
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

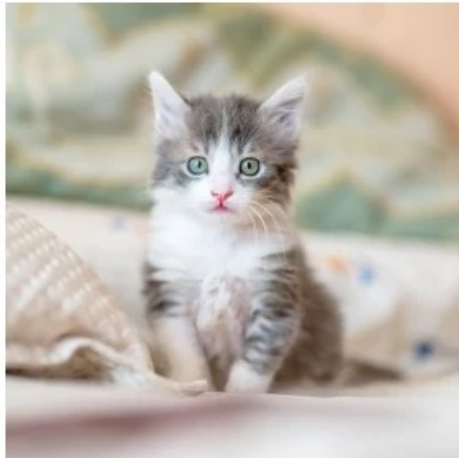
- hyeseunglee@korea.ac.kr

Introduction



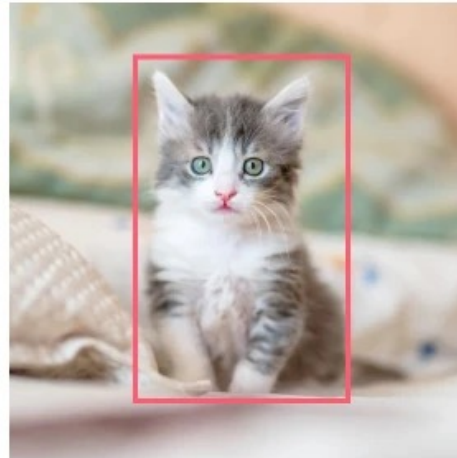
Image Segmentation

Classification



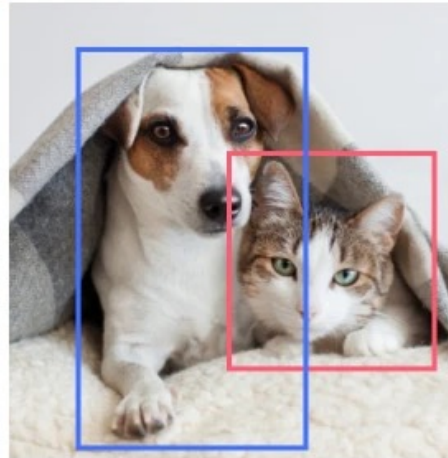
Cat

Classification + Localization



Cat

Object Detection



Cat, Dog

Segmentation



Cat, Dog



픽셀 기반의 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



“ChatGPT of the image segmentation field”

Segment Anything

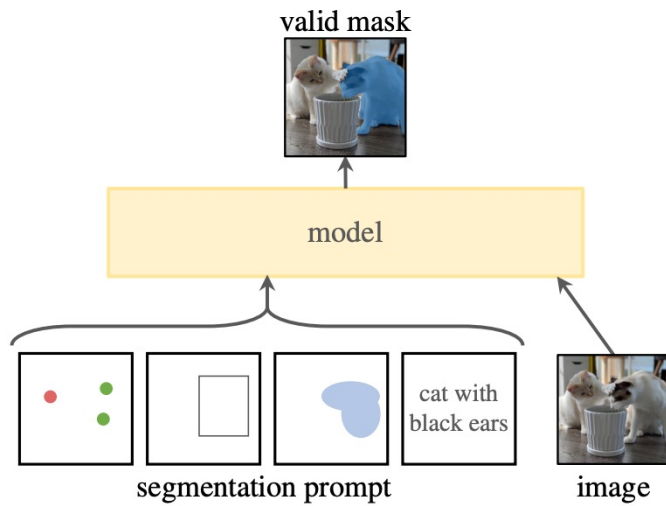
Alexander Kirillov^{1,2,4} Eric Mintun² Nikhila Ravi^{1,2} Hanzi Mao² Chloe Rolland³ Laura Gustafson³
 Tete Xiao³ Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár⁴ Ross Girshick⁴
¹project lead ²joint first author ³equal contribution ⁴directional lead

Meta AI Research, FAIR

Segment Anything

❖ Segment Anything (SAM)

- **Prompt-guided** Vision foundation model



Segment Anything

Research by Meta AI



Tools

Upload Gallery

Hover & Click

Box
Roughly draw a box around an object.

