# Lightweight Segment Anything

## DMQA Open Seminar

2025.03.28

Data Mining & Quality Analytics Lab.

이혜승

# 발표자 소개

## 이혜승 (Hyeseung Lee)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S. Student (2024.09 ~ Present)

## Research Interest

- Image Segmentation, Foundation Model
- Multi-Agent LLM

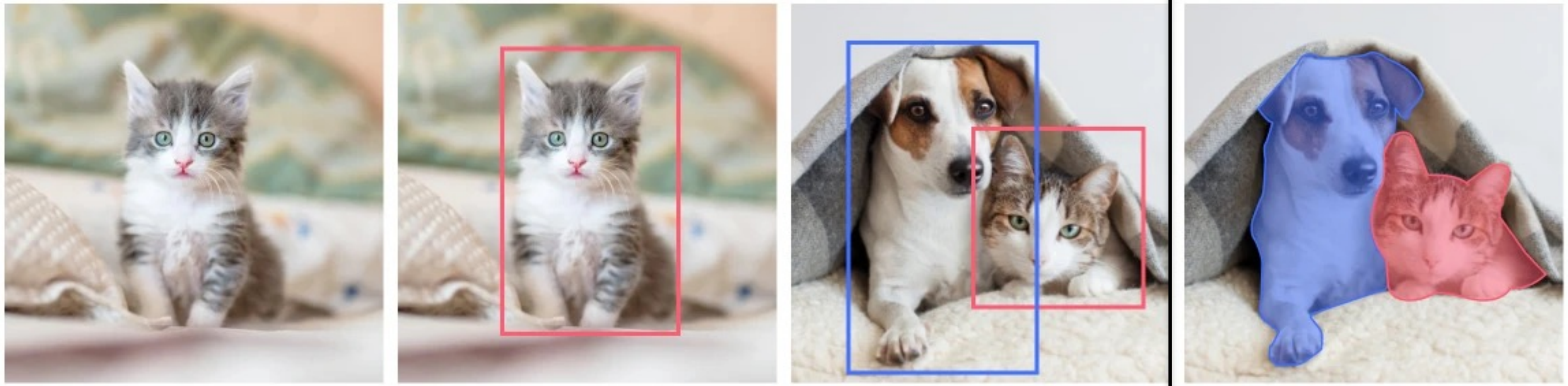
## Contact

- hyeseunglee@korea.ac.kr

# Introduction

# Image Segmentation

Classification | Classification + Localization | Object Detection | Segmentation



Cat | Cat | Cat, Dog | Cat, Dog

**픽셀 기반**의 이미지 분석

**각각의 픽셀들을** 특정 class로 **분류**

# Introduction

## ❖ Image Segmentation 의 종류

1) **Semantic Segmentation:** 같은 클래스에 속하면 하나로
2) **Instance Segmentation:** 같은 클래스 내에서도 객체 구분
3) **Panoptic Segmentation:** 배경과 객체를 모두 인식

Image

**Semantic** segmentation

**Instance** segmentation

**Panoptic** segmentation

# Segment Anything

## ❖Segment Anything (SAM)

- Prompt-guided **Vision foundation model** released by **Meta (ICCV, 2023)**
- 인용수: 9,515회

Meta

“**ChatGPT** of the **image segmentation** field”

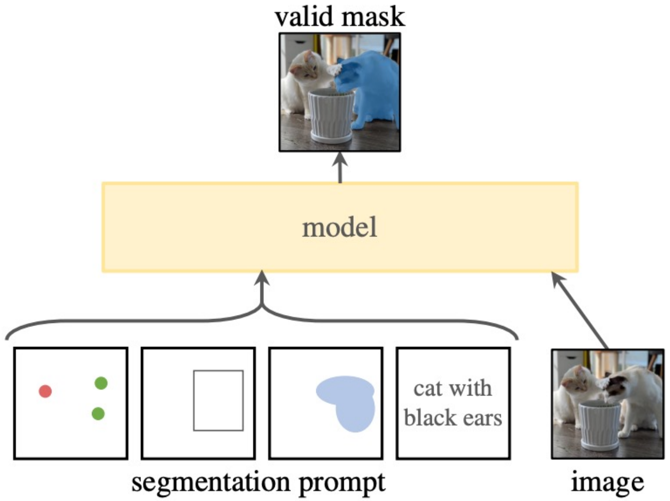
**Segment Anything**

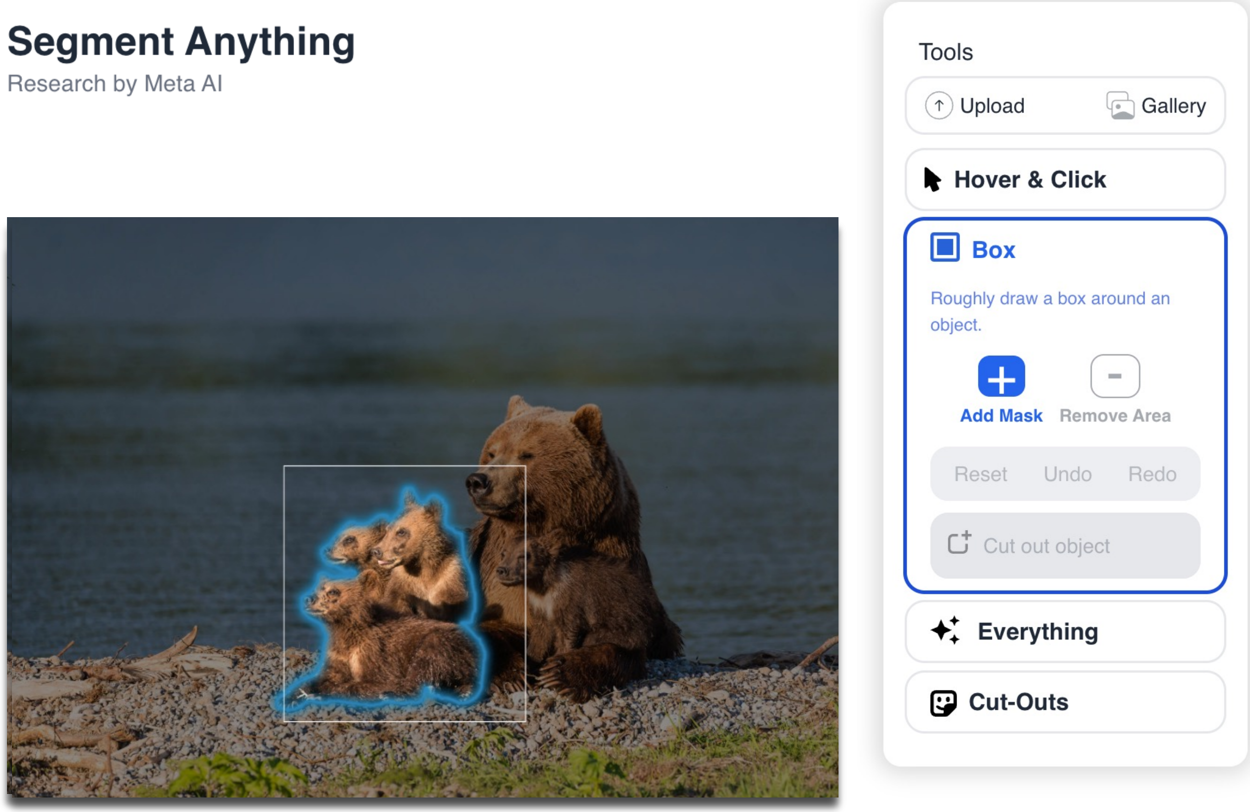
Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR

# Segment Anything

## ❖Segment Anything (SAM)

- Prompt-guided Vision foundation model

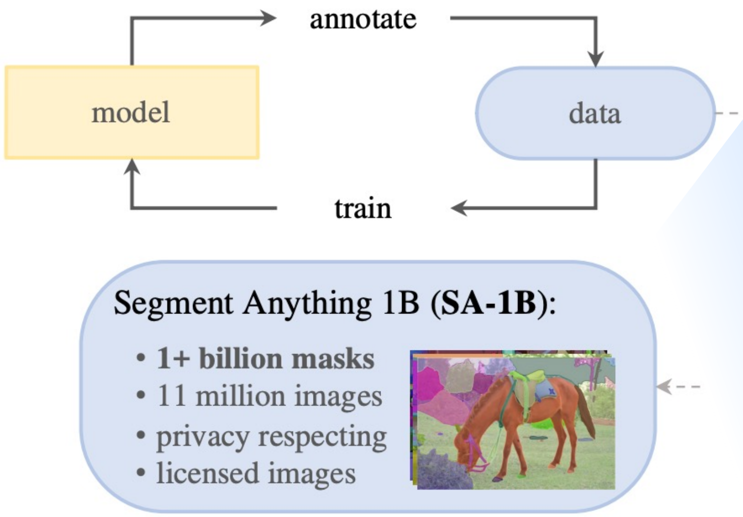


# Segment Anything

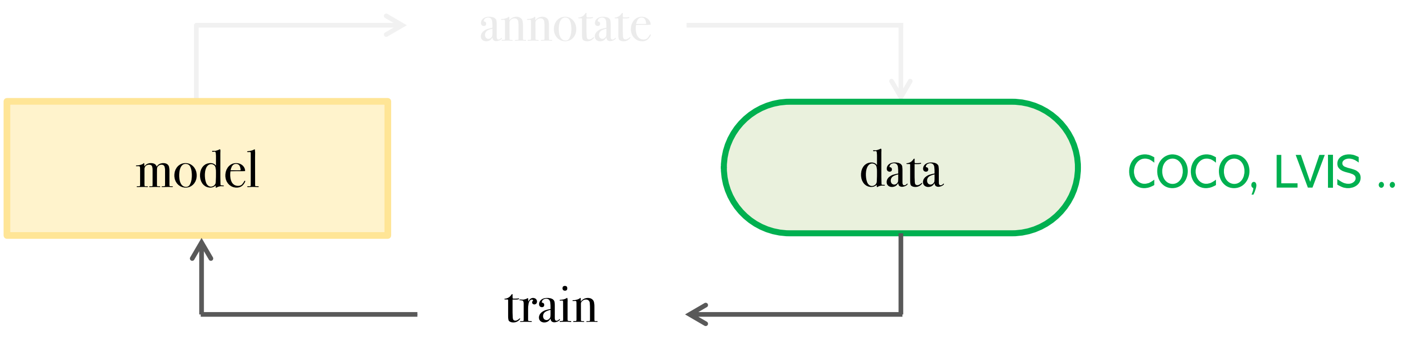
## ❖SA-1B dataset

- Prompt-guided Vision foundation model
- Trained on over 1 billion masks from 11 million image.




# Segment Anything

❖SA-1B dataset

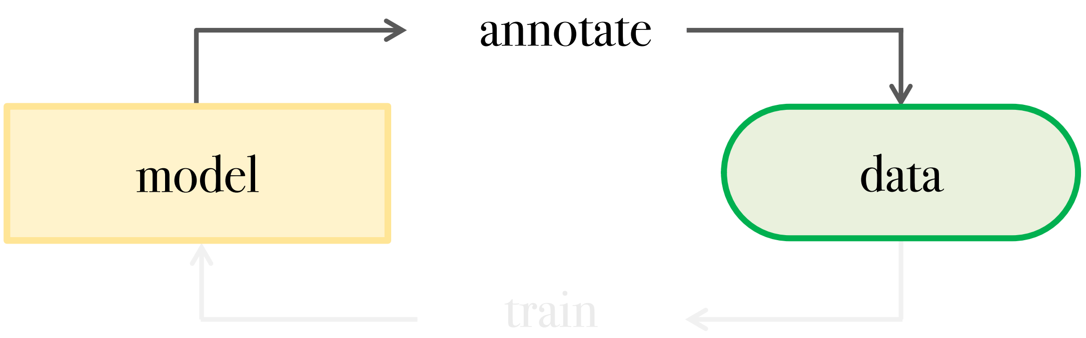



① Trained using common public segmentation datasets

# Segment Anything

## ❖SA-1B dataset

② model-assisted manual annotation stage

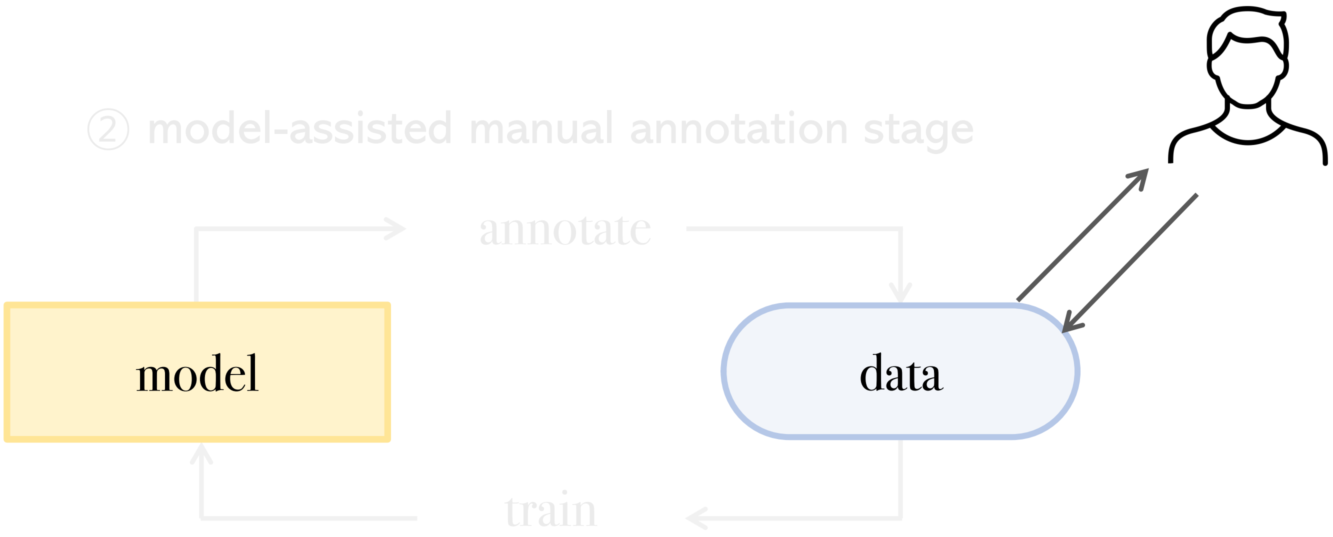



① Trained using common public segmentation datasets

# Segment Anything

## ❖SA-1B dataset

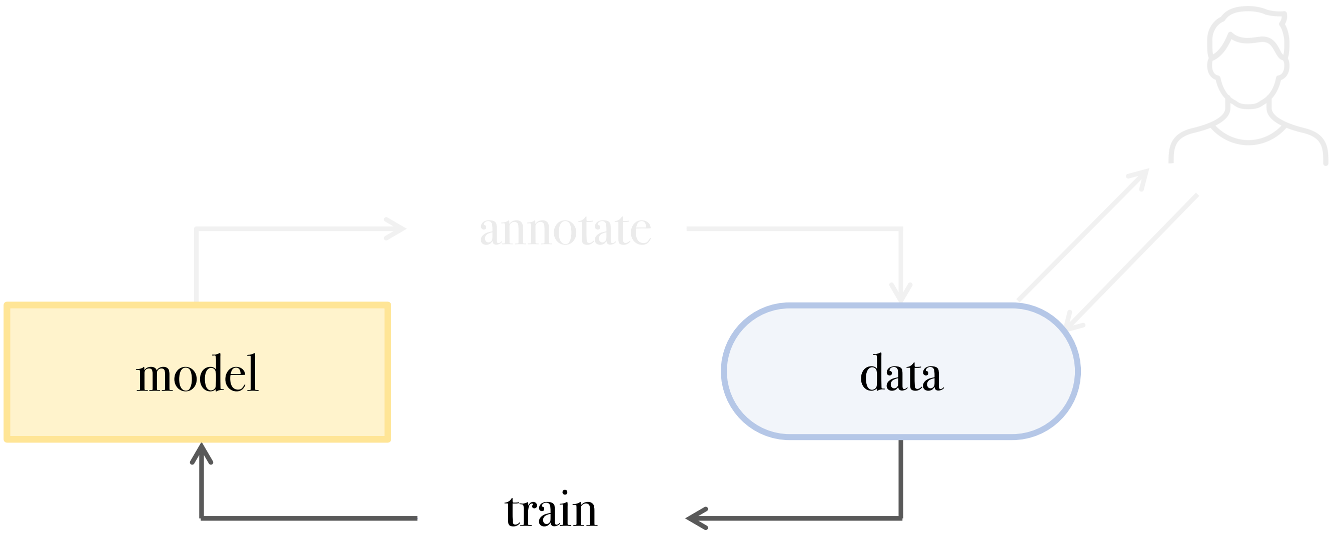
③ annotate any additional unannotated objects



# Segment Anything

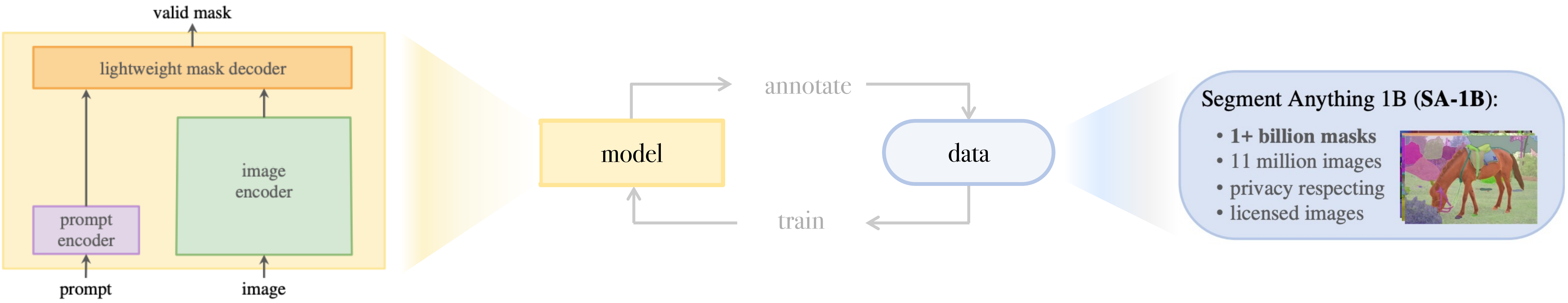
❖SA-1B dataset

③ annotate any additional unannotated objects




④ fully automatic stage
: model generates masks without annotator input

# Segment Anything

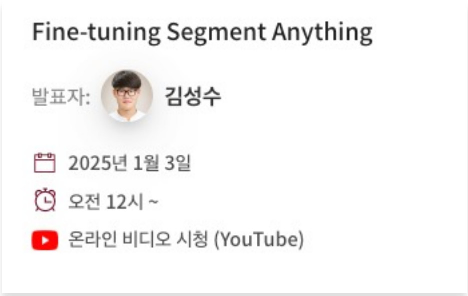


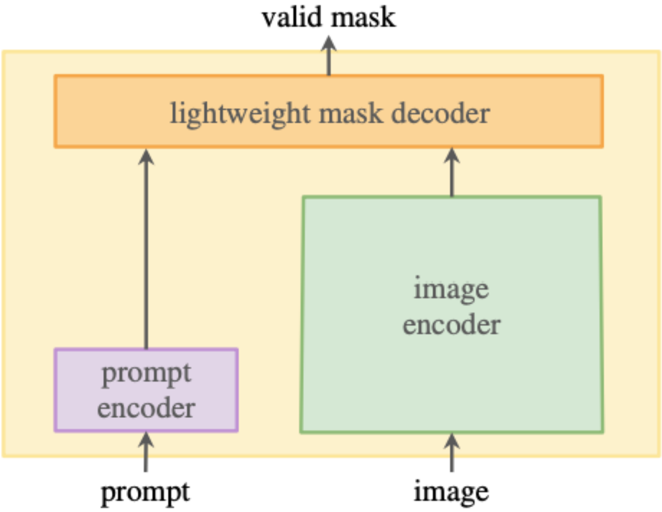

Segment Anything Model (SAM)

