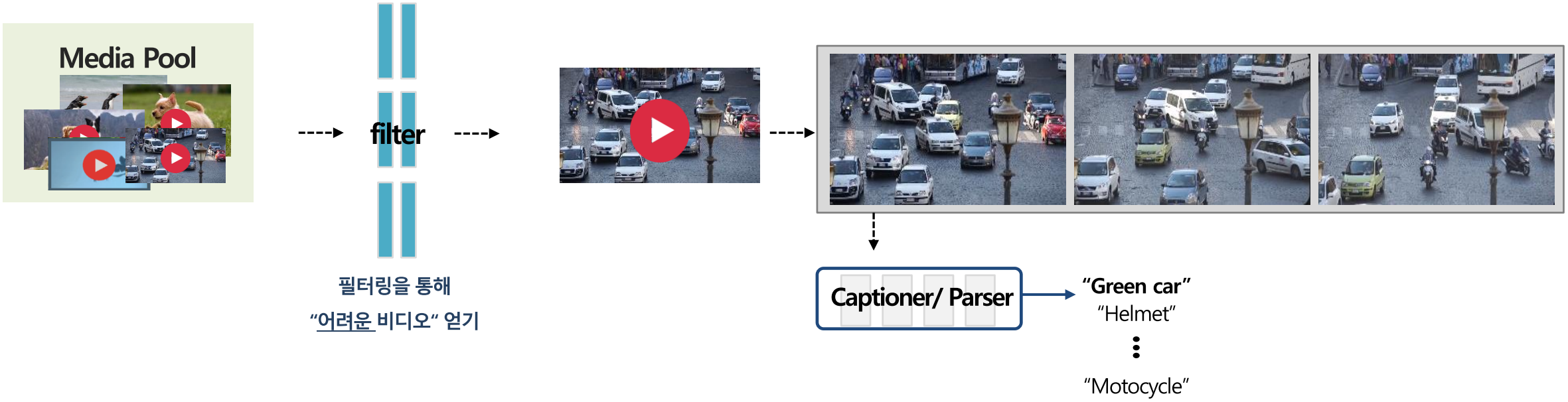
[Ontology]

Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- **Data engine 4단계 – 비디오로 확장**
  - 비디오 프레임 샘플링 → 기존 이미지 annotation pipeline 활용
  - SAM3를 활용한 **시간적 마스크 추적(masklet) 생성**
  - (결과) 5.25만개의 videos 46.7만 개 masklets
  - 복잡한 장면(다수 객체, 빠른 움직임 등) 중심으로 데이터 구성



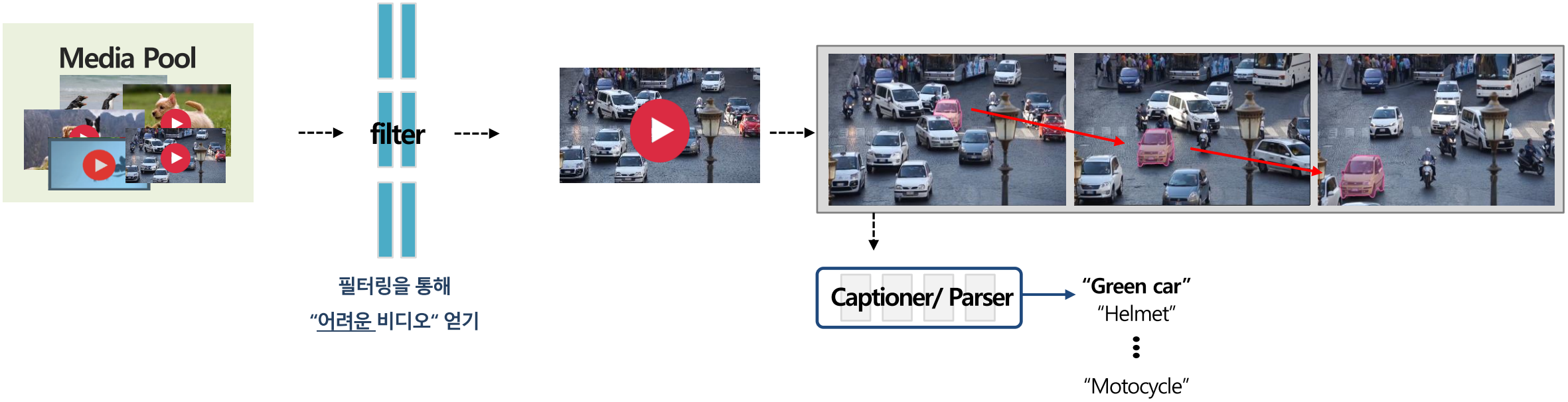
https://www.fitpetmall.com/wp-content/uploads/2023/10/GettyImages-492548888-1.png