+ Add Mask - Remove Area

Reset Undo Redo

Cut out object

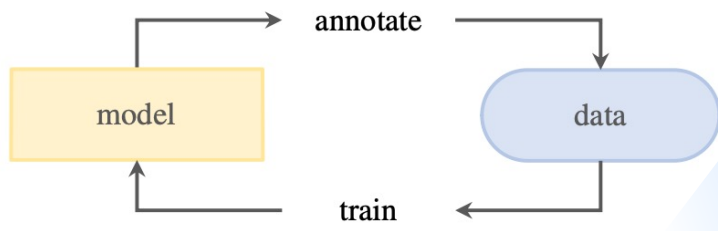
Everything

Cut-Outs

Segment Anything

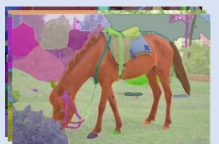
❖ SA-1B dataset

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



Segment Anything 1B (SA-1B):

- 1+ billion masks
- 11 million images
- privacy respecting
- licensed images



<50 masks



200-300 masks

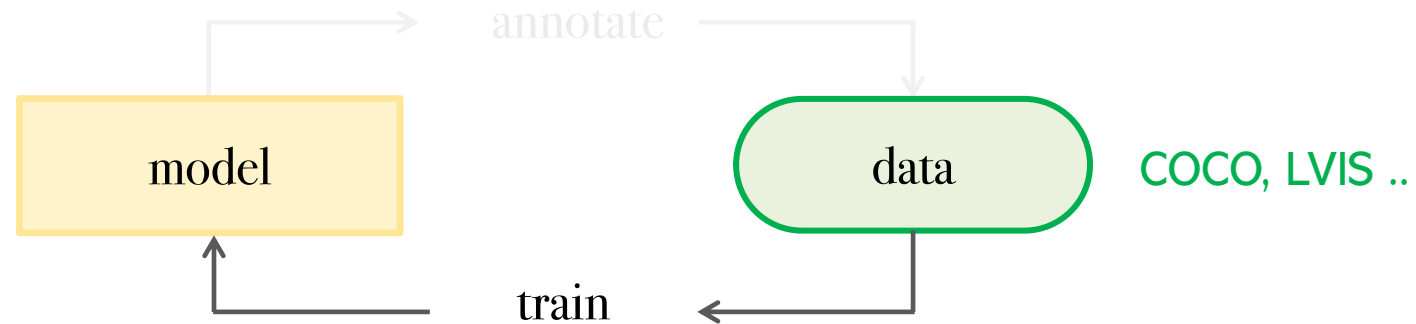
⋮