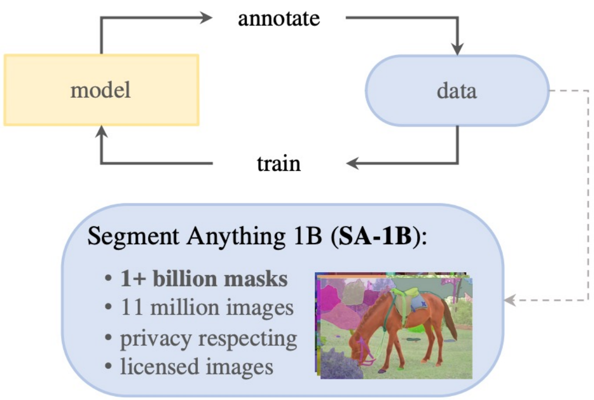
SA-1B dataset

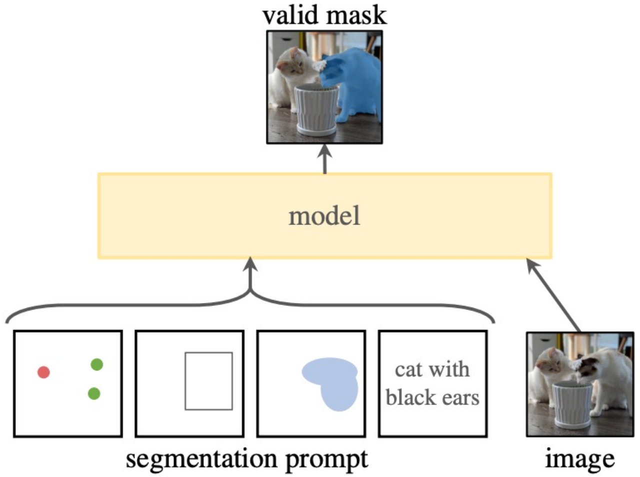
# Segment Anything

❖ Segment Anything (SAM)

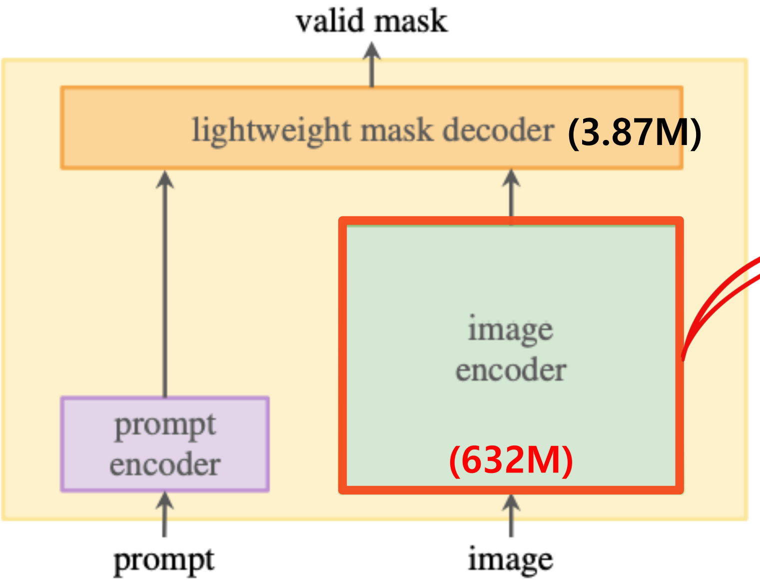



# Segment Anything

❖ Segment Anything (SAM)

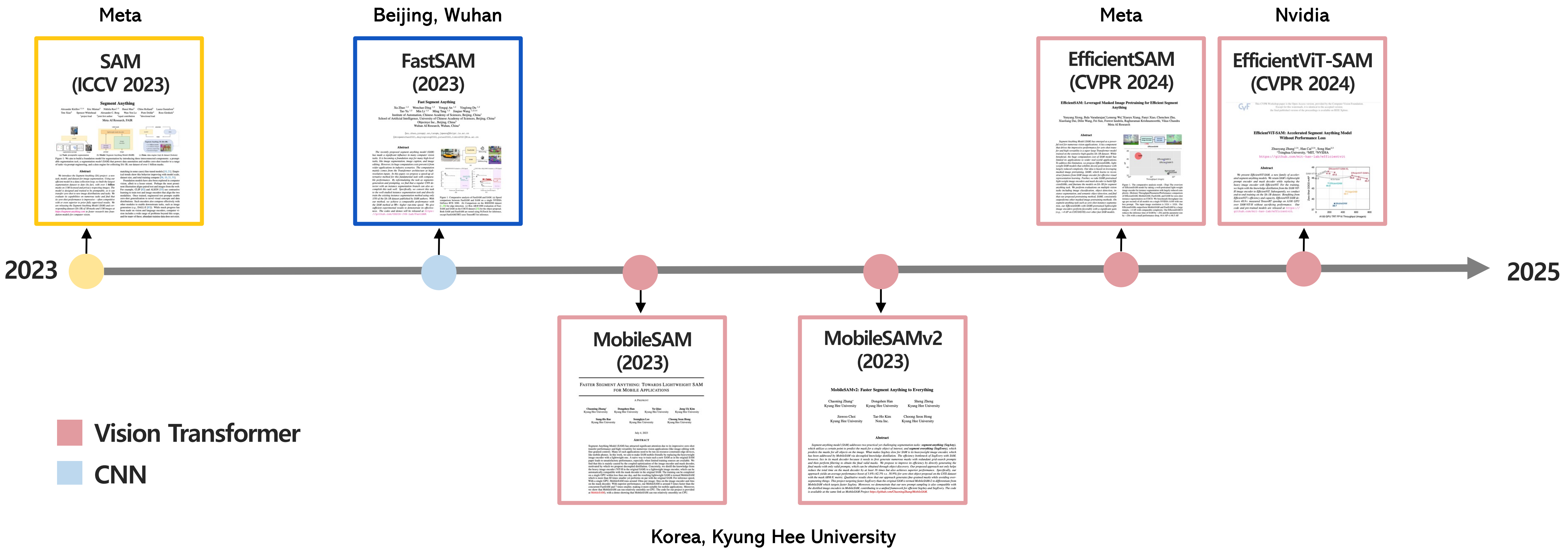
- Vision foundation model SAM의 막대한 계산 비용 지적
- SAM 경량화의 핵심: Image Encoder 경량화 !




"While beneficial,
the huge computation cost of SAM model
has limited its applications
to wider real-world applications."

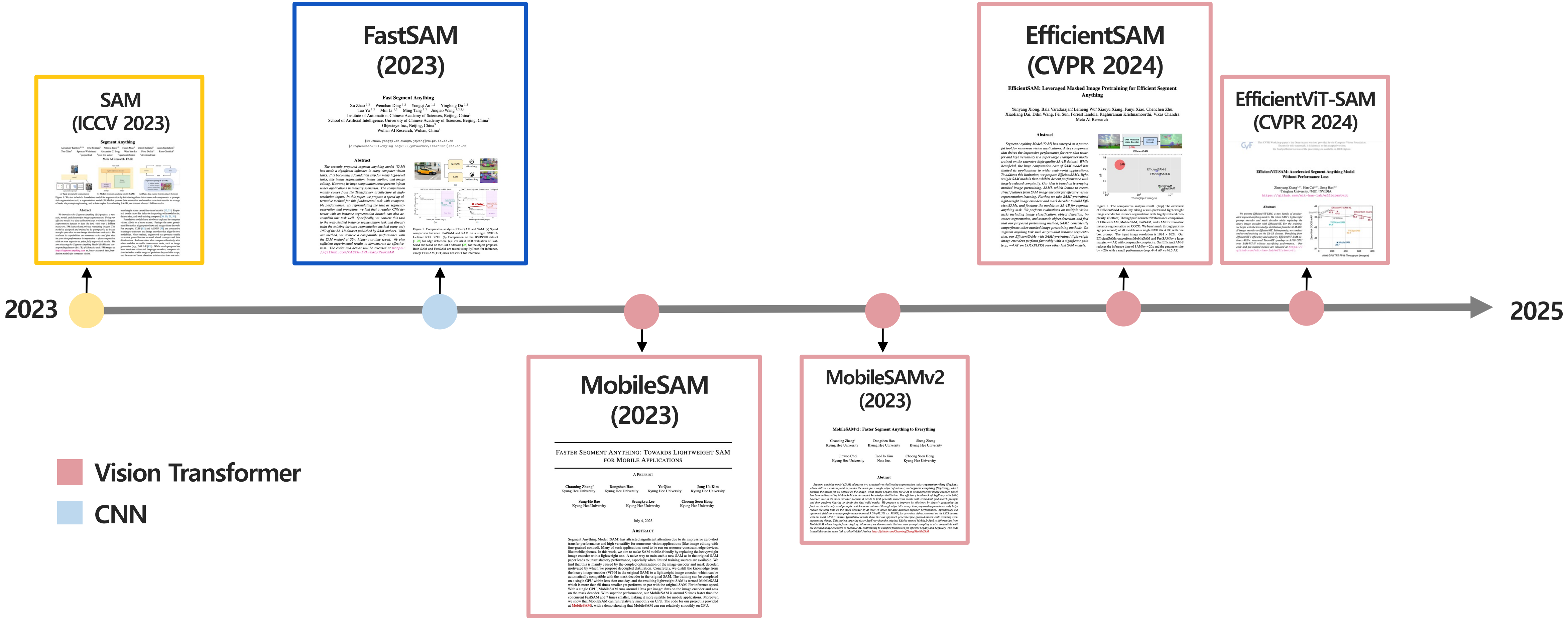
# Research Trends

❖ SAM 경량화 연구 동향



# Research Trends

❖ SAM 경량화 연구 동향



# FastSAM

# FastSAM

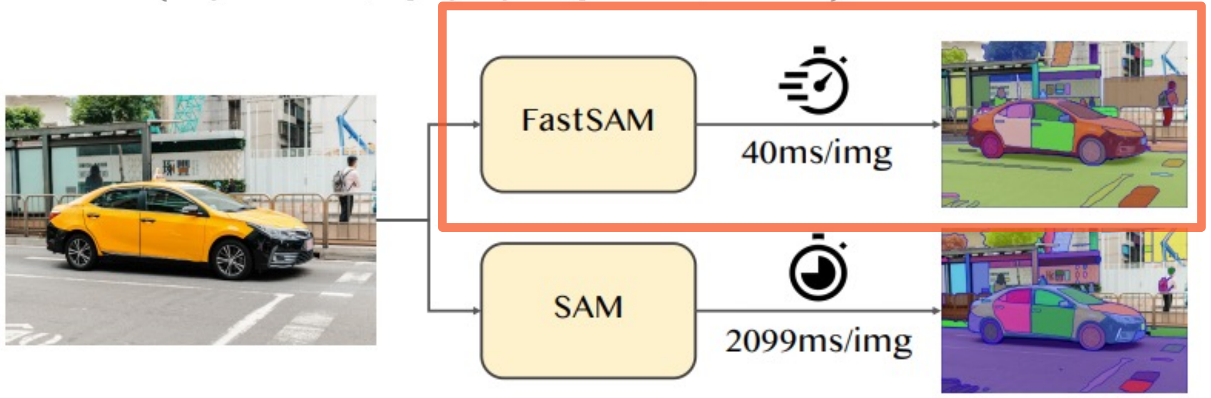
❖ Fast Segment Anything (2023.06.21)

- 인용수: 321회
- CNN backbone을 활용하여 image encoder 대체

**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]

{xu.zhao,yongqi.an,tangm,jqwang}@nlpr.ia.ac.cn
{dingwenchao2021,duyinglong2022,yutao2022,limin2021}@ia.ac.cn