Method

SAM3

❖ SAM3는 어떻게 다양한 개념 데이터를 구축할까?

- **Data engine 4단계 – 비디오로 확장**
  - 비디오 프레임 샘플링 → 기존 이미지 annotation pipeline 활용
  - SAM3를 활용한 **시간적 마스크 추적(masklet) 생성**
  - (결과) 5.25만개의 videos 46.7만 개 masklets
  - 복잡한 장면(다수 객체, 빠른 움직임 등) 중심으로 데이터 구성



https://www.fitpetmall.com/wp-content/uploads/2023/10/GettyImages-492548888-1.png

Experiments

SAM3

❖ Q1. SAM3는 기존 모델들보다 얼마나 잘 작동할까?

→ 다양한 이미지 기반의 객체 분할 및 탐지 작업에서 대부분의 기존 시스템보다 뛰어나거나 경쟁력 있는 성능

- cgF_1: 객체 **존재 여부** + **분할 정확도**
- AP: **객체**를 얼마나 **정확하게 탐지**하는지
- AP_o: 학습 데이터셋에 포함되지 않은 **새로운 유형의 객체나 데이터에 대해 얼마나** 잘 작동하는지
- pmF_1: positive mask (정답 객체 영역)에 대해서만 계산하는 F1
- $mIoU$: Semantic segmentation 작업에서 픽셀 단위의 **분할 정확도**를 나타내는 지표

	Instance Segmentation						Box Detection								Semantic Segmentation		
	LVIS		SA-Co				LVIS		COCO		SA-Co				ADE-847	PC-59	Cityscapes
Model	cgF_1	AP	Gold cgF_1	Silver cgF_1	Bronze cgF_1	Bio pmF_1	cgF_1	AP	AP	AP_o	Gold cgF_1	Silver cgF_1	Bronze cgF_1	Bio pmF_1	mIoU	mIoU	mIoU
Human	–	–	72.8	–	–	–	–	–	–	–	74.0	–	–	–	–	–	–
OWLv2	20.1	–	17.3	7.6	3.9	0.64	19.9	35.2	38.2	42.4	16.9	7.1	4.1	0.95	–	–	–
OWLv2*	29.3	43.4	24.6	11.5	11.7	0.04	30.2	45.5	46.1	23.9	24.5	11.0	12.0	0.08	–	–	–
gDino-T	14.7	–	3.3	2.7	7.0	0.34	15.1	20.5	45.7	35.3	3.4	2.5	7.6	0.35	–	–	–
LLMDet-L	35.1	36.3	6.5	7.1	12.5	0.15	39.3	42.0	55.6	49.8	6.8	6.7	14.0	0.17	–	–	–
APE-D*	–	53.0†	16.4	7.3	12.4	0.00	–	59.6†	58.3†	–	17.3	7.7	14.3	0.00	9.2†	58.5†	44.2†
DINO-X	–	38.5†	21.3δ	–	–	–	–	52.4†	56.0†	–	22.5δ	–	–	–	–	–	–
Gemini 2.5	13.4	–	13.0	8.3	7.3	10.7	16.1	–	–	–	14.4	9.4	8.2	12.4	–	–	–
SAM 3	37.2	48.5	54.1	49.6	42.6	55.4	40.6	53.6	56.4	55.7	55.7	50.0	47.1	56.3	13.8	60.8	65.2

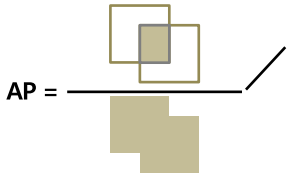
Experiments

SAM3

❖ Q1. SAM3는 기존 모델들보다 얼마나 잘 작동할까?