>500 masks

Segment Anything

❖ SA-1B dataset

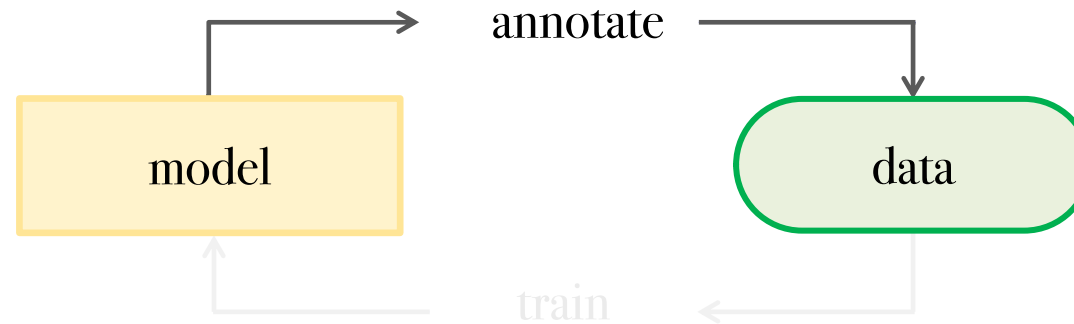


① Trained using common **public** segmentation datasets

Segment Anything

❖ SA-1B dataset

② model-assisted manual annotation stage

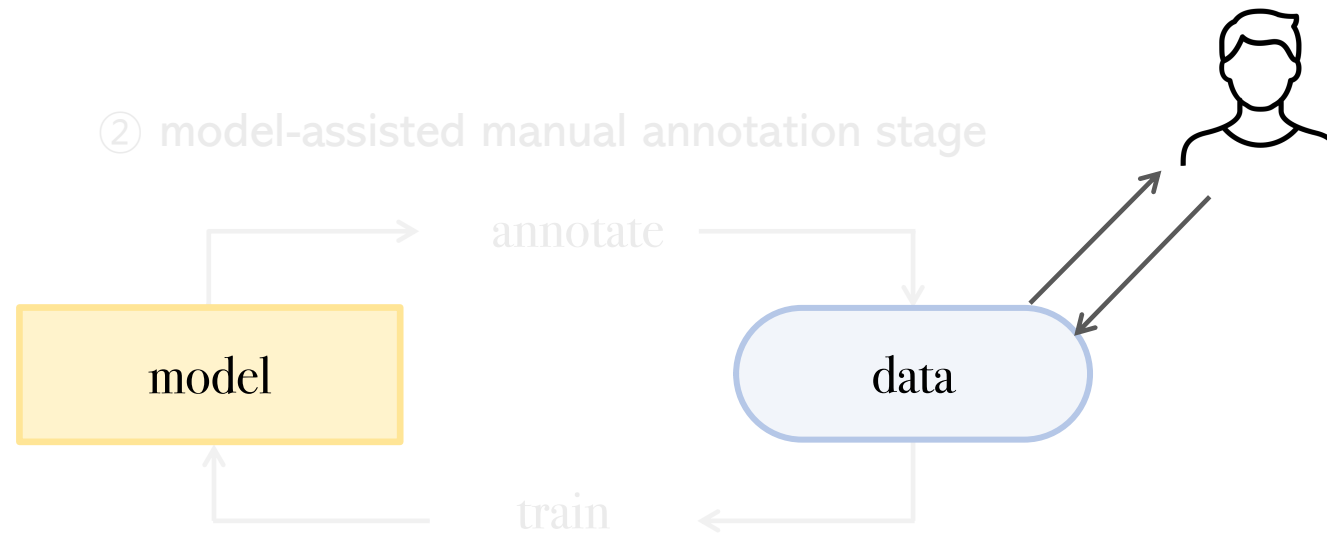


① Trained using common public segmentation datasets

Segment Anything

❖ SA-1B dataset

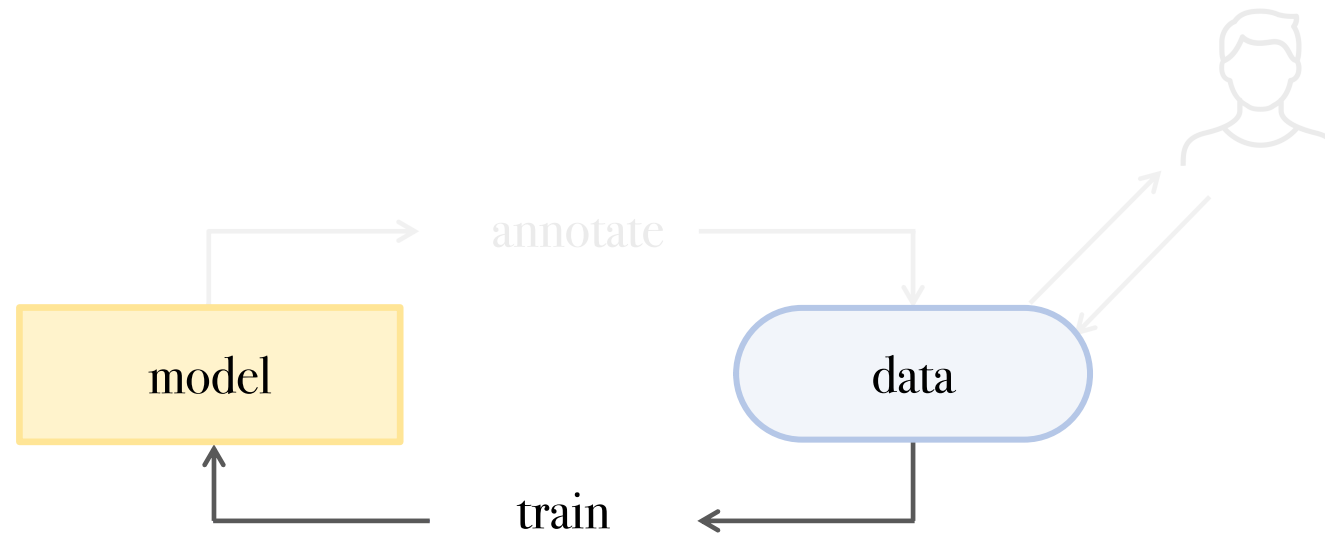
③ annotate any additional unannotated objects