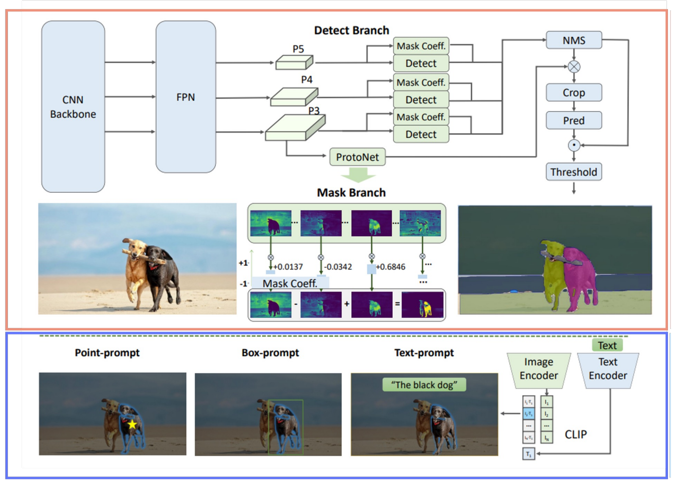


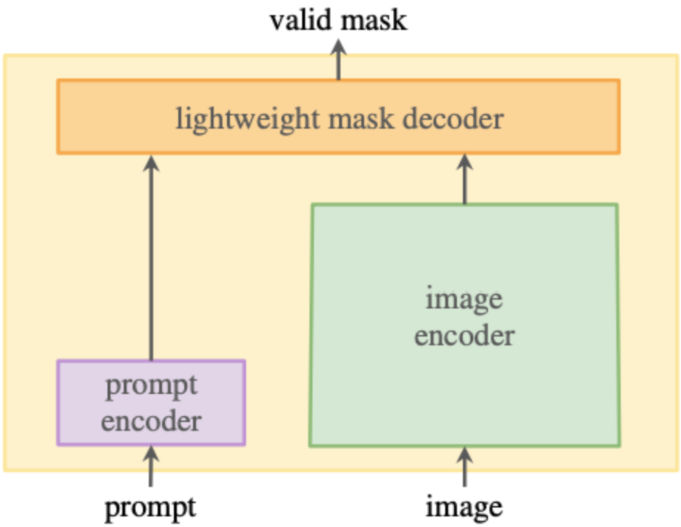
# FastSAM

## ❖ FastSAM vs SAM

- 파라미터 수 비교

| FastSAM | SAM |
| --- | --- |
| **(68M)** | **(632M)** |



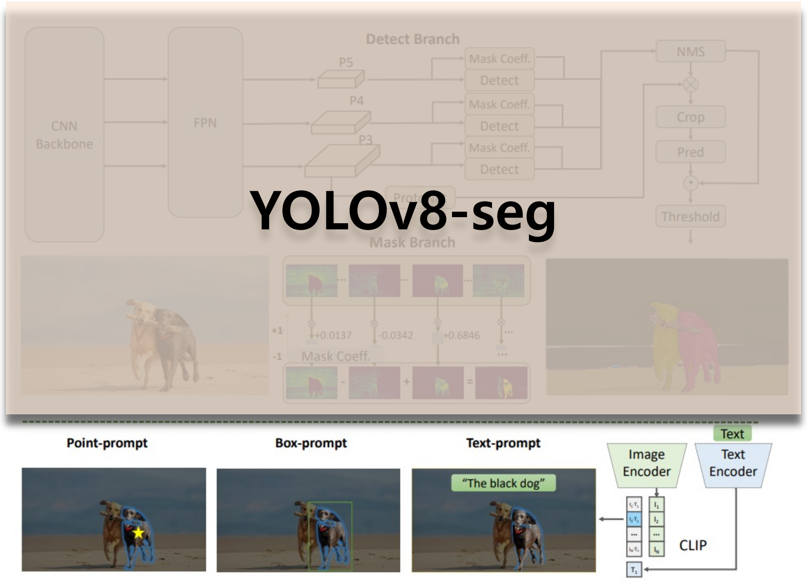



# FastSAM

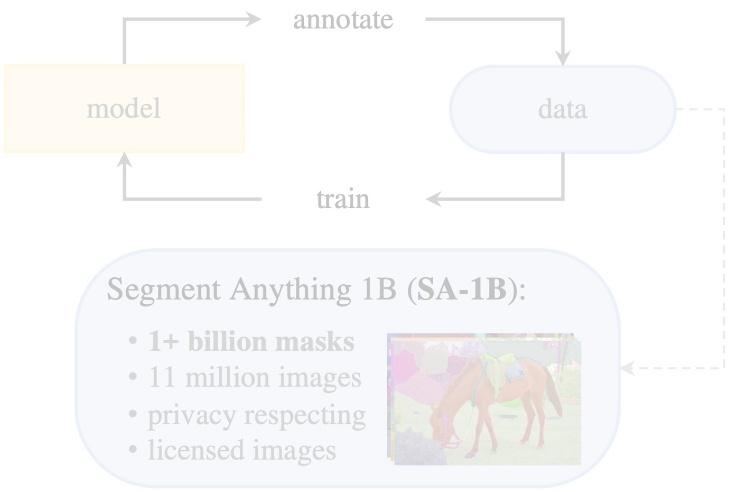
❖ Fast Segment Anything

- CNN backbone을 활용하여 image encoder 대체
- SAM에서 게시한 광범위 SA-1B 데이터셋 사용

**CNN backbone**




**Pretrain data**

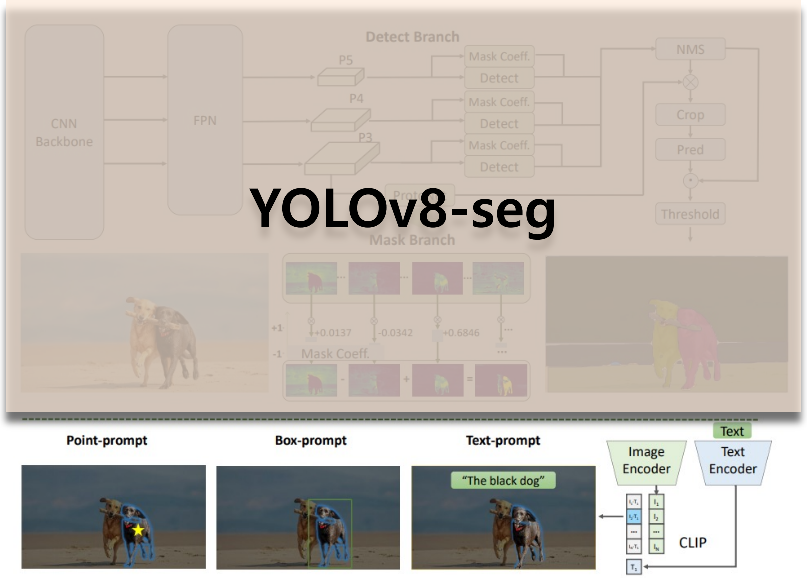


# FastSAM

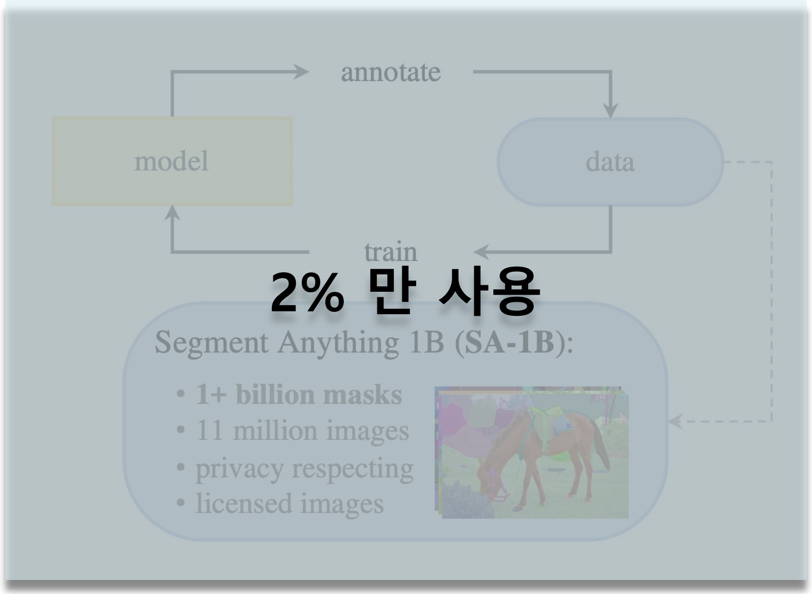
## ❖ Fast Segment Anything

- CNN backbone을 활용하여 image encoder 대체
- SAM에서 게시한 광범위 SA-1B 데이터셋 사용

### CNN backbone

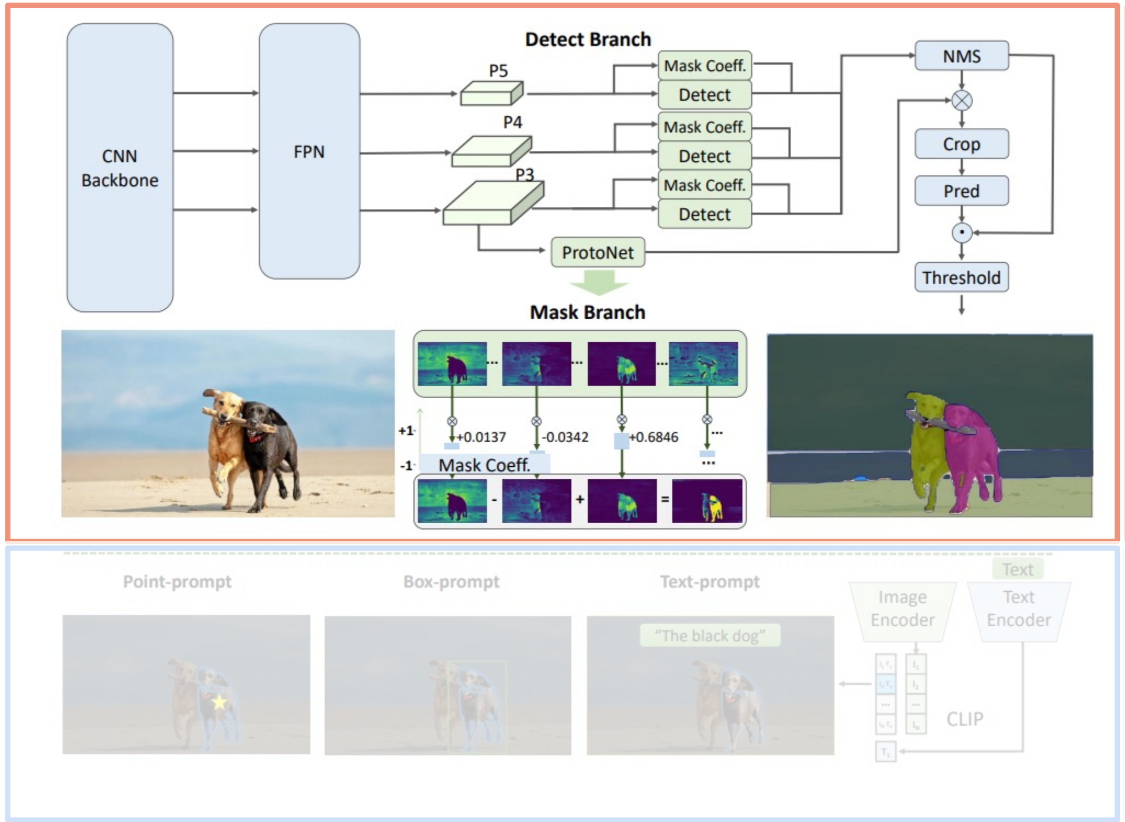



### Pretrain data



# FastSAM

❖ 2-Stage Framework




Stage 1

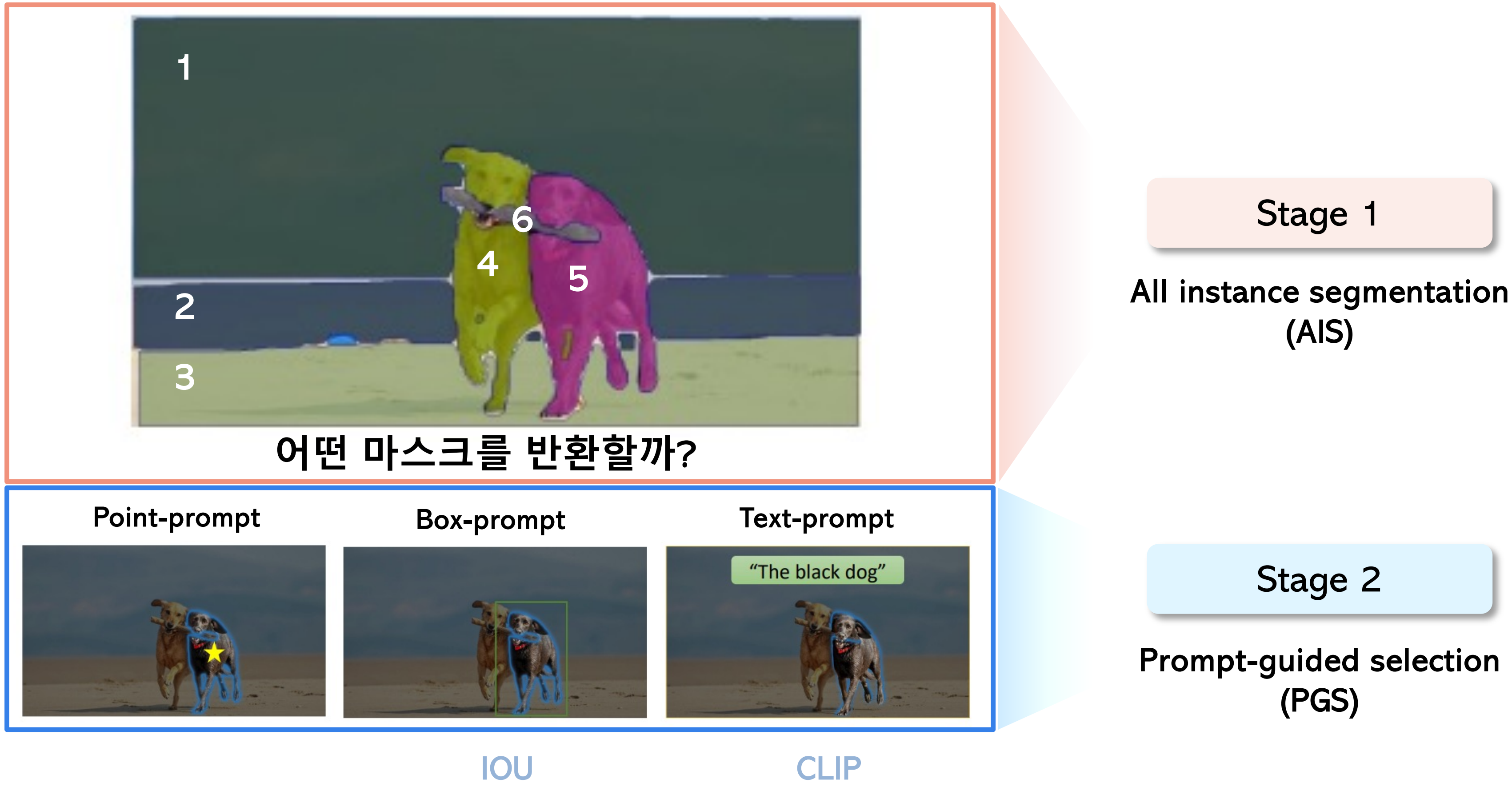
**All instance segmentation (AIS)**

Stage 2

Prompt-guided selection (PGS)

# FastSAM

❖ 2-Stage Framework



# FastSAM

## ❖ FastSAM vs SAM

- 추론 속도 비교: **50× higher run-time speed**

| 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 | | |

약 7초

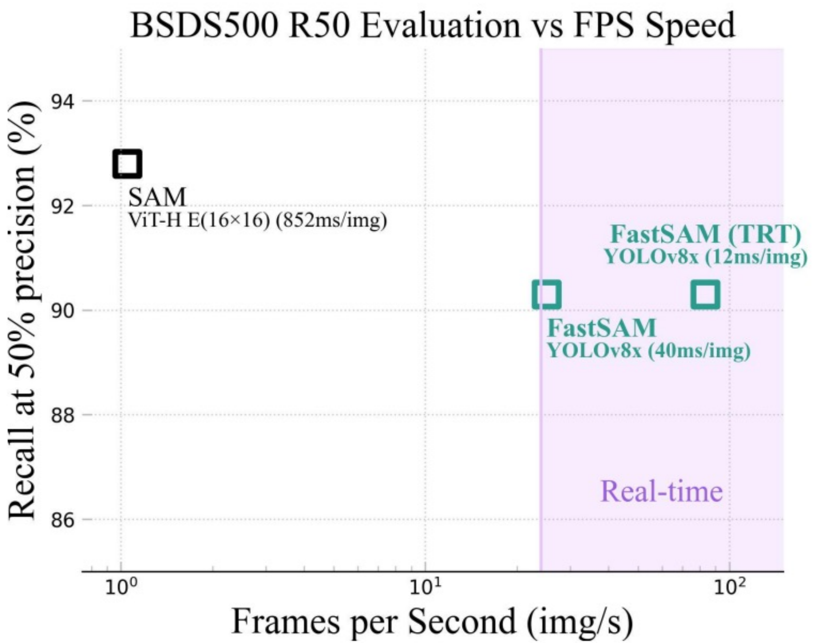
0.04초

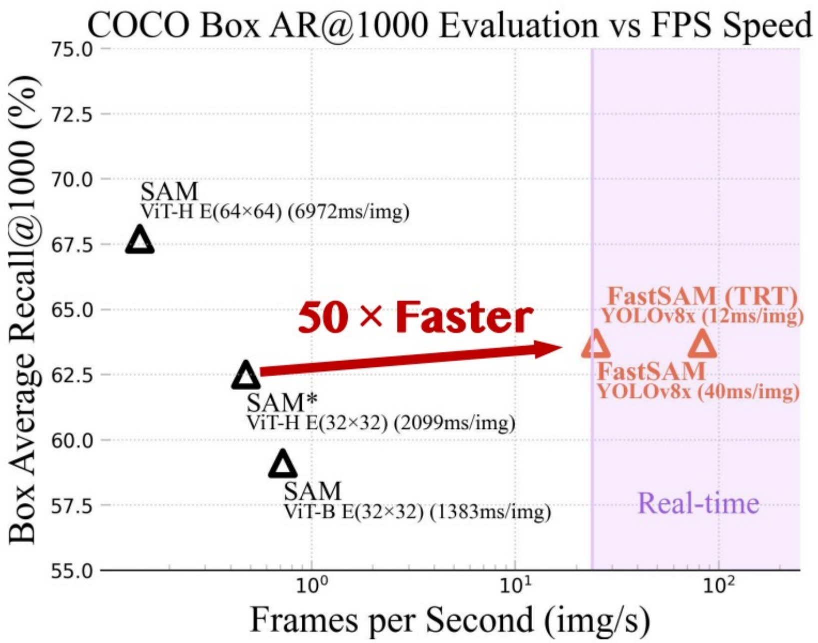
enabling real-time application

# FastSAM

## ❖ FastSAM vs SAM

- 추론 속도 비교: 50× higher run-time speed
- Achieve a comparable performance with the SAM



# MobileSAM

# MobileSAM

❖ Faster Segment Anything: Towards Lightweight SAM for Mobile Applications

- MobileSAM (2023.07.04 preprint)
- 인용수: 387회

## FASTER SEGMENT ANYTHING: TOWARDS LIGHTWEIGHT SAM FOR MOBILE APPLICATIONS

A PREPRINT

**Chaoning Zhang*** Kyung Hee University
**Dongshen Han** Kyung Hee University
**Yu Qiao** Kyung Hee University
**Jung Uk Kim** Kyung Hee University
**Sung-Ho Bae** Kyung Hee University
**Seungkyu Lee** Kyung Hee University
**Choong Seon Hong** Kyung Hee University

July 4, 2023

# MobileSAM

❖ **Faster Segment Anything: Towards Lightweight SAM for Mobile Applications**

## Segment Anything Model

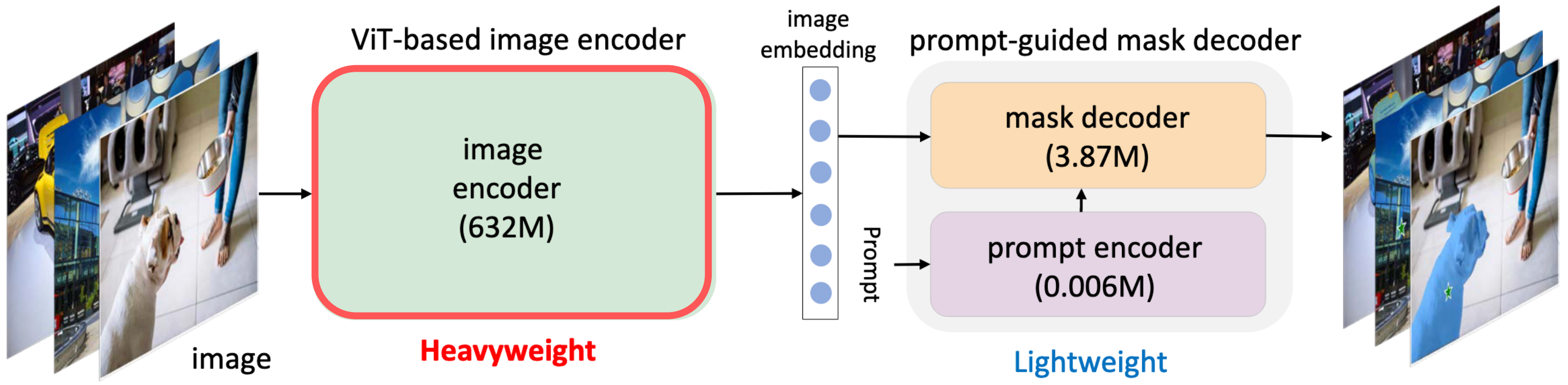
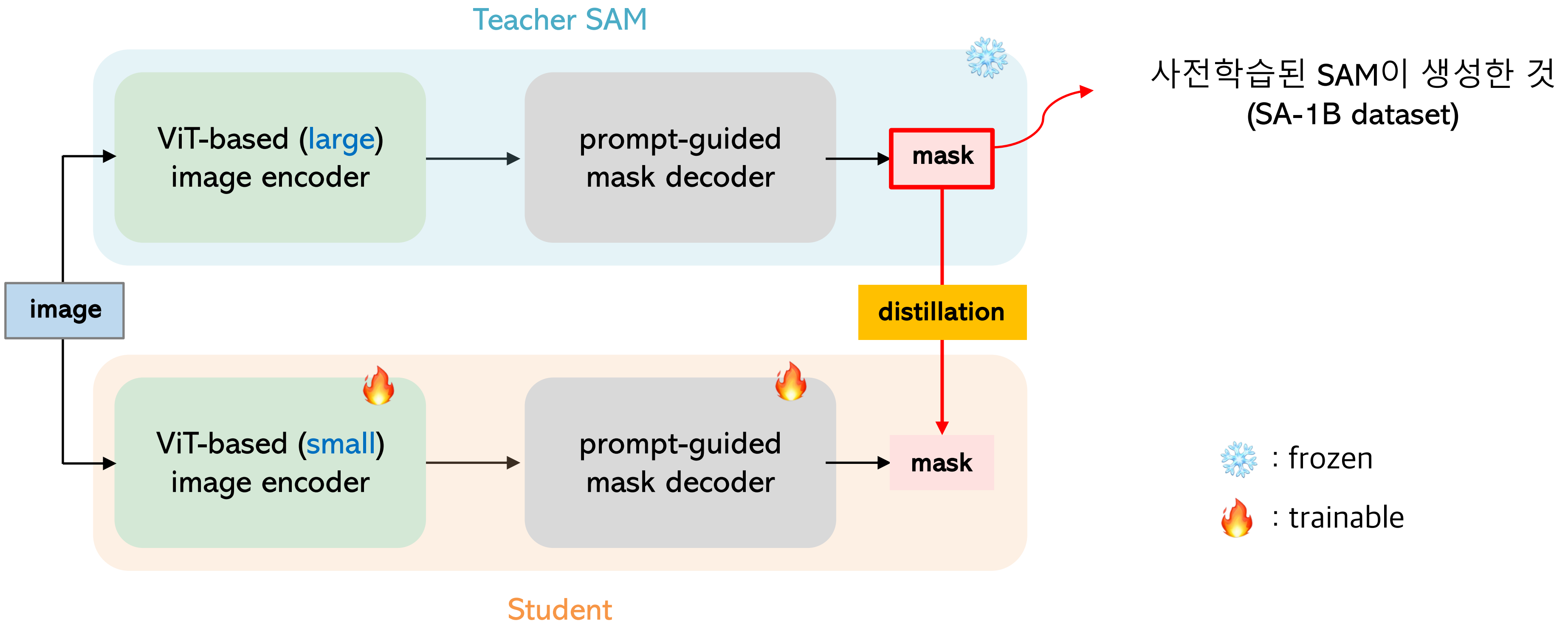



Figure 1: The overview of Segment Anything Model.