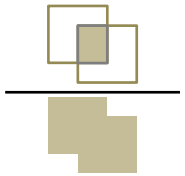
→ 다양한 이미지 기반의 객체 분할 및 탐지 작업에서 대부분의 기존 시스템보다 뛰어나거나 경쟁력 있는 성능

cgF1 = 객체가 존재하는가 x 위치를 잘 찾고 분할 했는가

AP = 

50% ↑ → 맞음 개수(TP) +1

맞음 개수 / 찾은 것들 전체 , 맞음 개수 / 실제로 있는 것 → 조화

mIoU= 

	Model	Instance Segmentation						Box Detection								Semantic Segmentation		
		LVIS (어려운 데이터)		SA-Co (SAM3가 만든 데이터(SA-CO))				LVIS (어려운 데이터)		COCO (기본 데이터)		SA-Co (SAM3가 만든 데이터(SA-CO))				ADE-847 (장면 데이터)	PC-59 (세밀한 데이터)	Cityscapes (자율주행 데이터)
		cgF1	AP	Gold cgF1	Silver cgF1	Bronze cgF1	Bio F1score	cgF1	AP	AP	AP^o	Gold cgF1	Silver cgF1	Bronze cgF1	Bio F1score	mIoU	mIoU	mIoU
	사람이 직접	–	–	72.8	–	–	–	–	–	–	–	74.0	–	–	–	–	–	–
텍스트 기반 Detection	OWLv2	20.1	–	17.3	7.6	3.9	0.64	19.9	35.2	38.2	42.4	16.9	7.1	4.1	0.95	–	–	–
텍스트 기반 Detection	OWLv2*(추가학습)	29.3	43.4	24.6	11.5	11.7	0.04	30.2	45.5	46.1	23.9	24.5	11.0	12.0	0.08	–	–	–
텍스트 기반 Detection	gDINO-T(경량화)	14.7	–	3.3	2.7	7.0	0.34	15.1	20.5	45.7	35.3	3.4	2.5	7.6	0.35	–	–	–
텍스트 기반 Detection	LLMDet-L(LLM+detection)	35.1	36.3	6.5	7.1	12.5	0.15	39.3	42.0	55.6	49.8	6.8	6.7	14.0	0.17	–	–	–
텍스트 기반 Detection	APE-D*(고성능/seg가능)	–	53.0†	16.4	7.3	12.4	0.00	–	59.6†	58.3†	–	17.3	7.7	14.3	0.00	9.2†	58.5†	44.2†
텍스트 기반 Detection	DINO-X(고성능)	–	38.5†	21.3δ	–	–	–	–	52.4†	56.0†	–	22.5δ	–	–	–	–	–	–
멀티모달	Gemini2.5(멀티모달 LLM)	13.4	–	13.0	8.3	7.3	10.7	16.1	–	–	–	14.4	9.4	8.2	12.4	–	–	–
텍스트 기반 Segmentation	SAM 3	37.2	48.5	54.1	49.6	42.6	55.4	40.6	53.6	56.4	55.7	55.7	50.0	47.1	56.3	13.8	60.8	65.2

Experiments

SAM3

❖ Q1. SAM3는 기존 모델들보다 얼마나 잘 작동할까?

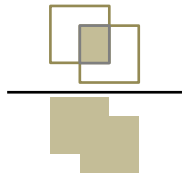
→ 다양한 이미지 기반의 객체 분할 및 탐지 작업에서 대부분의 기존 시스템보다 뛰어나거나 경쟁력 있는 성능