Segment Anything

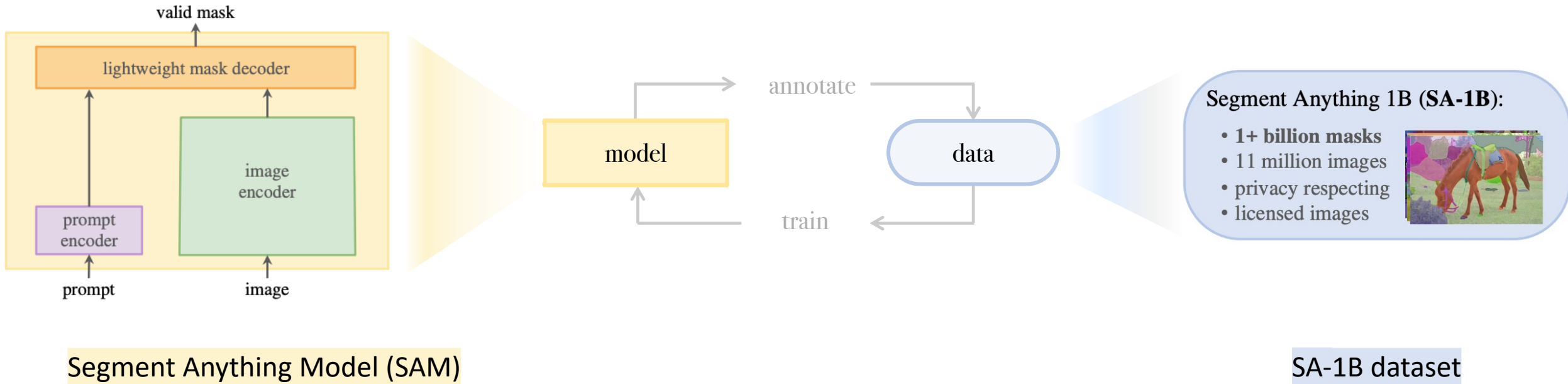
❖ SA-1B dataset

③ annotate any additional unannotated objects



④ fully automatic stage
: model generates masks without annotator input


Segment Anything



Segment Anything

❖ Segment Anything (SAM)

Segment Anything and its Adapter

발표자:  조용원

📅 2023년 12월 8일

🕒 오전 12시 ~

📍 고려대학교 신공학관 218호

▶ 온라인 비디오 시청 (YouTube)