# MobileSAM

❖ Knowledge distillation (지식 증류)

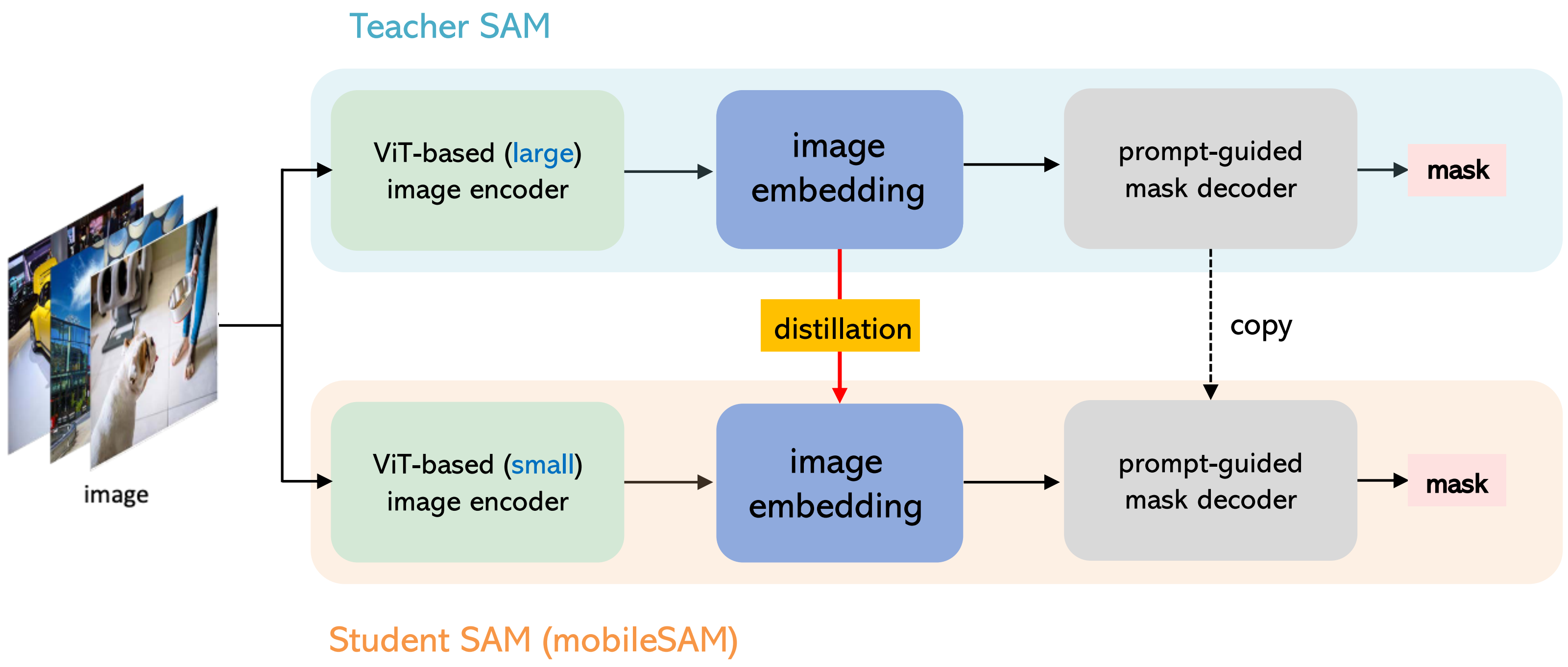
- ViT-H 기반 SAM의 지식을 더 작은 이미지 인코더를 사용하는 SAM으로 전이
- 인코더 파라미터를 100배, 전체 파라미터를 60배 줄임



# MobileSAM

❖ Decoupled Distillation (Divide KD into two sub-tasks)

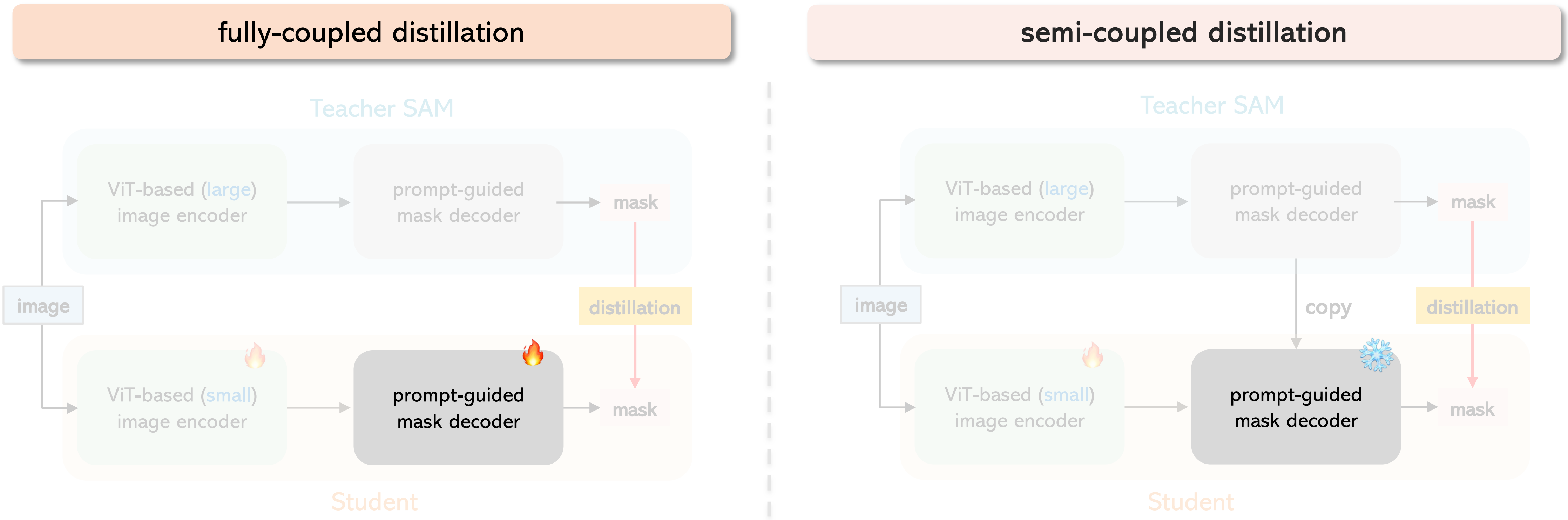
1) image encoder distillation
2) mask decoder finetuning



# MobileSAM

❖ Knowledge distillation (지식 증류)

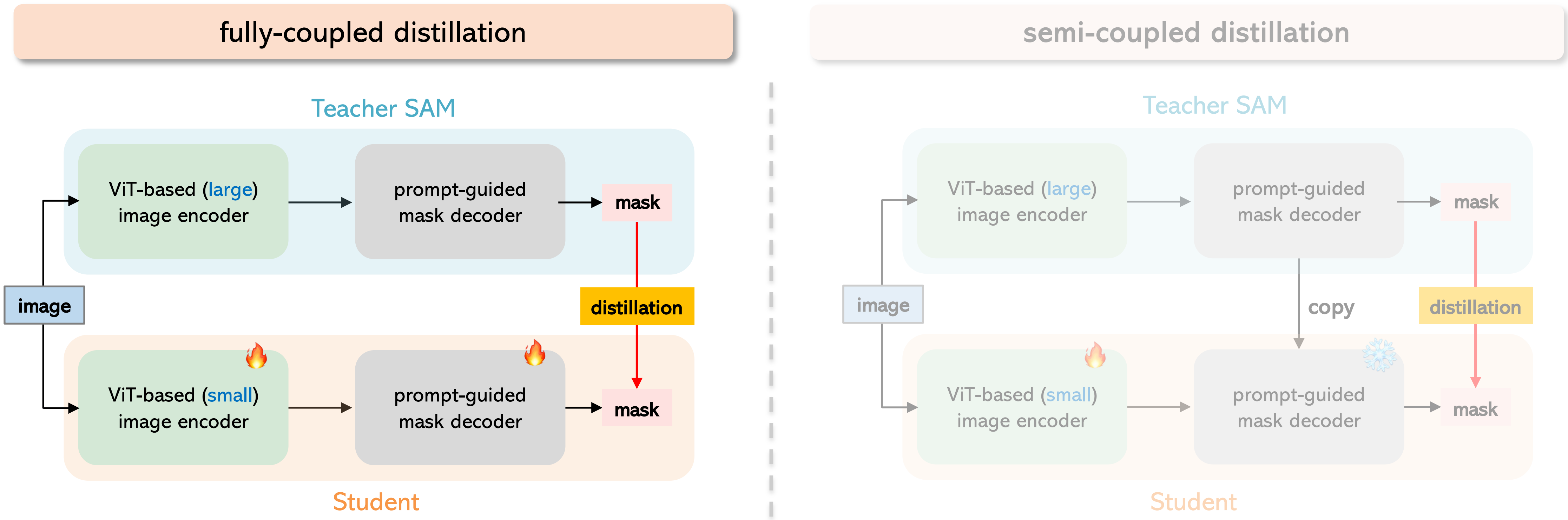
- **Coupled optimization of the image encoder and combined decoder**



# MobileSAM

❖ Fully-coupled distillation

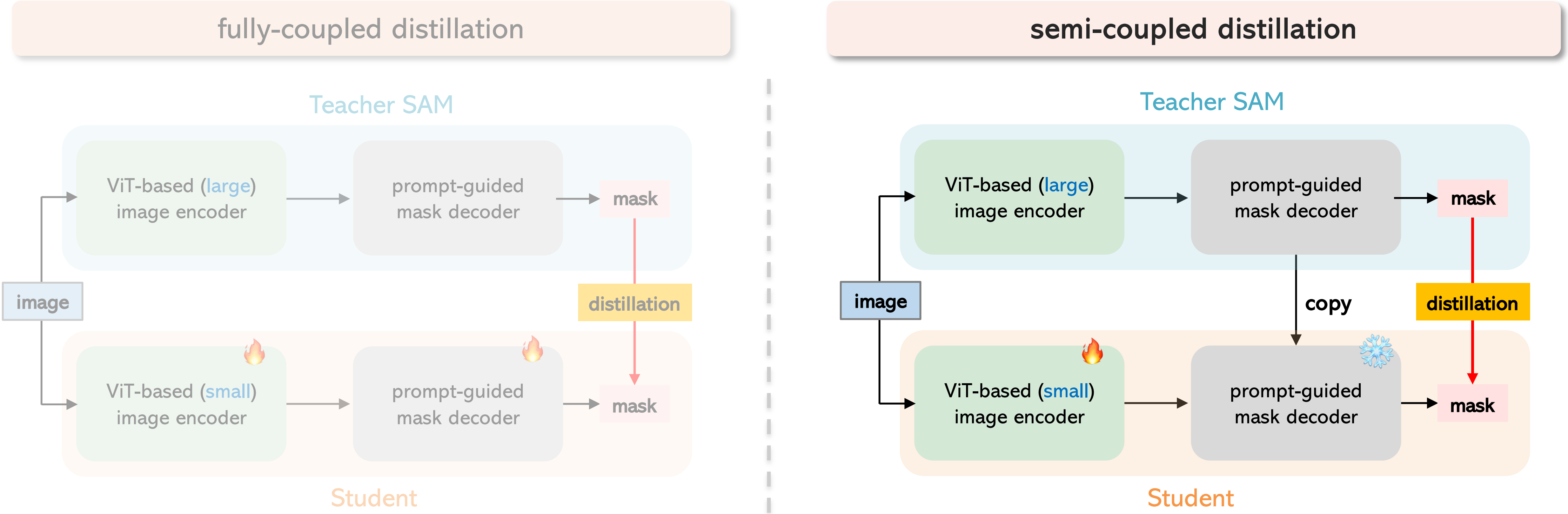
- Encoder와 decoder가 서로 의존적 → 두 모듈이 동시에 학습되기까지 시간이 오래 걸림
- Semi-coupled 방식으로 개선 시도



# MobileSAM

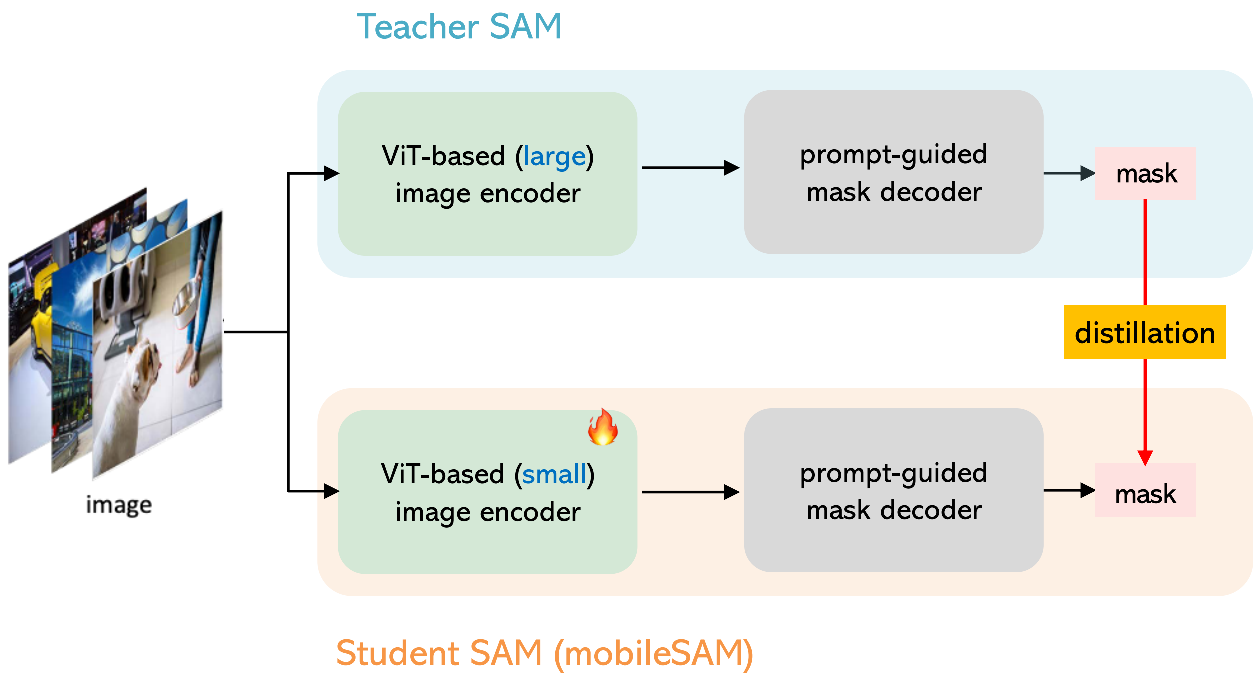
❖ Semi-coupled distillation

- Decoder가 흔들리지 않아 학습 안정성 증가
- 하지만, decoder의 출력이 prompt에 따라 달라짐 → 학습 과정에서 출력 불안정성 존재



# MobileSAM

❖ Coupled Distillation의 단점을 보완한 Decoupled Distillation !

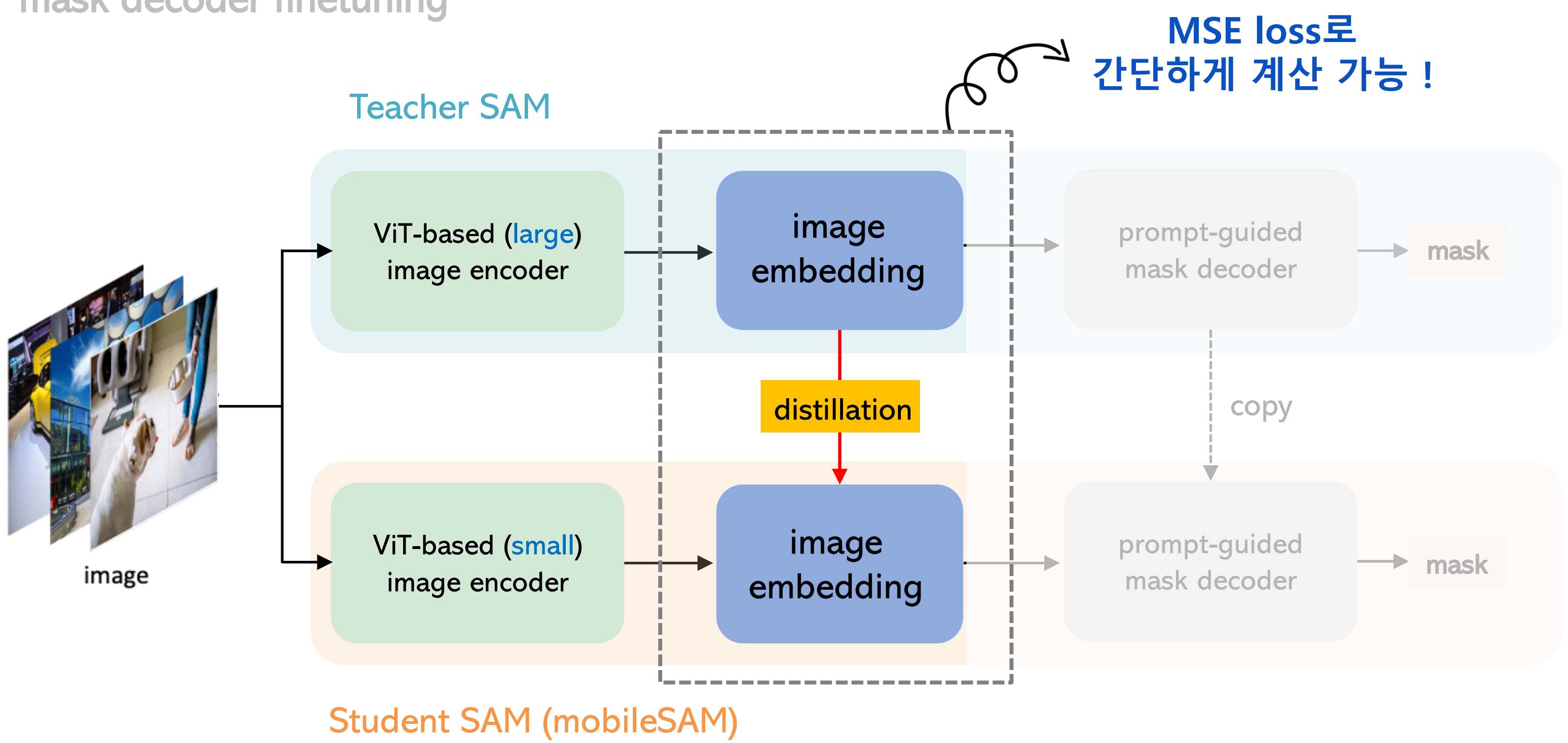



<Coupled Distillation>

# MobileSAM

❖ Decoupled Distillation (Divide KD into two sub-tasks)