cgF1 = 객체가 존재하는가 x 위치를 잘 찾고 분할 했는가

AP =

50% ↑ → 맞음 개수(TP) +1

맞음 개수 / 찾은 것들 전체 — 맞음 개수 / 실제로 있는 것 → 조화

mIoU= 

	Model	Instance Segmentation						Box Detection								Semantic Segmentation		
		어려운 데이터		SAM3가 만든 데이터(SA-CO)				어려운 데이터		기본 데이터		SAM3가 만든 데이터(SA-CO)				장면 데이터	세밀한 데이터	자율주행 데이터-
		cgF1	AP	Gold cgF1	Silver cgF1	Bronze cgF1	Bio F1score	cgF1	AP	AP	AP_o	Gold cgF1	Silver cgF1	Bronze cgF1	Bio F1score	mIoU	mIoU	mIoU
	사람이 직접	–	–	72.8	–	–	–	–	–	–	–	74.0	–	–	–	–	–	–
텍스트 기반 Detection	OWLv2	20.1	–	17.3	7.6	3.9	0.64	19.9	35.2	38.2	42.4	16.9	7.1	4.1	0.95	–	–	–
텍스트 기반 Detection	OWLv2*(추가학습)	29.3	43.4	24.6	11.5	11.7	0.04	30.2	45.5	46.1	23.9	24.5	11.0	12.0	0.08	–	–	–
텍스트 기반 Detection	gDINO-T(경량화)	14.7	–	3.3	2.7	7.0	0.34	15.1	20.5	45.7	35.3	3.4	2.5	7.6	0.35	–	–	–
텍스트 기반 Detection	LLMDet-L(LLM+detection)	35.1	36.3	6.5	7.1	12.5	0.15	39.3	42.0	55.6	49.8	6.8	6.7	14.0	0.17	–	–	–
텍스트 기반 Detection	APE-D*(고성능/seg가능)	–	53.0†	16.4	7.3	12.4	0.00	–	59.6†	58.3†	–	17.3	7.7	14.3	0.00	9.2†	58.5†	44.2†
텍스트 기반 Detection	DINO-X(고성능)	–	38.5†	21.3δ	–	–	–	–	52.4†	56.0†	–	22.5δ	–	–	–	–	–	–
멀티모달	Gemini2.5(멀티모달 LLM)	13.4	–	13.0	8.3	7.3	10.7	16.1	–	–	–	14.4	9.4	8.2	12.4	–	–	–
텍스트 기반 Segmentation	SAM 3	37.2	48.5	54.1	49.6	42.6	55.4	40.6	53.6	56.4	55.7	55.7	50.0	47.1	56.3	13.8	60.8	65.2

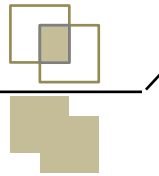
Experiments

SAM3

❖ Q1. SAM3는 기존 모델들보다 얼마나 잘 작동할까?

→ 다양한 이미지 기반의 객체 분할 및 탐지 작업에서 대부분의 기존 시스템보다 뛰어나거나 경쟁력 있는 성능

cgF1 = 객체가 존재하는가 x 위치를 잘 찾고 분할 했는가

AP = 

50% ↑ → 맞음 개수(TP) +1

맞음 개수 / 찾은 것들 전체

맞음 개수 / 실제로 있는 것

→ 조화

mIoU=

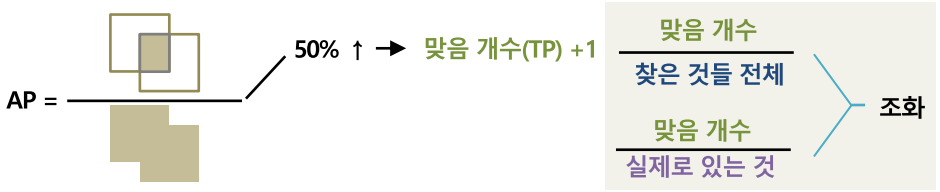
	Model	Instance Segmentation						Box Detection								Semantic Segmentation		
		어려운 데이터		SAM3가 만든 데이터(SA-CO)				어려운 데이터		기본 데이터		SAM3가 만든 데이터(SA-CO)				장면 데이터	세밀한 데이터	자율주행 데이터
		cgF1	AP	Gold cgF1	Silver cgF1	Bronze cgF1	Bio F1score	cgF1	AP	AP	AP_O	Gold cgF1	Silver cgF1	Bronze cgF1	Bio F1score	mIoU	mIoU	mIoU
	사람이 직접	–	–	72.8	–	–	–	–	–	–	–	74.0	–	–	–	–	–	–
텍스트 기반 Detection	OWLv2	20.1	–	17.3	7.6	3.9	0.64	19.9	35.2	38.2	42.4	16.9	7.1	4.1	0.95	–	–	–
텍스트 기반 Detection	OWLv2*(추가학습)	29.3	43.4	24.6	11.5	11.7	0.04	30.2	45.5	46.1	23.9	24.5	11.0	12.0	0.08	–	–	–
텍스트 기반 Detection	gDINO-T(경량화)	14.7	–	3.3	2.7	7.0	0.34	15.1	20.5	45.7	35.3	3.4	2.5	7.6	0.35	–	–	–
텍스트 기반 Detection	LLMDet-L(LLM+detection)	35.1	36.3	6.5	7.1	12.5	0.15	39.3	42.0	55.6	49.8	6.8	6.7	14.0	0.17	–	–	–
텍스트 기반 Detection	APE-D*(고성능/seg가능)	–	53.0†	16.4	7.3	12.4	0.00	–	59.6†	58.3†	–	17.3	7.7	14.3	0.00	9.2†	58.5†	44.2†
텍스트 기반 Detection	DINO-X(고성능)	–	38.5†	21.3δ	–	–	–	–	52.4†	56.0†	–	22.5δ	–	–	–	–	–	–
멀티모달	Gemini2.5(멀티모달 LLM)	13.4	–	13.0	8.3	7.3	10.7	16.1	–	–	–	14.4	9.4	8.2	12.4	–	–	–
텍스트 기반 Segmentation	SAM 3	37.2	48.5	54.1	49.6	42.6	55.4	40.6	53.6	56.4	55.7	55.7	50.0	47.1	56.3	13.8	60.8	65.2