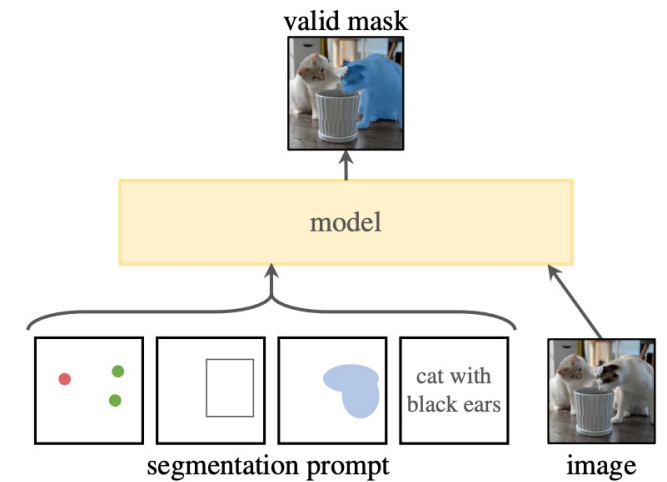
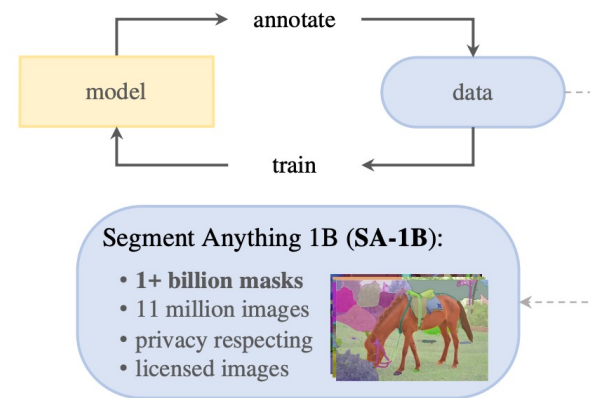
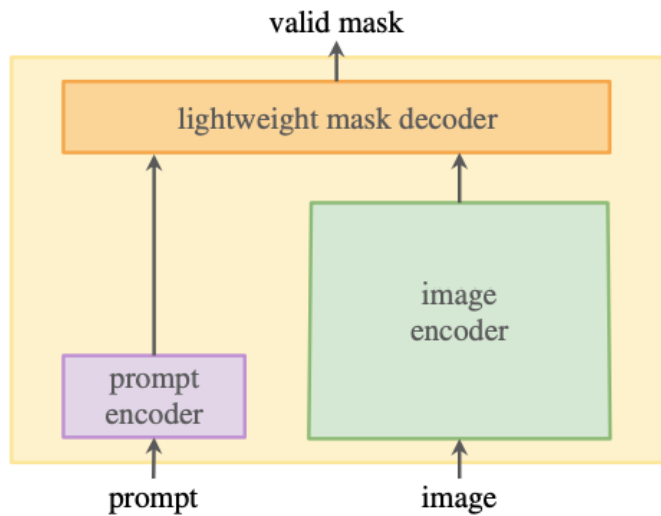
Fine-tuning Segment Anything

발표자:  김성수

📅 2025년 1월 3일

🕒 오전 12시 ~

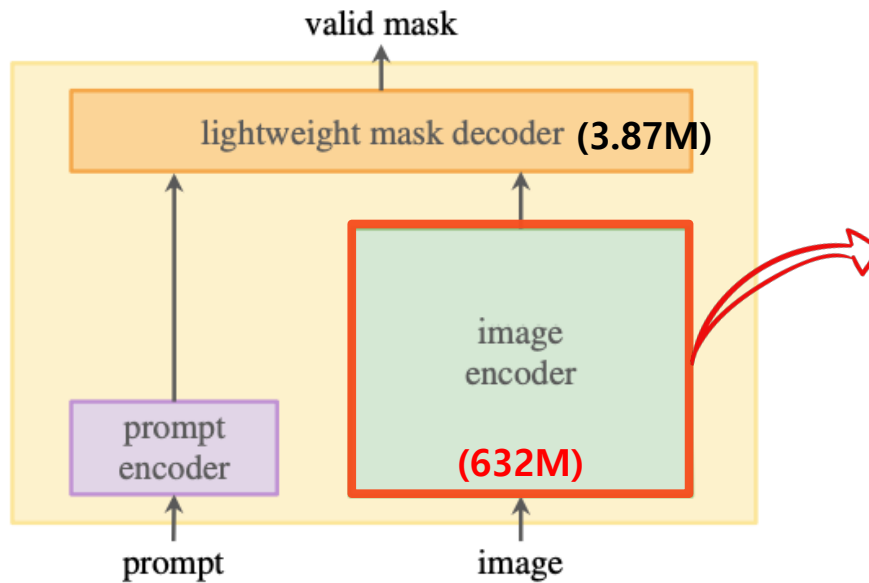
▶ 온라인 비디오 시청 (YouTube)