1) image encoder distillation
2) mask decoder finetuning

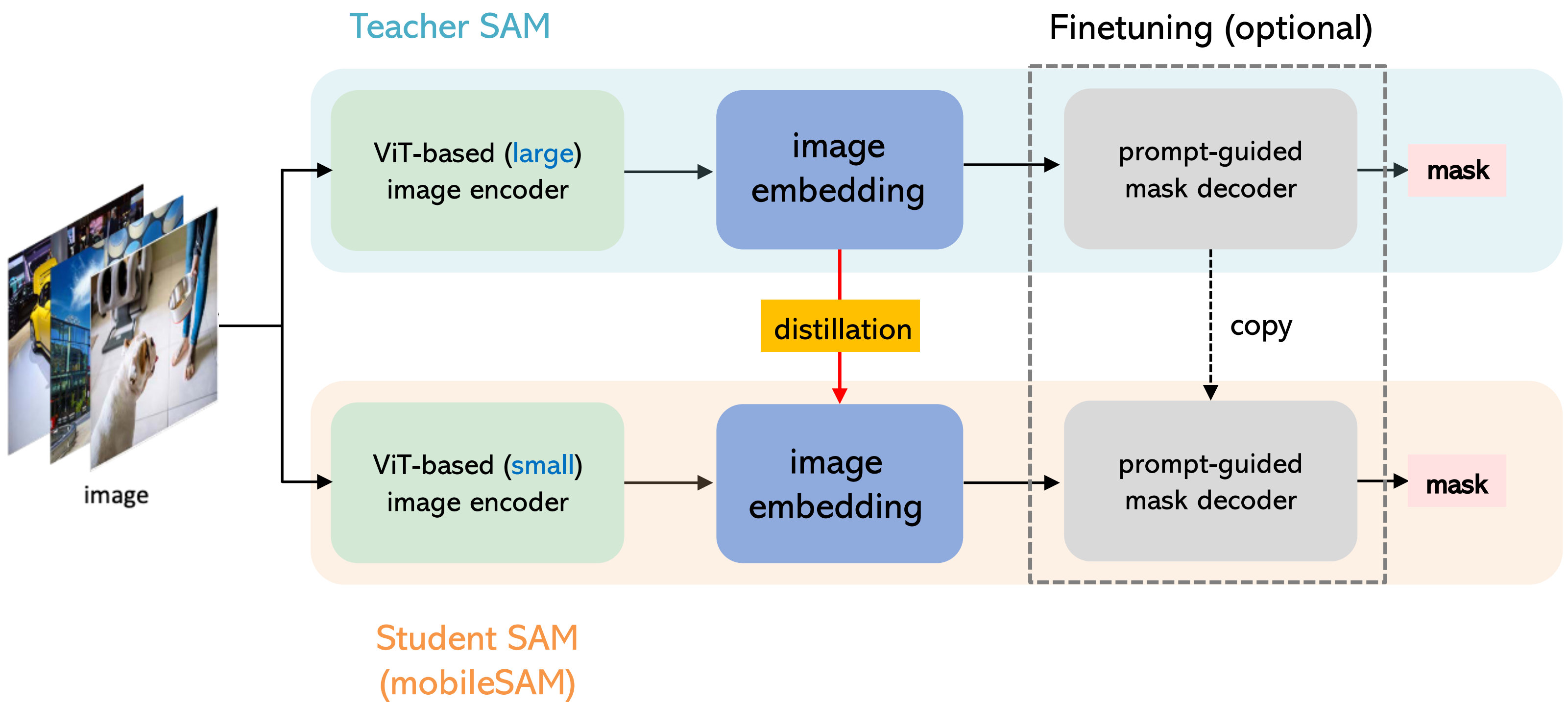


# MobileSAM

❖ **Decoupled Distillation (Divide KD into two sub-tasks)**

1) image encoder distillation

2) **mask decoder finetuning (optional)**



# MobileSAM

❖ **MobileSAM performs on par with the orignal SAM**

(a) Original SAM (b) MobileSAM (c) Original SAM (d) MobileSAM

Figure 5: Mask prediction with a box as the prompt.

# MobileSAM

## ❖ FastSAM vs MobileSAM

Table 6: Comparison between FastSAM and MobileSAM.

| | FastSAM | MobileSAM | Ratio |
|---|---|---|---|
| Size | 68M | 9.66M | ≈ 7 |
| Speed | 64ms | 12ms | ≈ 5 |

7 times samller
5 times faster

Table 7: mIoU comparison. With the assumption that the predicted mask from the original SAM is ground-truth, a higher mIoU indicates a better performance.

| | 100 | 200 | 300 | 400 | 500 |
|---|---|---|---|---|---|
| FastSAM | 0.27 | 0.33 | 0.37 | 0.41 | 0.41 |
| MobileSAM | 0.73 | 0.71 | 0.74 | 0.73 | 0.73 |

superior performance

more suitable for mobile applications

# EfficientSAM

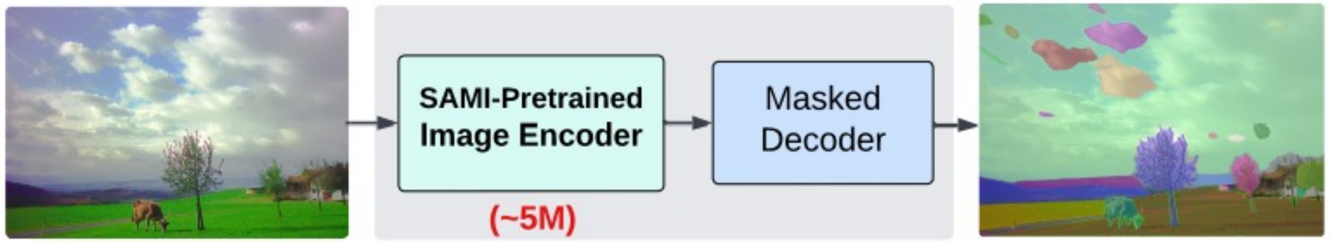
# EfficientSAM

❖ EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything

- Meta AI Research (2024, CVPR)
- 인용수: 153회

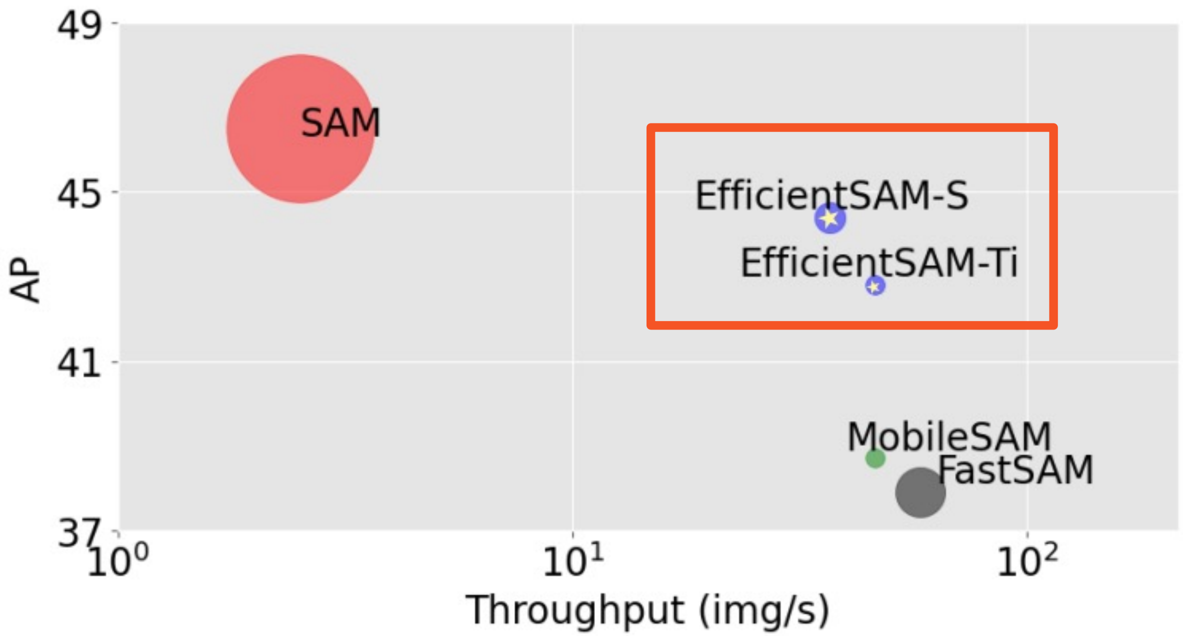
**EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything**

Yunyang Xiong, Bala Varadarajan*, Lemeng Wu*, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, Raghuraman Krishnamoorthi, Vikas Chandra
Meta AI Research



# EfficientSAM

❖ **EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything**

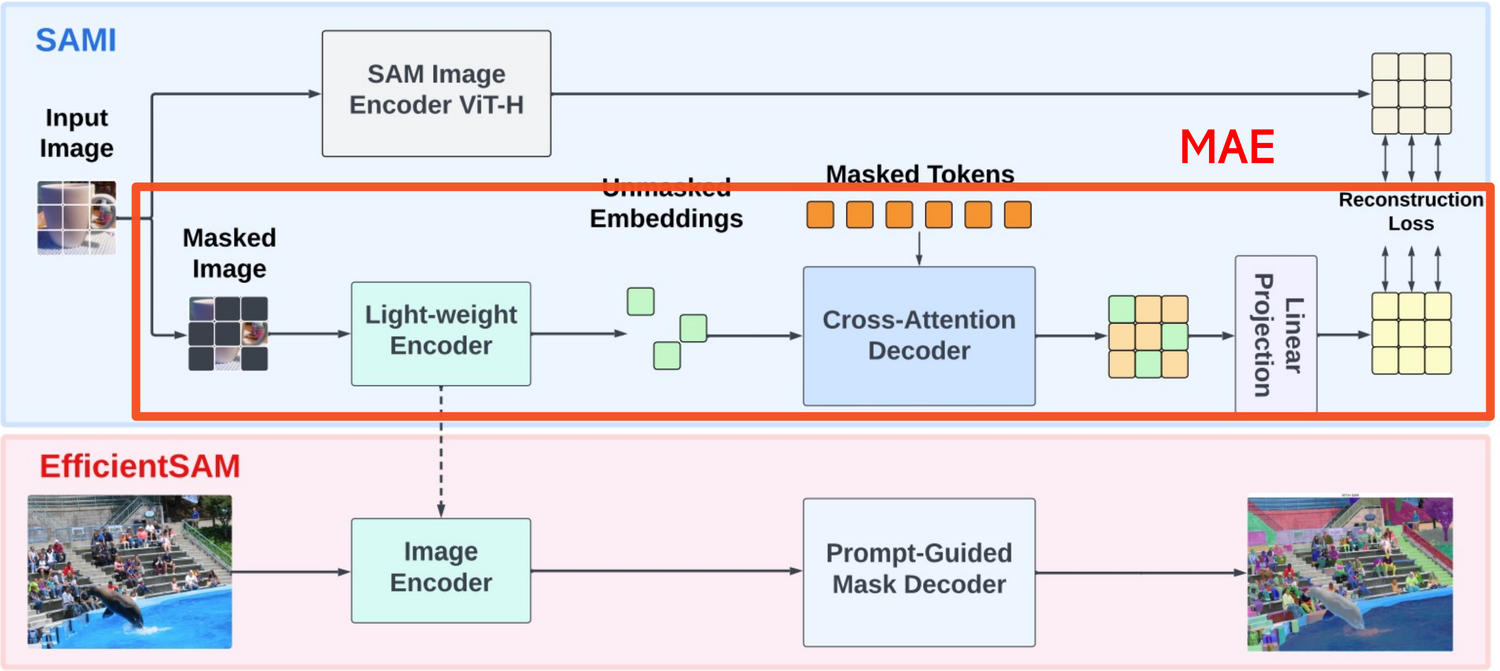


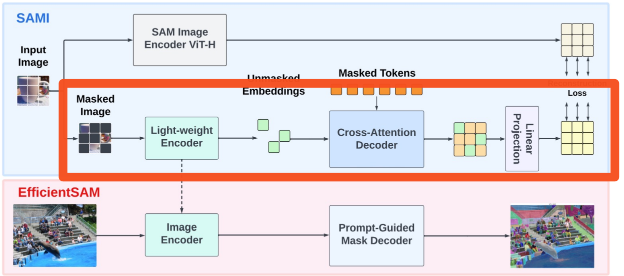
# EfficientSAM

❖ Two stage framework

Stage 1
SAMI pretraining

Stage 2
SAM finetuning

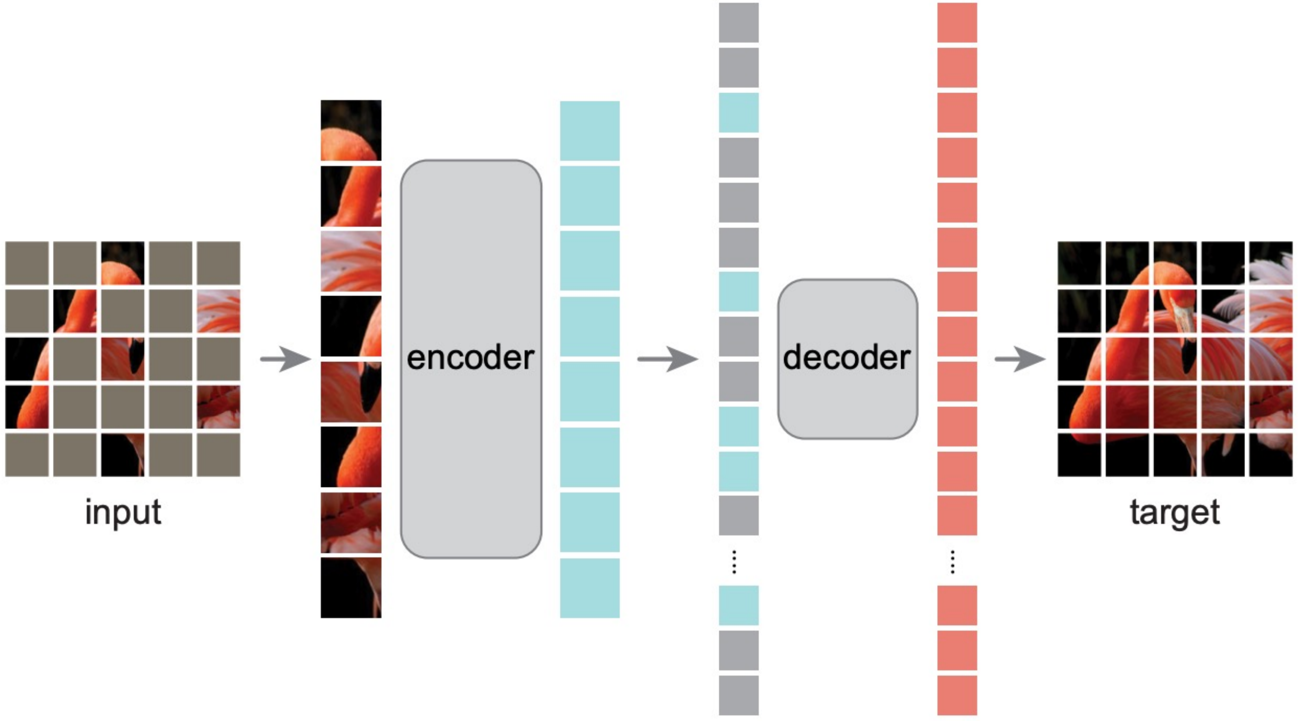


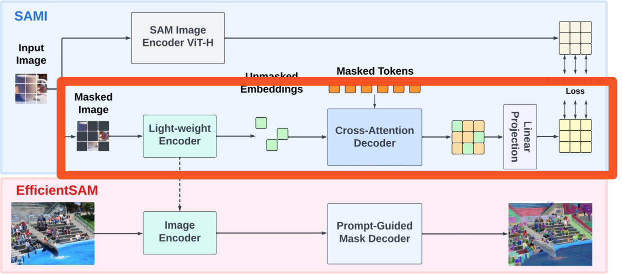



# EfficientSAM

❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners






# EfficientSAM

## ❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners

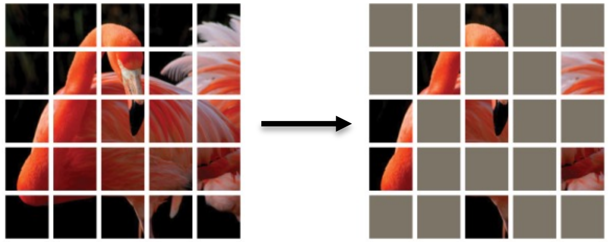
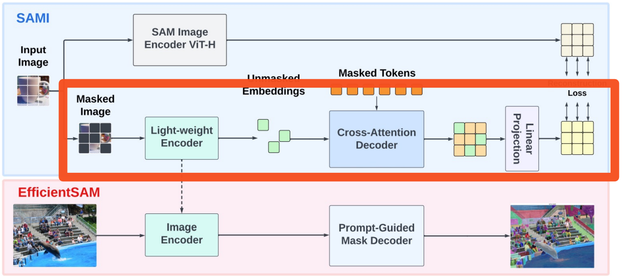