Experiments

SAM3

❖ Q1. SAM3는 기존 모델들보다 얼마나 잘 작동할까?

→ 다양한 이미지 기반의 객체 분할 및 탐지 작업에서 대부분의 기존 시스템보다 뛰어나거나 경쟁력 있는 성능



	Model	Box Detection								Semantic Segmentation		
		어려운 데이터		기본 데이터		SAM3가 만든 데이터(SA-CO)				장면 데이터	세밀한 데이터	자율주행 데이터
		cgF1	AP	AP	AP_O	Gold cgF1	Silver cgF1	Bronze cgF1	Bio F1score	mIoU	mIoU	mIoU
	사람이 직접	–	–	–	–	74.0	–	–	–	–	–	–
텍스트 기반 Detection	OWLv2	19.9	35.2	38.2	42.4	16.9	7.1	4.1	0.95	–	–	–
텍스트 기반 Detection	OWLv2*(추가학습)	30.2	45.5	46.1	23.9	24.5	11.0	12.0	0.08	–	–	–
텍스트 기반 Detection	gDINO-T(경량화)	15.1	20.5	45.7	35.3	3.4	2.5	7.6	0.35	–	–	–
텍스트 기반 Detection	LLMDet-L(LLM+detection)	39.3	42.0	55.6	49.8	6.8	6.7	14.0	0.17	–	–	–
텍스트 기반 Detection	APE-D*(고성능/seg가능)	–	$59.6^{\dagger}$	$58.3^{\dagger}$	–	17.3	7.7	14.3	0.00	$9.2^{\dagger}$	$58.5^{\dagger}$	$44.2^{\dagger}$
텍스트 기반 Detection	DINO-X(고성능)	–	$52.4^{\dagger}$	$56.0^{\dagger}$	–	22.5^{δ}	–	–	–	–	–	–
멀티모달	Gemini2.5(멀티모달 LLM)	16.1	–	–	–	14.4	9.4	8.2	12.4	–	–	–
텍스트 기반 Segmentation	SAM 3	40.6	53.6	56.4	55.7	55.7	50.0	47.1	56.3	13.8	60.8	65.2

Experiments

SAM3

❖ Q2. SAM3는 새로운 데이터에서도 잘 작동할까?