Segment Anything

❖ Segment Anything (SAM)

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

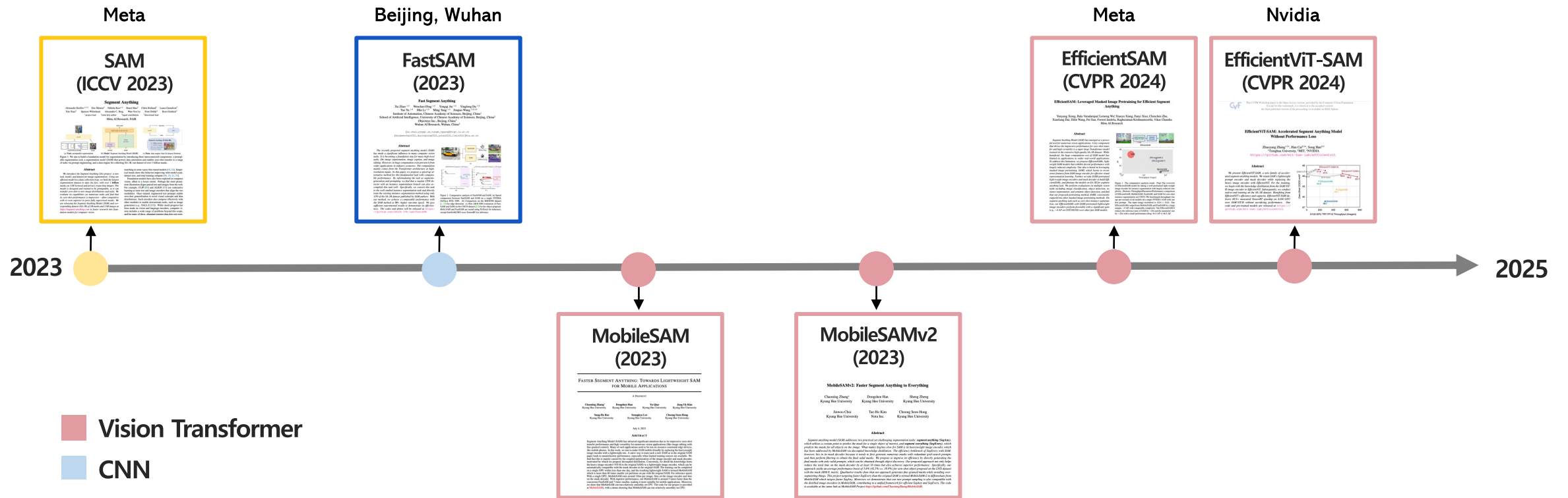


!