image masked image

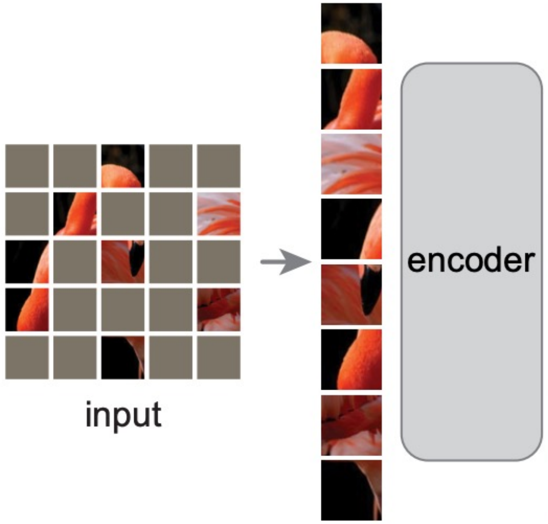
visibile patch mask token

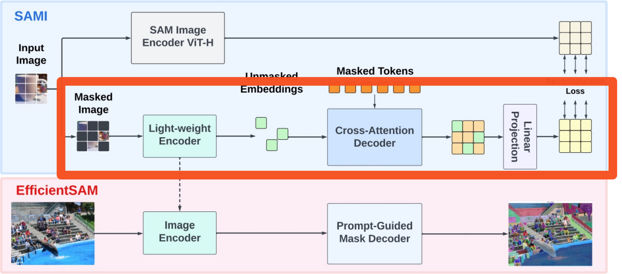



# EfficientSAM

## ❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners

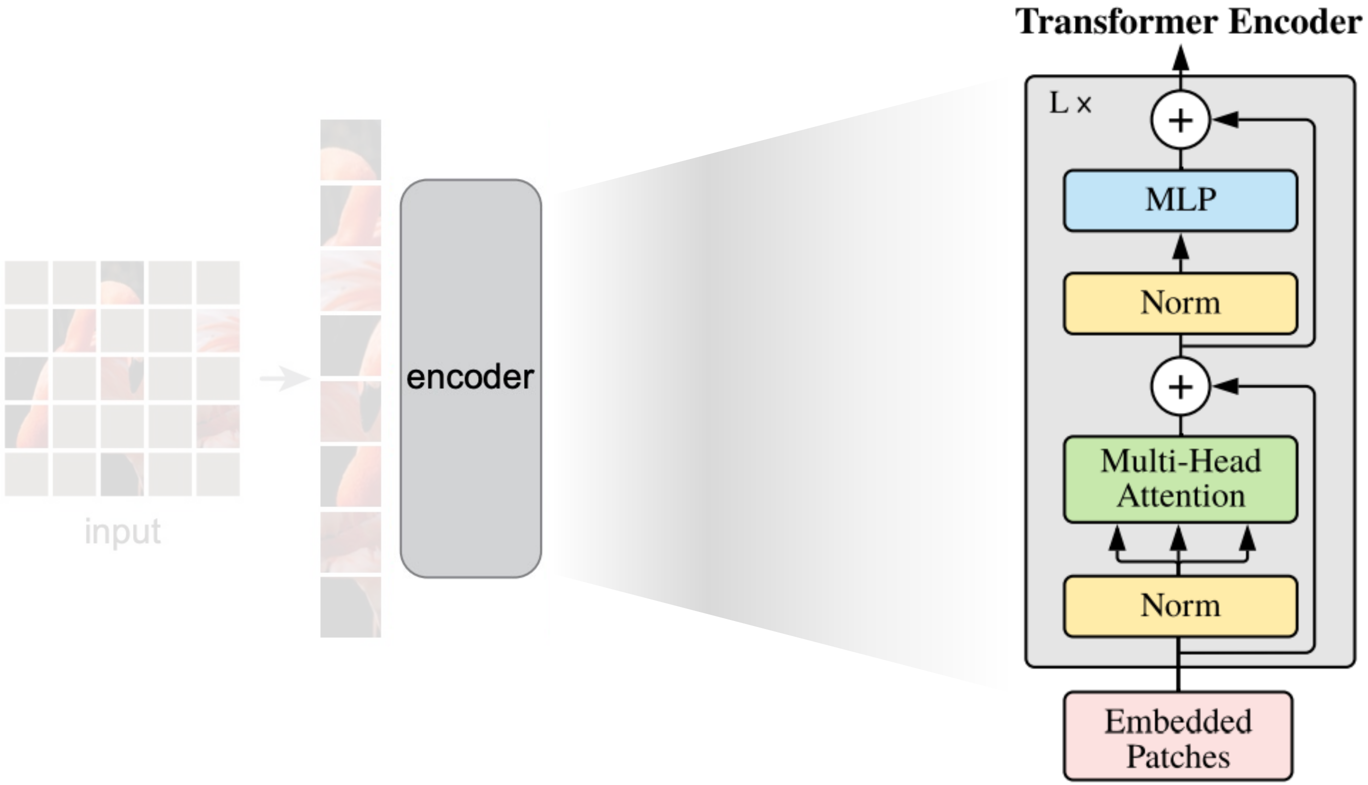


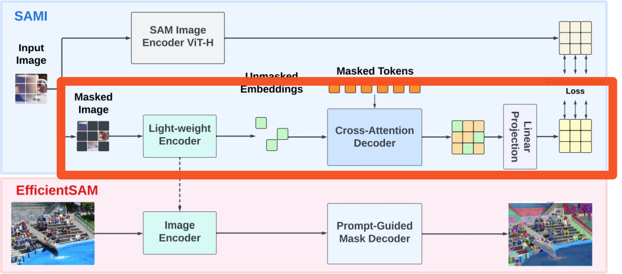

# EfficientSAM

## ❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners



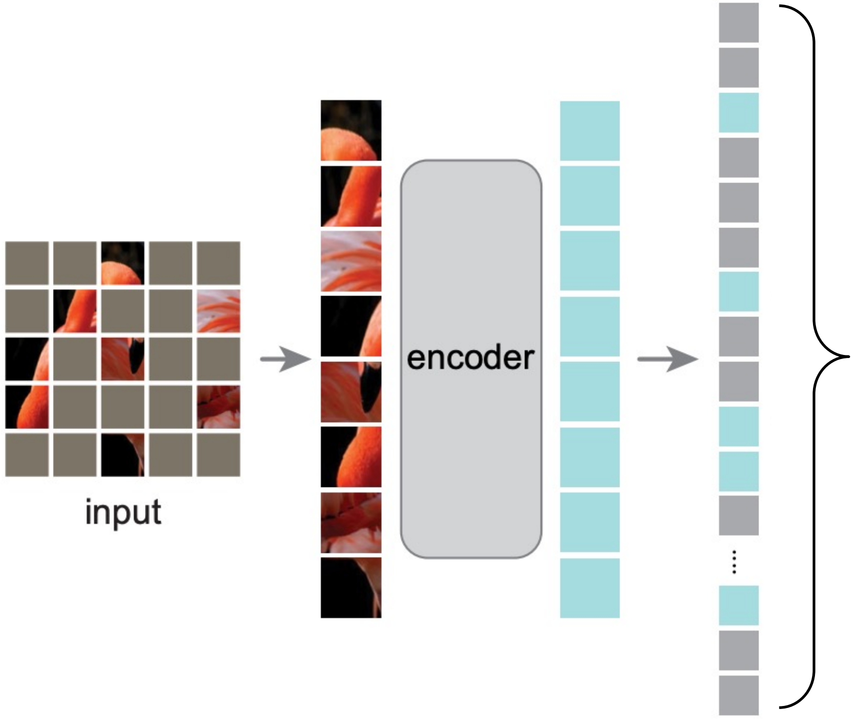

ViT




# EfficientSAM

## ❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners

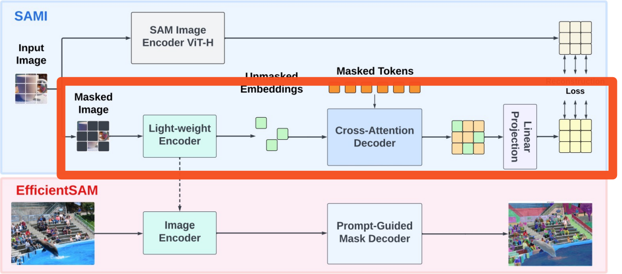



Positional embedding

encoded visible patch

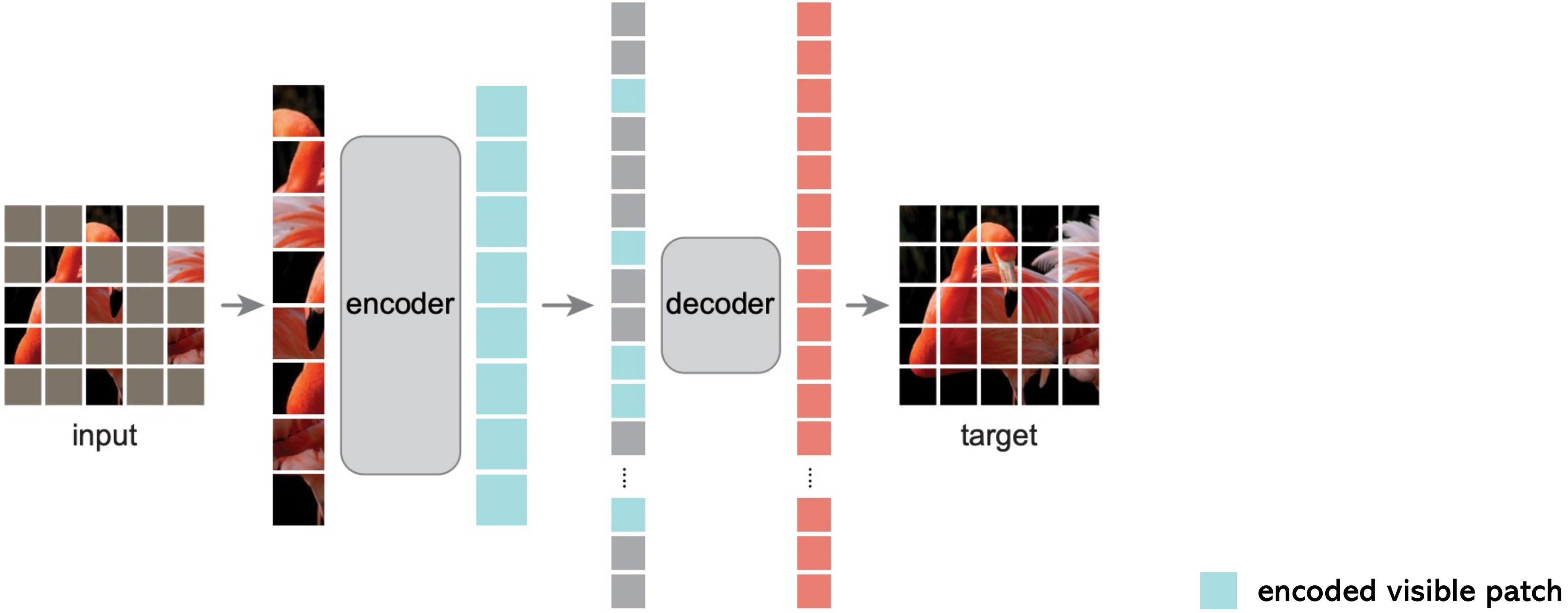
mask token

# EfficientSAM




## ❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners

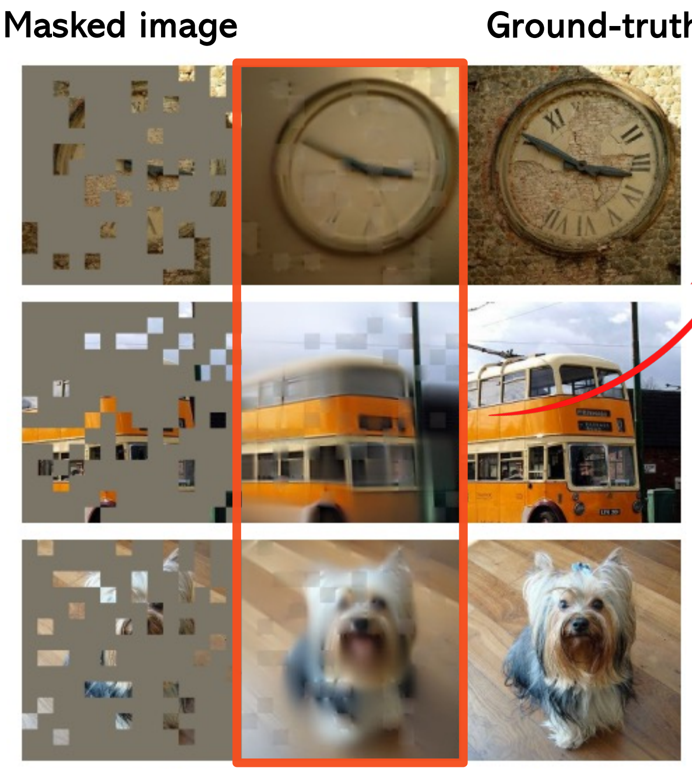



**decoder에서만 mask token을 사용하는 비대칭 구조**

# EfficientSAM

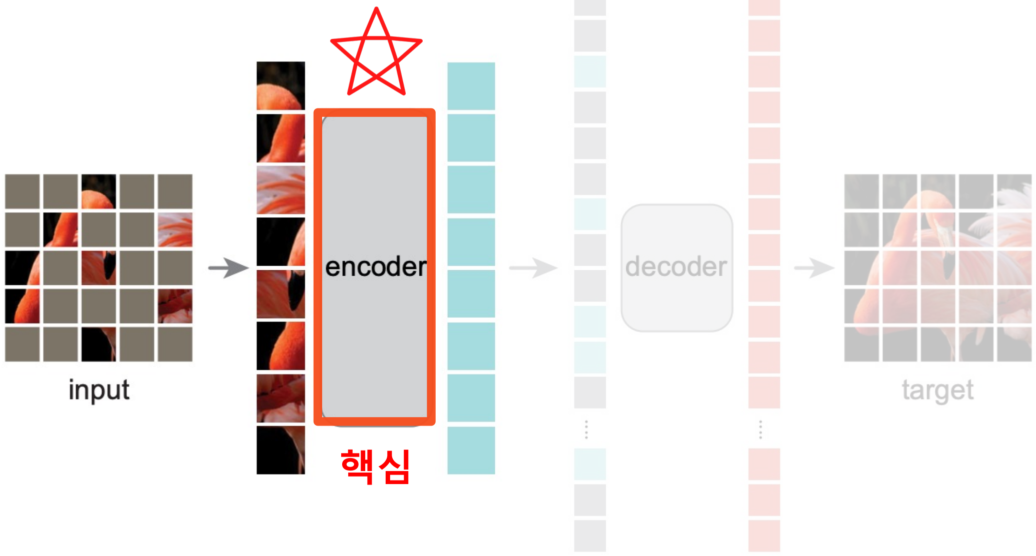
## ❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners




MAE reconstruction

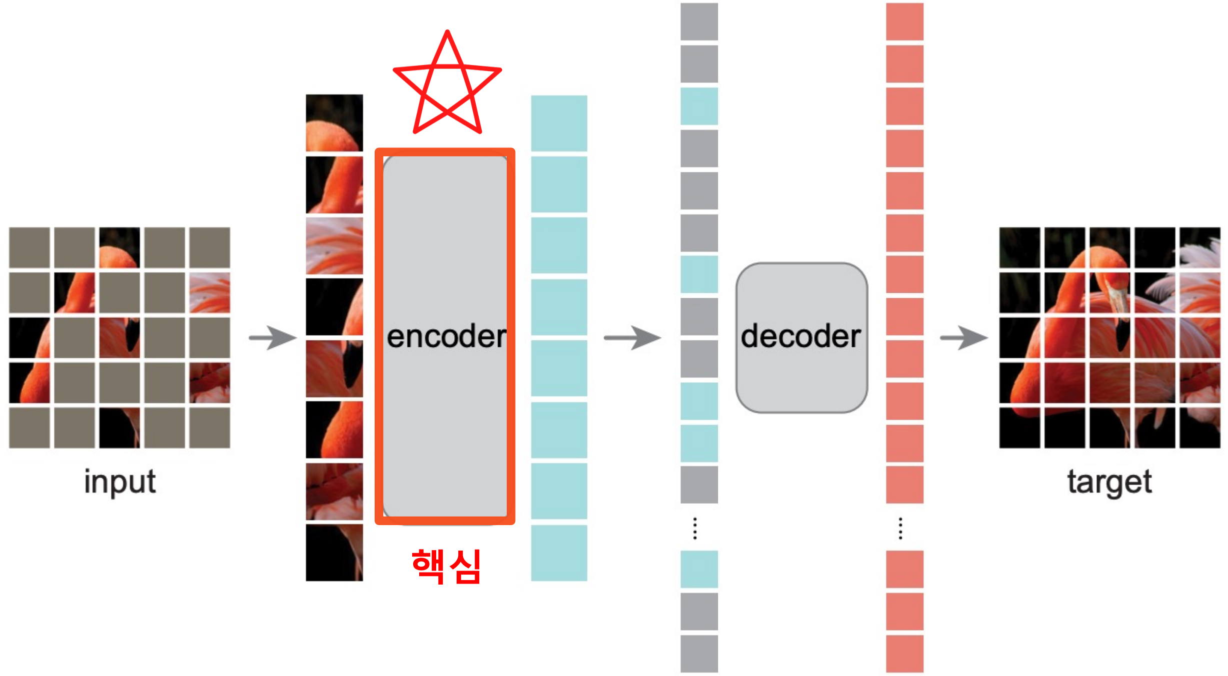
Blurry ..