→ Zero-shot / Few-shot에서도 강한 성능

- **ODinW13**: 13개의 object detection 데이터셋을 모아놓은 benchmark
- **RF-100VL**: 100개의 다양한 object detection 데이터셋 모음

		정제된 데이터 ODinW13		현실적인 데이터 RF-100VL	
	Model	AP_0	AP_{10}	AP_0	AP_{10}
	Gemini2.5-Pro	33.7	–	11.6	9.8
텍스트 기반 detection 모델	gDino-T	49.7	–	**15.7**	33.7
텍스트 기반 detection 모델	gDino1.5-Pro	58.7	67.9	–	–
	SAM 3	61.0	71.8	15.2	36.5

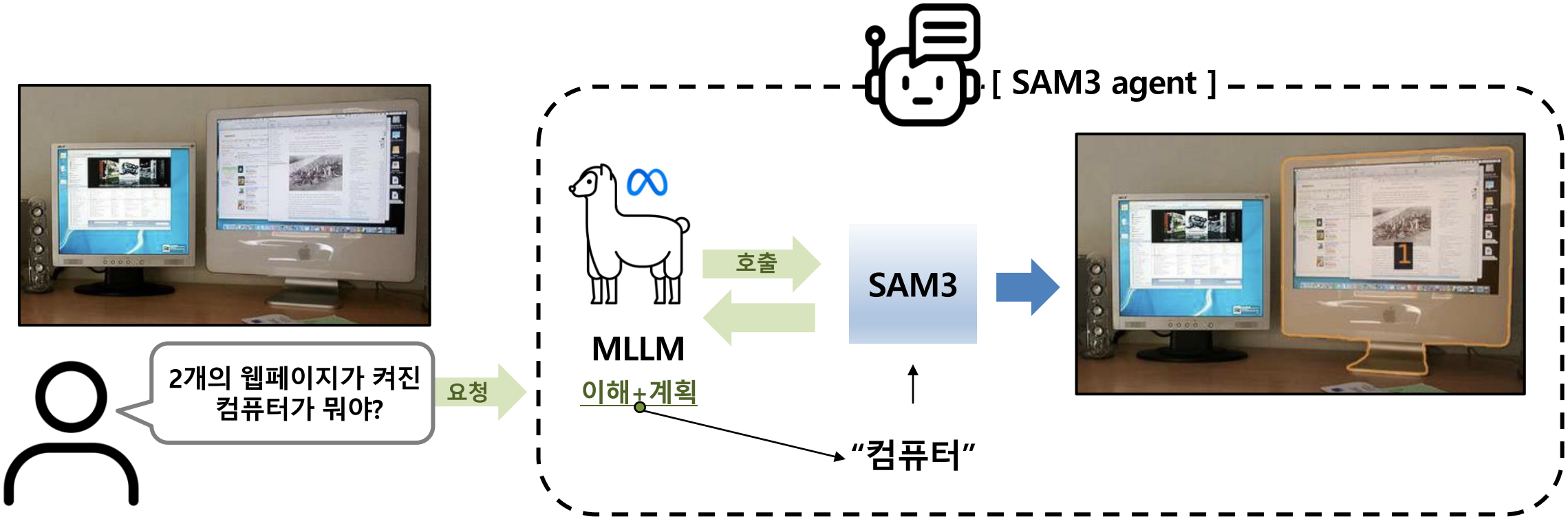
10개의 데이터셋으로 fine tuning 후 성능

Application

SAM3

❖ SAM3 agent

- **MLLM + SAM3 결합 구조**
  - "생각(LLM) + 실행(SAM3)"
- **자연어 요청 → segmentation mask**



Application

SAM3

❖ SAM3 agent

- **왜 SAM3 agent가 필요할까?**
  - ➢ SAM3는 입력 텍스트가 명확한 경우에는 잘 동작함
  - ➢ 추상적이거나 비교적 표현이 포함되면 이해에 한계가 있음

  ⇒ SAM3 agent로 해결!