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

Research Trends

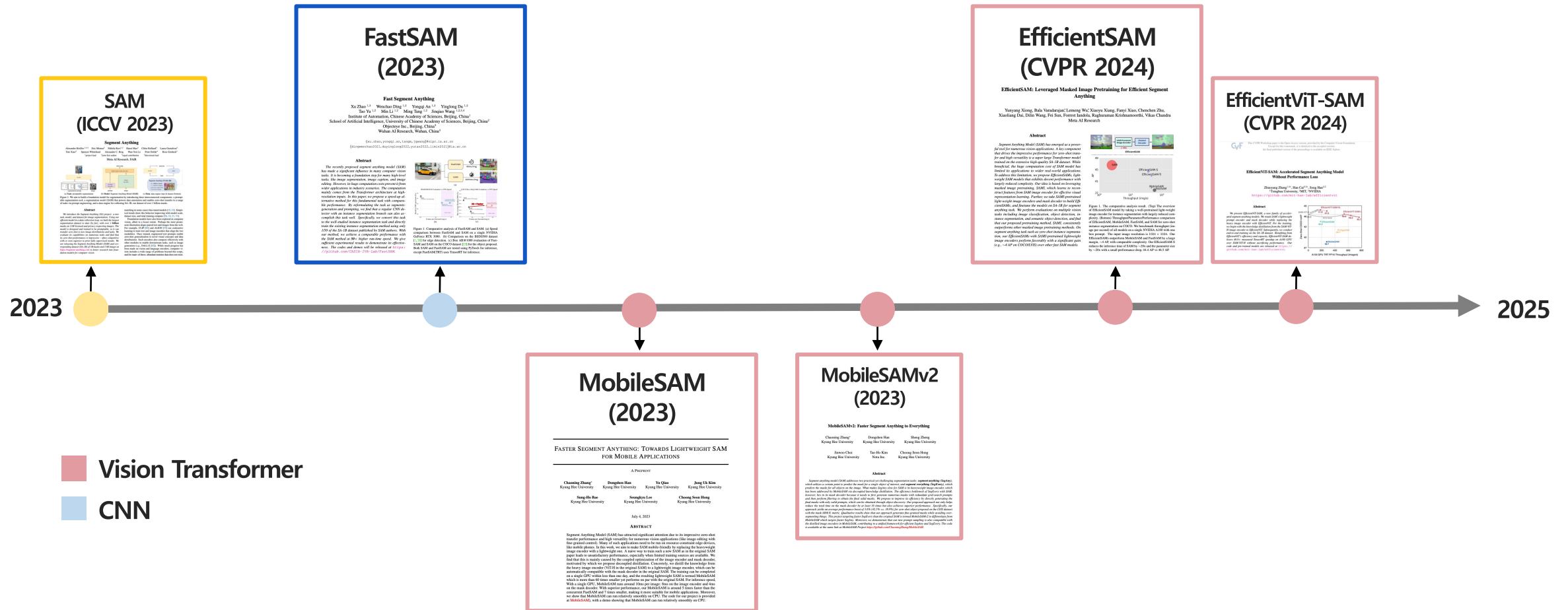
❖ SAM 경량화 연구 동향



Korea, Kyung Hee University

Research Trends

❖ SAM 경량화 연구 동향



Vision Transformer
CNN

SAM (ICCV 2023)

Segment Anything

Xu et al. (2023)

The recently proposed segment anything model (SAM) has made a significant advance in many computer vision tasks. It is becoming a foundation step for many high-level tasks, like image segmentation, image caption, and image editing. However, it has some computational overheads in some applications in industry scenarios. The computation mainly comes from the Transformer architecture or high-resolution input. In this paper, we propose a general and scalable method for accelerating SAM and improve the performance. By introducing the use of attention-free architecture, we reduce the use of attention-free architecture. By introducing the use of attention-free architecture, we reduce the use of attention-free architecture. By introducing the use of attention-free architecture, we reduce the use of attention-free architecture.

FastSAM (2023)

Fast Segment Anything

Xu Zhao^{1,2}, Wenzhe Ding^{1,2}, Yongqi Ai^{1,2}, Yinghao Du^{1,2}, Tao Yu^{1,2}, Ming Tang^{1,2}, Junqiang Wang^{1,2,3*}, Institute of Automation, Chinese Academy of Sciences, Beijing, China¹, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China², Wuhan AI Research, Wuhan, China³

fastsam@caict.ac.cn, {dingwenzhe, ai, du, tangming, wangjunqiang}@caict.ac.cn

Abstract

The recently proposed segment anything model (SAM) has made a significant advance in many computer vision tasks. It is becoming a foundation step for many high-level tasks, like image segmentation, image caption, and image editing. However, it has some computational overheads in some applications in industry scenarios. The computation mainly comes from the Transformer architecture or high-resolution input. In this paper, we propose a general and scalable method for accelerating SAM and improve the performance. By introducing the use of attention-free architecture, we reduce the use of attention-free architecture. By introducing the use of attention-free architecture, we reduce the use of attention-free architecture.