# EfficientSAM

## ❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners

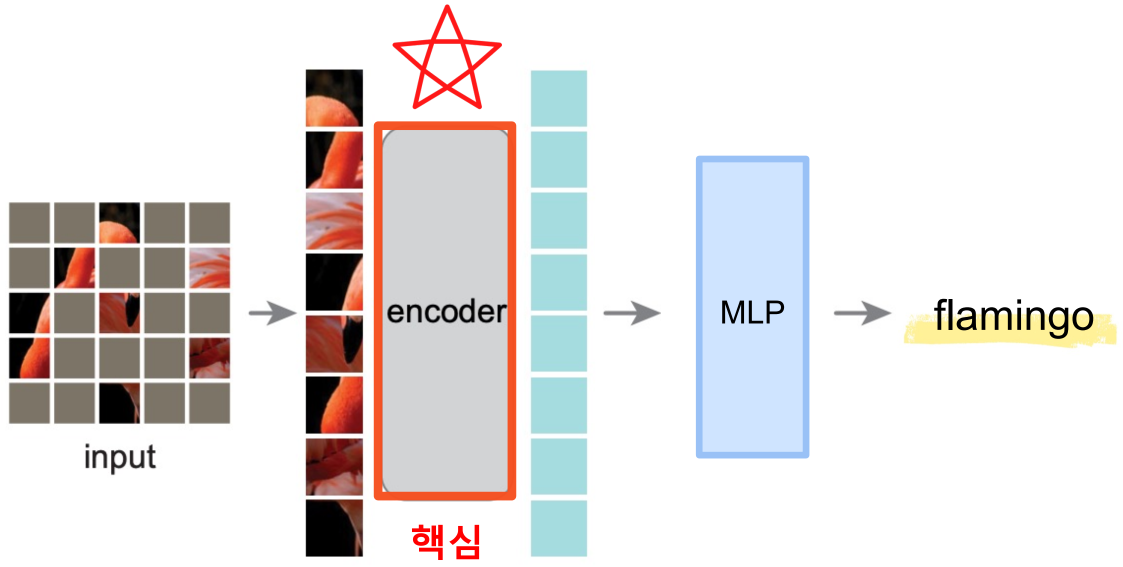


# EfficientSAM

❖ MAE Architecture

Downstream Task

ex) Classification

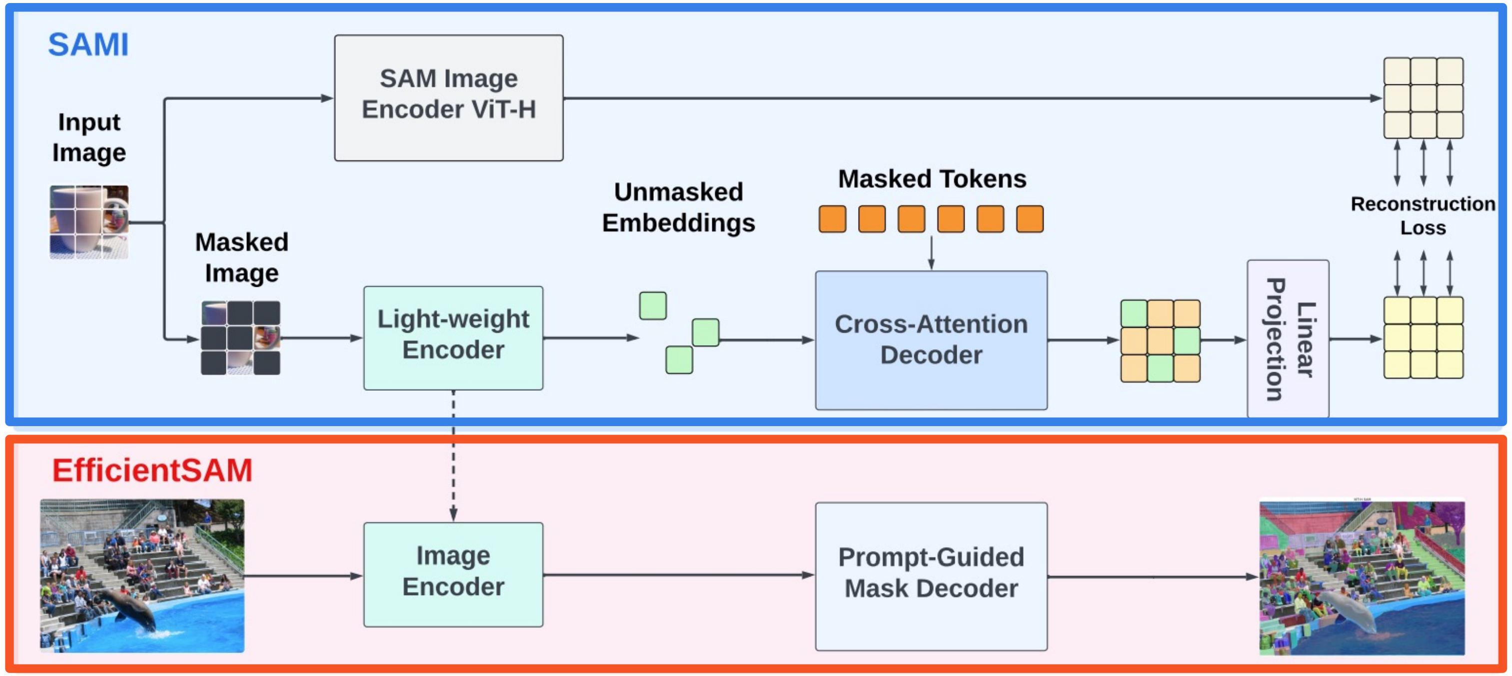


# EfficientSAM

❖ 2-Stage

**Stage 1**
SAMI pretraining

**Stage 2**
SAM finetuning




MAE

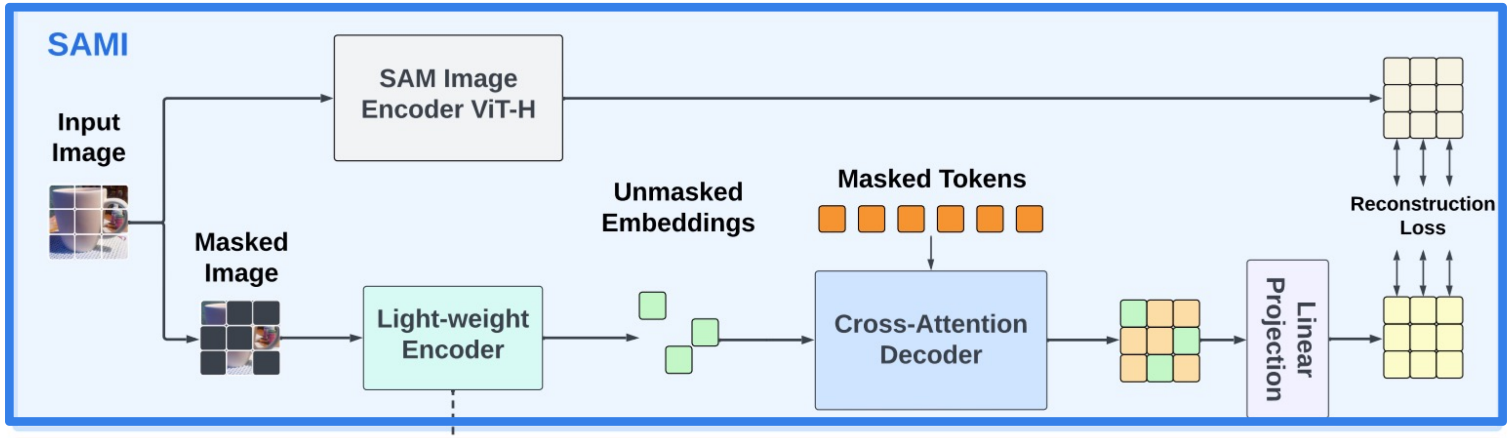
Downstream task

# EfficientSAM

❖ SAMI pretraining (Stage 1)

- 기존 SAM 모델의 거대한 Vit-H 인코더를 직접 사용하는 대신, MAE를 적용해 경량화된 인코더를 학습하는 과정
- SAMI: SAM-leveraged masked image pretraining

Stage 1
SAMI pretraining

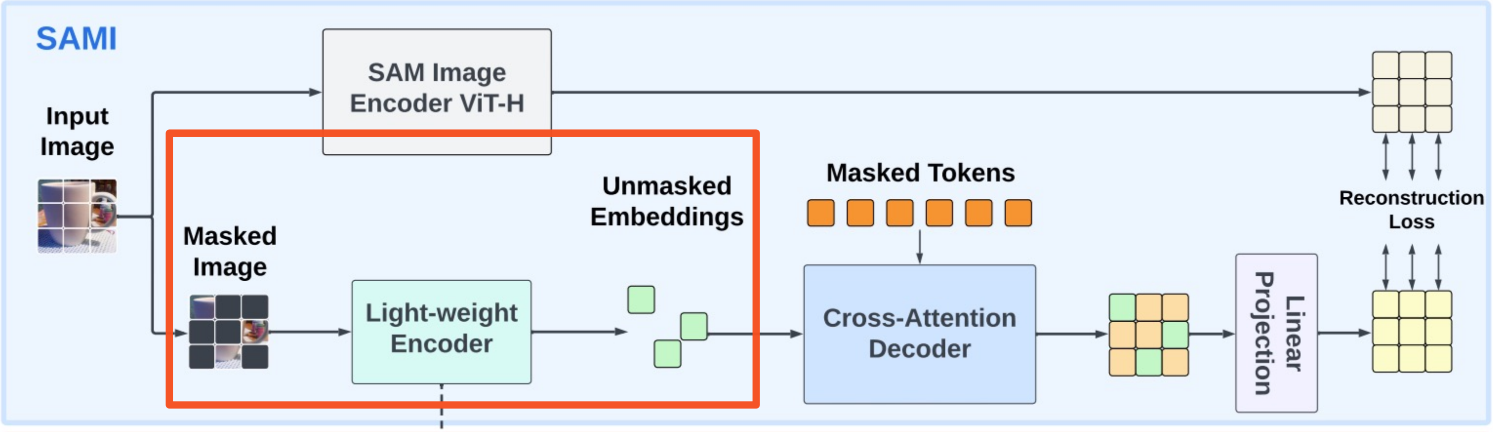


# EfficientSAM

❖ SAMI pretraining (Stage 1)

- 기존 SAM 모델의 거대한 Vit-H 인코더를 직접 사용하는 대신, MAE를 적용해 경량화된 인코더를 학습하는 과정
- SAMI: SAM-leveraged masked image pretraining

Stage 1
SAMI pretraining



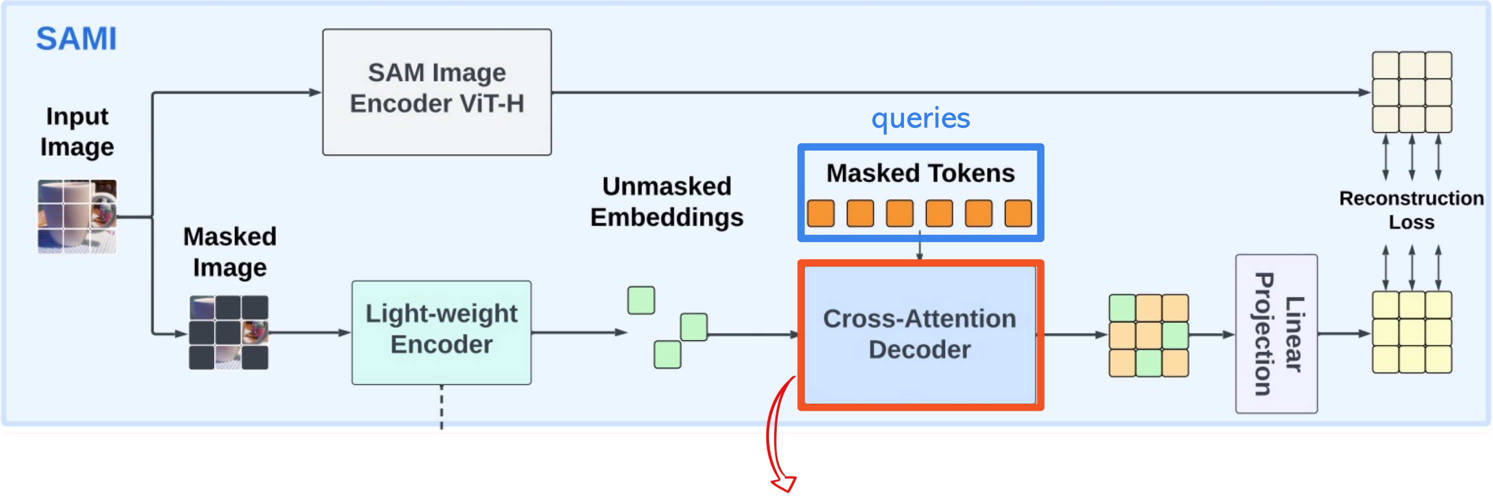

Encoder의 input:
unmasked patch

# EfficientSAM

❖ SAMI pretraining (Stage 1)

- 기존 SAM 모델의 거대한 Vit-H 인코더를 직접 사용하는 대신, MAE를 적용해 경량화된 인코더를 학습하는 과정
- SAMI: SAM-leveraged masked image pretraining

Stage 1
SAMI pretraining



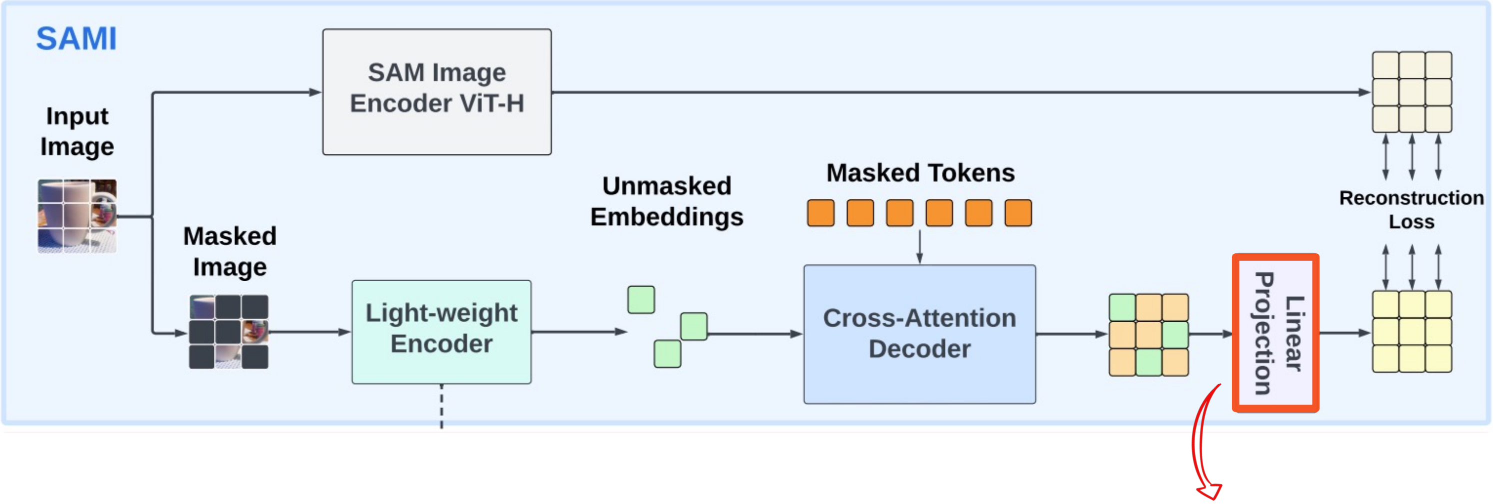

Queries: masked tokens
Keys & values: unmasked features from encoder & masked features

# EfficientSAM

❖ SAMI pretraining (Stage 1)

- 기존 SAM 모델의 거대한 Vit-H 인코더를 직접 사용하는 대신, MAE를 적용해 경량화된 인코더를 학습하는 과정
- SAMI: SAM-leveraged masked image pretraining

Stage 1
SAMI pretraining



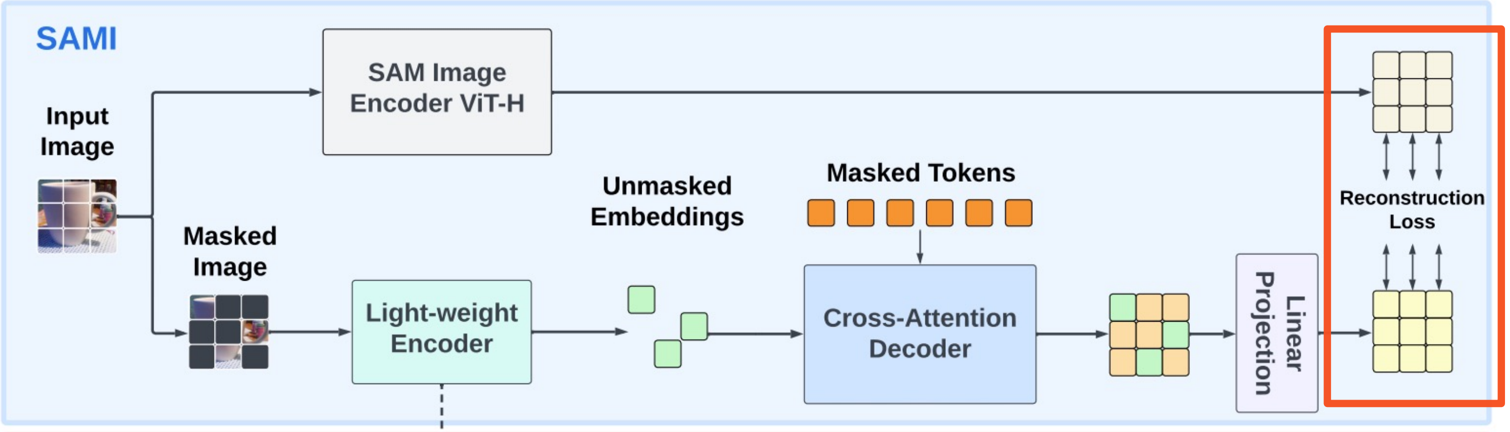

Linear Projection:
SAM image encoder & MAE output
**Feature dimension mismatch 해결**

# EfficientSAM

❖ SAMI pretraining (Stage 1)

- 기존 SAM 모델의 거대한 Vit-H 인코더를 직접 사용하는 대신, MAE를 적용해 경량화된 인코더를 학습하는 과정
- SAMI: SAM-leveraged masked image pretraining

Stage 1
SAMI pretraining




$L_{W_e,W_d,W_\theta}$

$$L_{W_e,W_d,W_\theta} = \frac{1}{N} \cdot \sum_{j=1}^{N} ||f^{\text{sam}}(\mathbf{x}) - f^{h}(\mathbf{x})||^2,$$

where $N$ is the number of input tokens

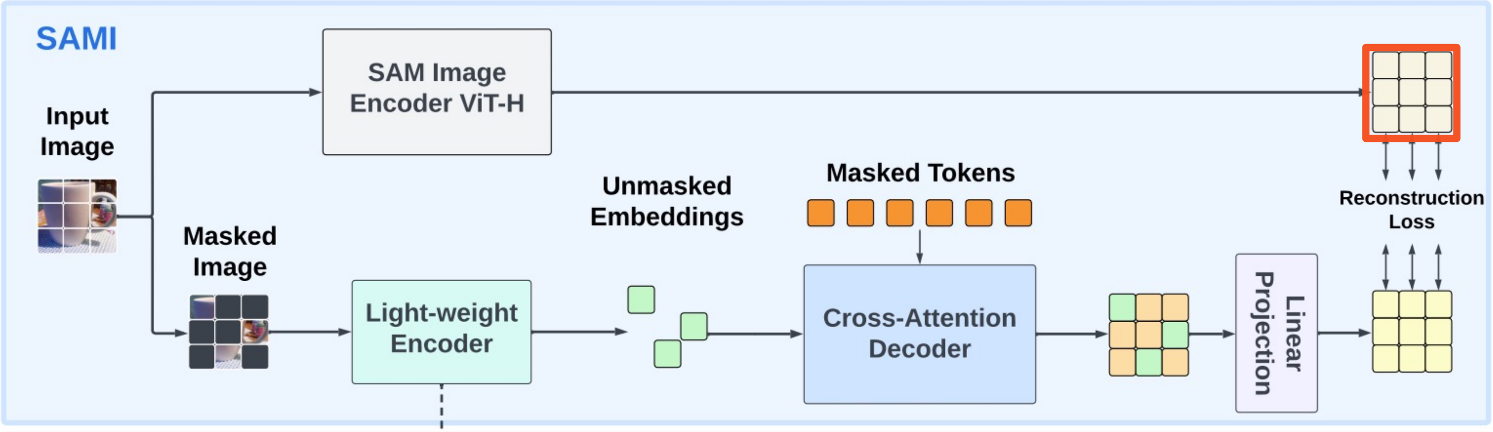
# EfficientSAM

## ❖ SAMI pretraining (Stage 1)

- SAM의 encoder에서 추출한 고차원적인 특징 (Feature Embeddings)을 복원하는 방식

**Stage 1**
**SAMI pretraining**

강력한 표현력을 가진 특징




원본 이미지

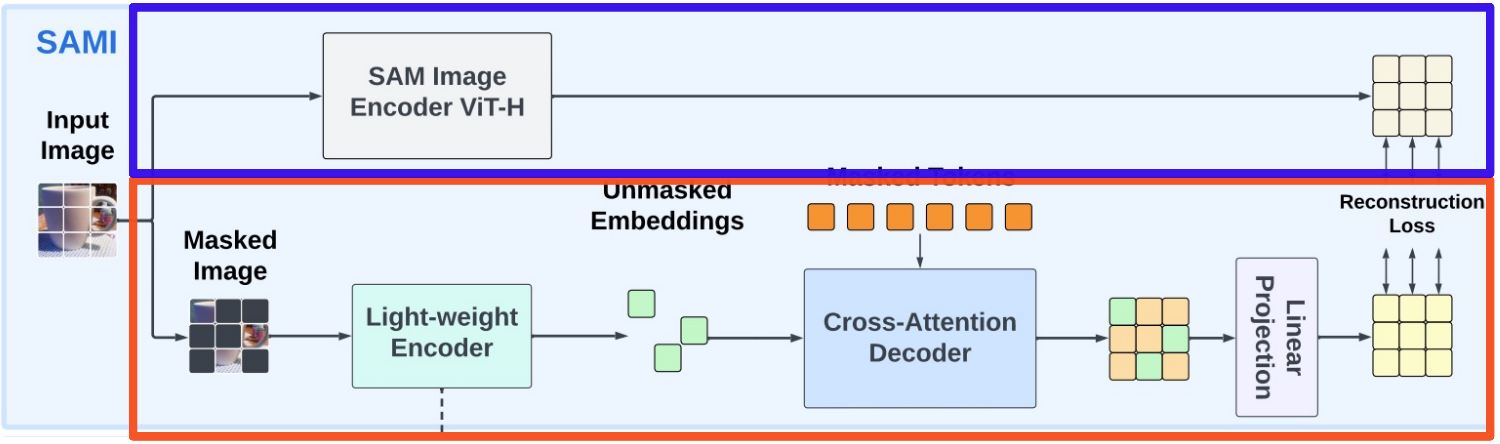
# EfficientSAM

❖ SAMI pretraining (Stage 1)