"green car" → SAM3 →

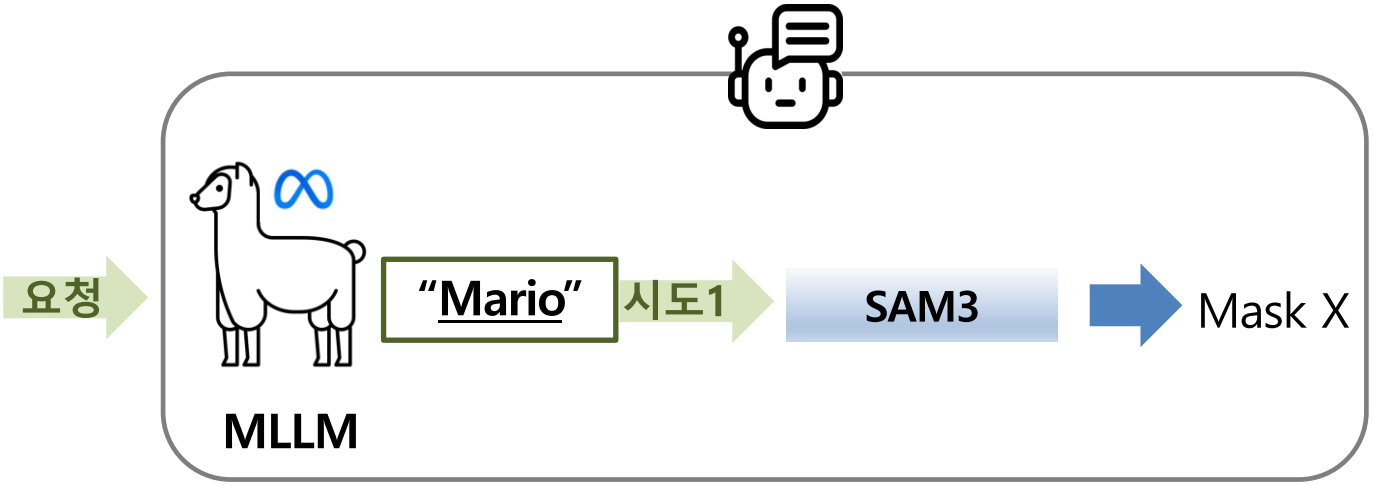
https://hips.hearstapps.com/hmg-prod/images/vol-27-pcoty-2025-group-1-677dbc61d16c0.jpg?crop=0.889333333333334xw:1xh;center,top&resize=1200:*

Application

SAM3

❖ SAM3 agent

- **예시: "The stronger mario"**
- SAM3 agent는 이를 이해하고 계획함 ("**Mario**" →)
- 최대 60번까지 trial-and-error를 반복하며 복잡한 쿼리를 점진적으로 해결

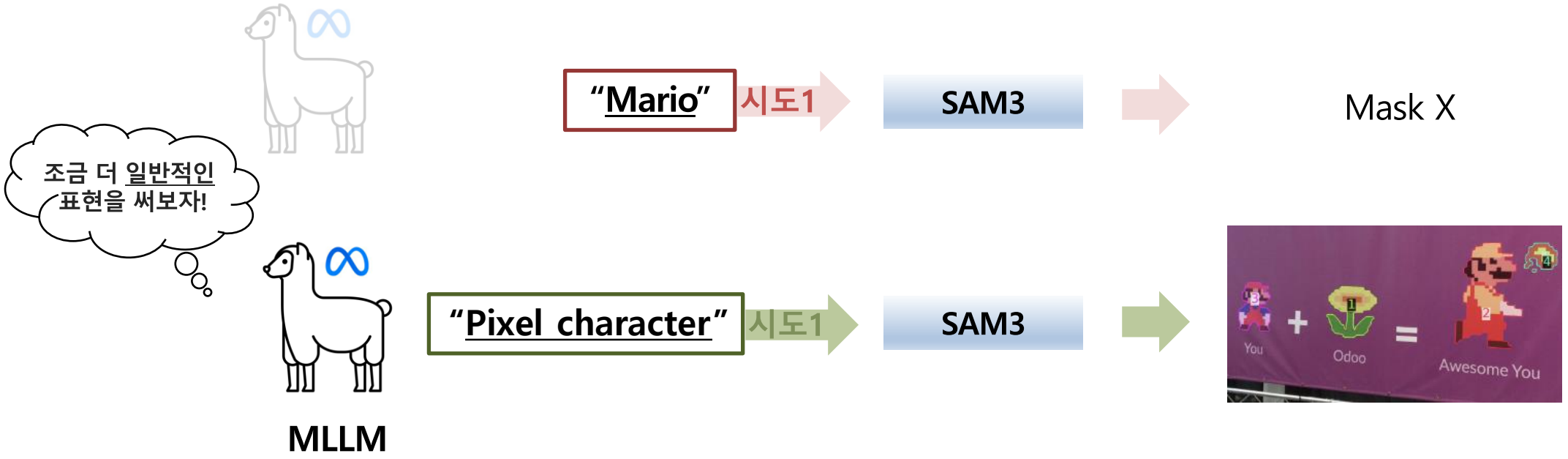


Application

SAM3

❖ SAM3 agent

- **예시: "The stronger mario"**
- SAM3 agent는 이를 이해하고 계획함 ("Mario" → "**pixel character**")
- 최대 60번까지 trial-and-error를 반복하며 복잡한 쿼리를 점진적으로 해결