- SAM의 encoder에서 추출한 고차원적인 특징 (Feature Embeddings)을 복원하는 방식

**Stage 1**
SAMI pretraining

feedforward



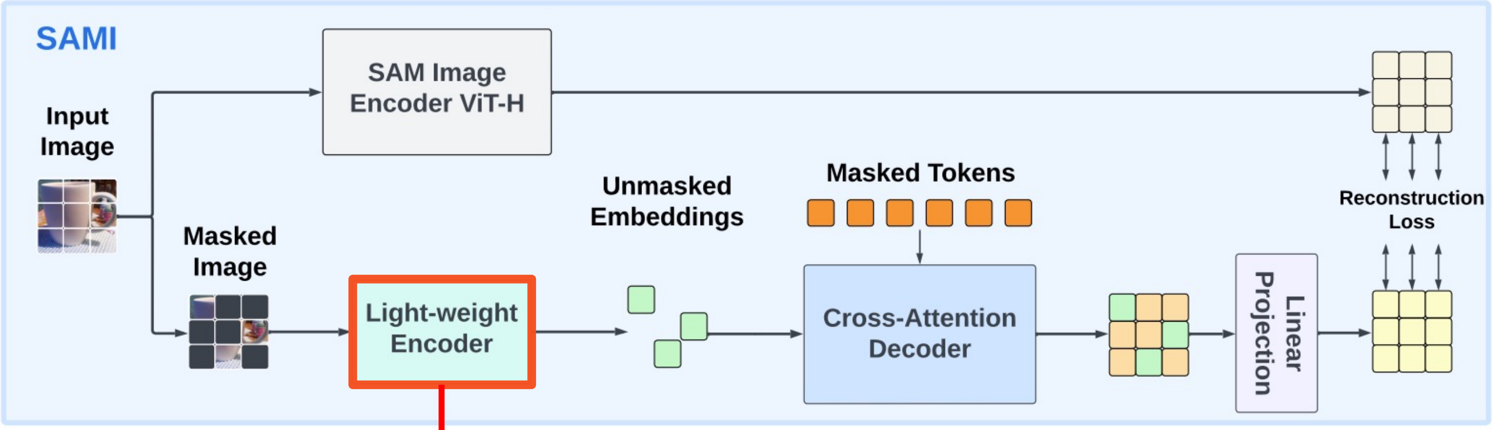

feedforward
& backpropagation

# EfficientSAM

❖ SAMI pretraining (Stage 1)

- SAM의 encoder에서 추출한 고차원적인 특징 (Feature Embeddings)을 복원하는 방식

Stage 1
SAMI pretraining



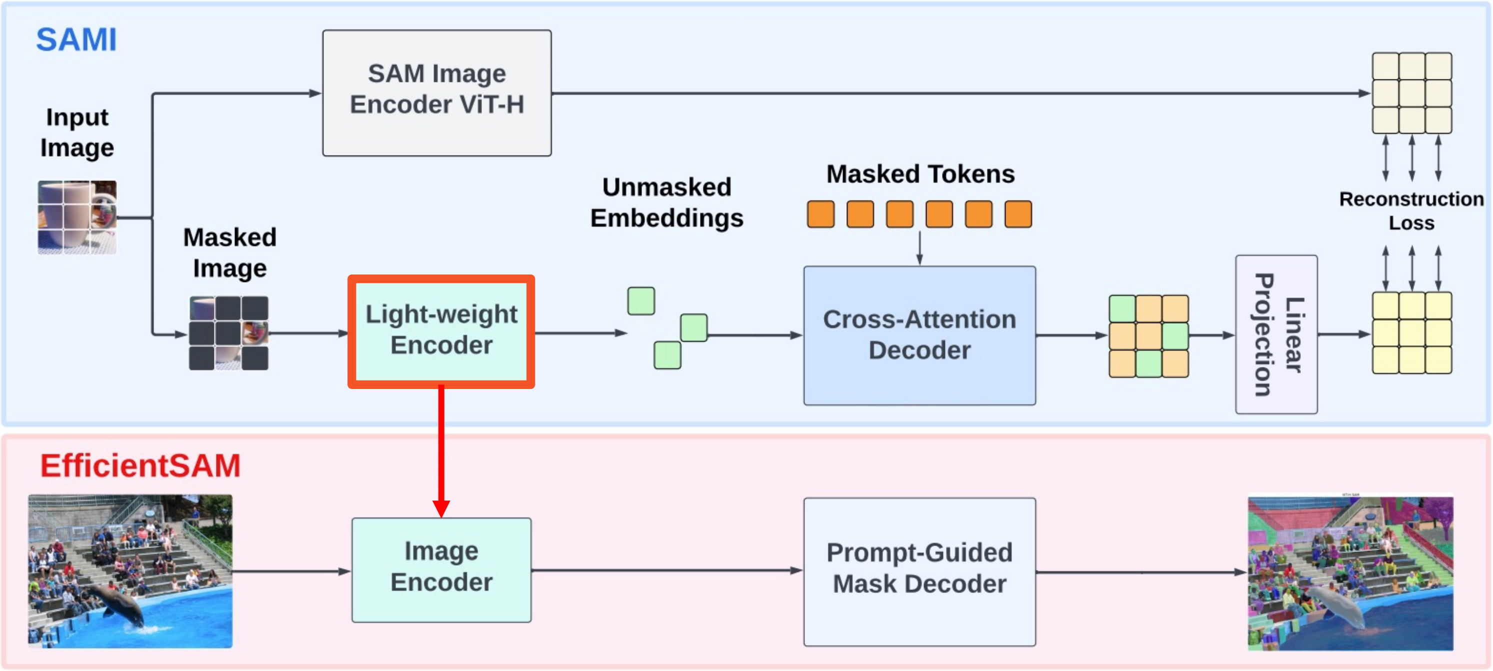

extract feature representations
for various vision task

# EfficientSAM

❖ SAM finetuning (Stage 2)

- 경량화된 인코더를 SAM의 mask decoder와 결합하여 segmentation 작업 수행하도록 finetuning

Stage 2
SAM finetuning

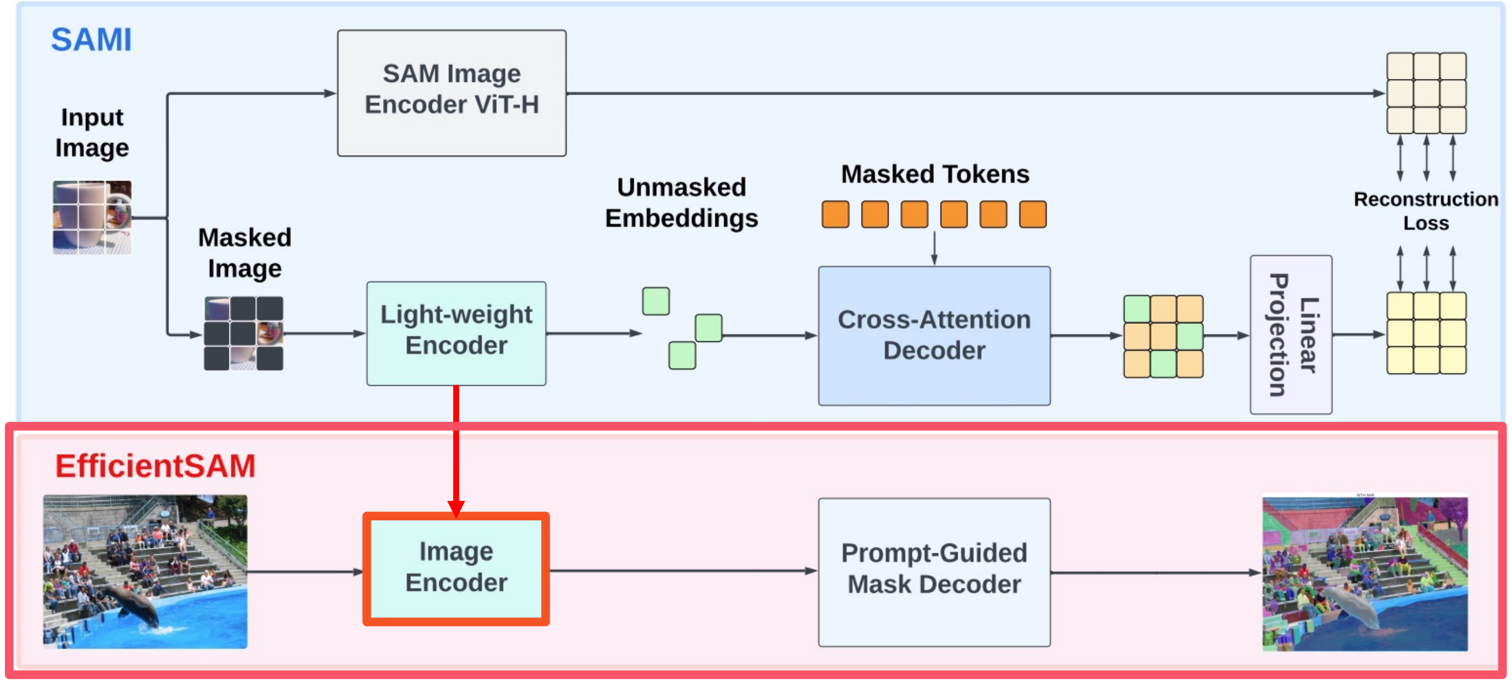


# EfficientSAM

❖ SAM finetuning (Stage 2)

- 경량화된 인코더를 SAM의 mask decoder와 결합하여 segmentation 작업 수행하도록 finetuning

Stage 2
SAM finetuning




finetune on SA-1B dataset

# Main Results

❖ Zero-shot single point calid mask evaluation results

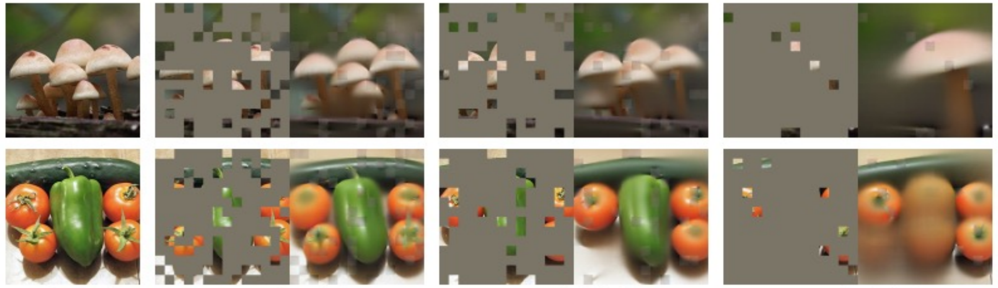
- Only underperforms SAM by 1.5 mIOU

| Method | COCO | | | LVIS | | |
|---|---|---|---|---|---|---|
| | box | 1 click | 3 click | box | 1 click | 3 click |
| SAM[31] | 78.4 | 55.6 | 74.1 | 78.9 | 59.8 | 75.2 |
| MobileSAM[68] | 74.2 | 43.7 | 59.7 | 73.8 | 51.0 | 54.4 |
| SAM-MAE-Ti[31] | 74.7 | 43.3 | 65.8 | 73.8 | 50.6 | 65.3 |
| EfficientSAM-Ti (ours) | 75.7 | 45.5 | 67.2 | 74.3 | 52.7 | 66.8 |
| EfficientSAM-S (ours) | 76.9 | 50.0 | 69.8 | 75.4 | 56.2 | 68.7 |

# Main Results

## ❖ Ablation Studies

Masking ratio 75%가 가장 적절




| Mask Ratio | 50% | 75% | 85% |
|---|---|---|---|
| Top-1 Acc.(%) | 84.6 | **84.8** | 84.7 |

Table 7. Ablation on the mask ratio for SAMI-B on ImageNet-1K.

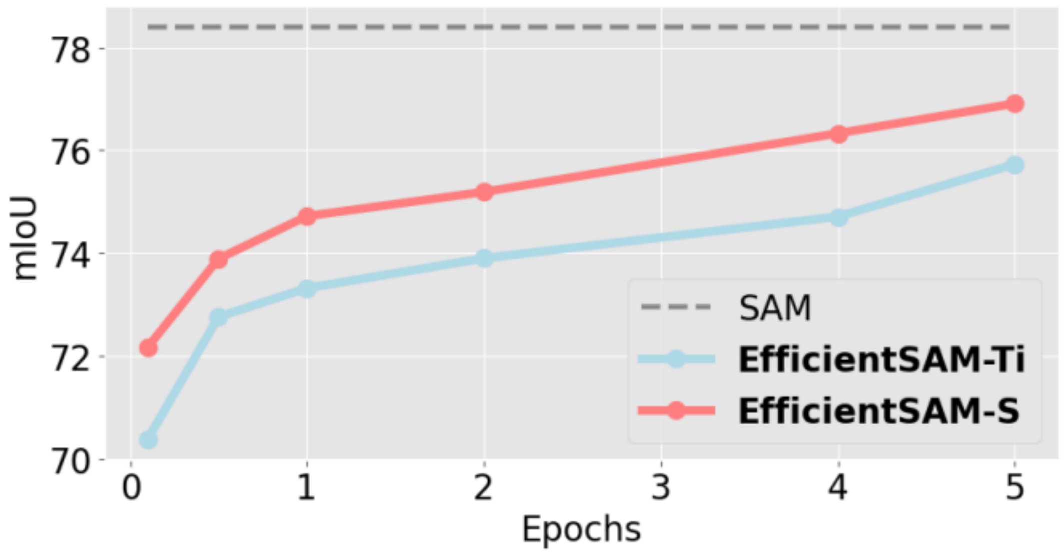
**적은 정보**만을 가지고도 더 **강력한 일반화 능력**을 갖도록 훈련되며,
과적합을 방지하고 효율적인 **특징 표현**을 학습

# Main Results

❖ **Ablation Studies**

**빠른 성능 향상**

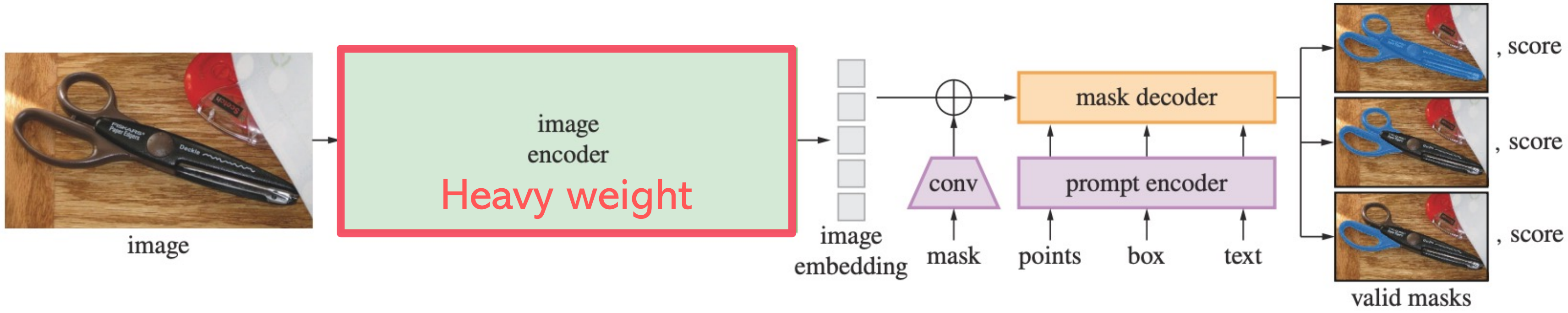
Advantages of SAMI-pretrained image encoders




EfficientSAM 모델이 SAMI를 활용한 사전 학습된 이미지 인코더 덕분에
빠르게 좋은 성능을 발휘할 수 있음

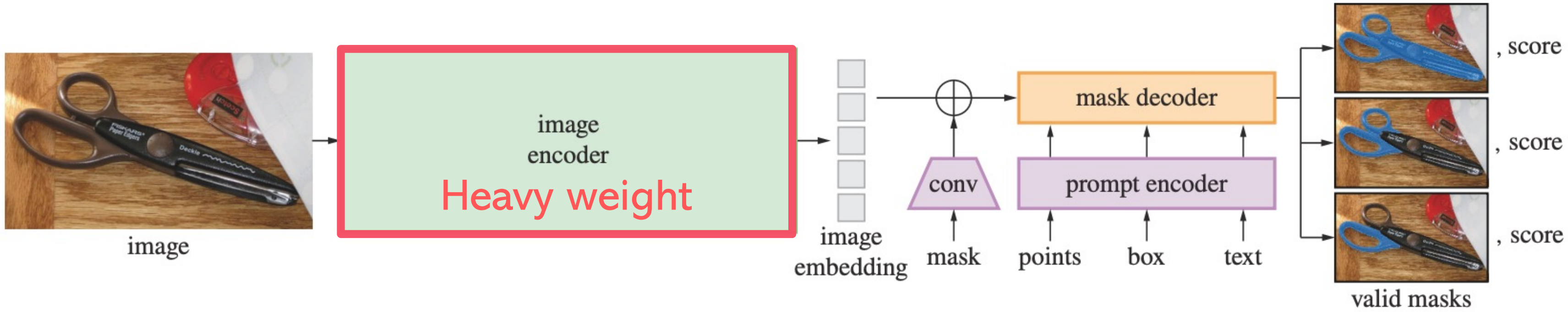
# Conclusion

## Segment Anything

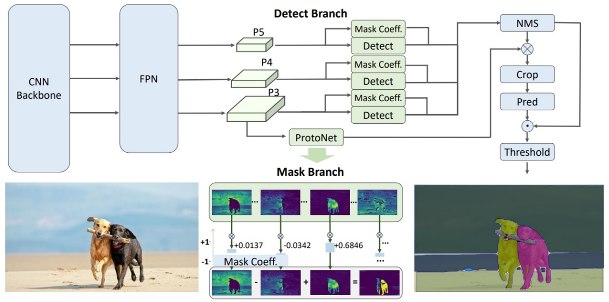


# Conclusion

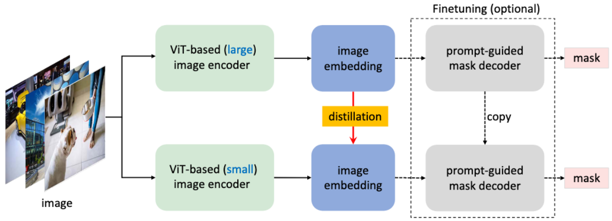
## Segment Anything



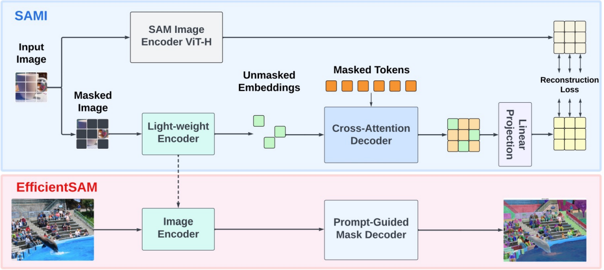

## FastSAM




## MobileSAM




## EfficientSAM



# Thank you