Application

SAM3

❖ SAM3 agent

- **예시: "The stronger mario"**
- SAM3 agent는 이를 이해하고 계획함 **("Mario" → "pixel character")**
- 최대 60번까지 trial-and-error를 반복하며 복잡한 쿼리를 점진적으로 해결

후보 mask

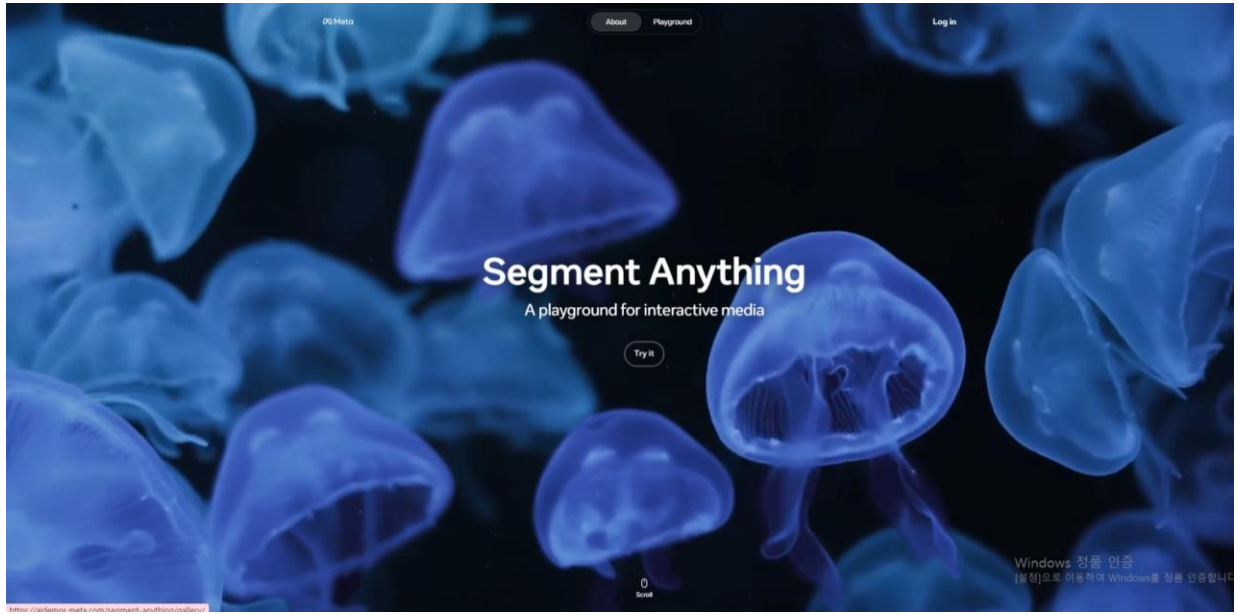
Final mask

Conclusion

SAM3

❖ Summary

- PCS task와 SA-Co benchmark를 새롭게 제안
- 기존 SAM2를 확장하여 concept 기반 segmentation을 수행하면서도 기존의 segmentation 성능을 유지
- 사람과 AI annotator의 장점을 결합한 고품질·고효율 데이터 엔진을 구축.



고맙습니다

References

[1] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4015-4026).

[2] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., ... & Feichtenhofer, C. (2025). SAM 2: Segment Anything in images and videos. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Ha6RTeWMd0

[3] Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., ... & Feichtenhofer, C. (2026). SAM 3: Segment Anything with concepts. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=r35clVtGzw

[4] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.

[5] Bolya, D., Huang, P.-Y., Sun, P., Cho, J. H., Madotto, A., Wei, C., ... & Feichtenhofer, C. (2026). Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=INqBOmwIpG