EfficientSAM (CVPR 2024)

EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything

Yunpeng Dong, Bohu Nandorff, Linyang Wu, Xinyu Xing, Fanyi Sun, Chuanxin Zhu, Xiaoliang Dai, Dili Wang, Pei Sun, Forrest Iandola, Radhakrishnan Krishnamoorti, Xiao Chudong, Meta AI Research

Abstract

Segment Anything Model (SAM) has emerged as a general and fast segmenter across applications. A key component that drives the impressive performance for zero-shot transfer and high versatility is a large heavy Transformer model trained on the dataset of about 1B SAM prompts. While benefiting its huge representation power, SAM model has limited the applications in wider real-world applications. In this paper, we propose EfficientSAM, a lightweight SAM model that enables diverse performance with lower computational overheads. EfficientSAM leverages masked image pretraining, SAMV2, which learns to reconstruct features from SAM image patches for efficient model inference. Further, we take SAMV2 pretraining as a foundation for EfficientSAM. EfficientSAM introduces a novel architecture, which is more efficient and faster than SAMV2. The positive evaluation on multiple scenes with varying image characteristics, various prompts, in-domain segmentation, and semantic object detection, and that our proposed pretraining method, SAMV2, consistently outperforms other masked image pretraining methods. Our proposed method and our proposed SAMV2 pretraining method consistently outperform other methods. Our proposed method and our proposed SAMV2 pretraining method consistently outperform other methods.

EfficientViT-SAM (CVPR 2024)

EfficientViT-SAM: Accelerated Segment Anything Model Without Performance Loss

Zhangqing Zhang^{1*}, Han Cai^{1*}, Song Han¹

zhangqingzhang@uwaterloo.ca, {hancai, songhan}@uwaterloo.ca

Abstract

All recent Segment Anything Model (SAM) variants have achieved impressive performance on the SAM benchmark. However, they still suffer from high inference latency and large model size, which hinders their deployment in resource-constrained scenarios. In this paper, we propose EfficientViT-SAM, a lightweight SAM model that achieves competitive performance with lower computational overheads. EfficientViT-SAM leverages ViT architecture and introduces a novel architecture, which is more efficient and faster than SAM. The positive evaluation on multiple scenes with varying image characteristics, various prompts, in-domain segmentation, and semantic object detection, and that our proposed pretraining method, SAMV2, consistently outperforms other masked image pretraining methods. Our proposed method and our proposed SAMV2 pretraining method consistently outperform other methods.

MobileSAM (2023)

FASTER SEGMENT ANYTHING: TOWARDS LIGHTWEIGHT SAM FOR MOBILE APPLICATIONS

Changsheng Zhang¹, Dongchen Han¹, Yu Qian¹, Jing Li Kim¹, Kyung Hee University, Kyung Hee University, Kyung Hee University, Kyung Hee University

kyunghe@khu.ac.kr, dongchenhan@khu.ac.kr, qianyu@khu.ac.kr, jingli@khu.ac.kr

Abstract

Segment Anything Model (SAM) has attracted significant attention due to its impressive zero-shot transfer performance and its versatility for numerous computer vision applications that range from image editing to image captioning. However, its large model size and high inference latency hinder its deployment in resource-constrained scenarios. In this paper, we propose MobileSAM, a lightweight SAM model that achieves competitive performance with lower computational overheads. MobileSAM leverages ViT architecture and introduces a novel architecture, which is more efficient and faster than SAM. The positive evaluation on multiple scenes with varying image characteristics, various prompts, in-domain segmentation, and semantic object detection, and that our proposed pretraining method, SAMV2, consistently outperforms other masked image pretraining methods. Our proposed method and our proposed SAMV2 pretraining method consistently outperform other methods.

MobileSAMv2 (2023)

MobileSAMv2: Faster Segment Anything to Everything

Changsheng Zhang¹, Dongchen Han¹, Song Zhang¹, Kyung Hee University, Kyung Hee University, Kyung Hee University

kyunghe@khu.ac.kr, dongchenhan@khu.ac.kr, songzhang@khu.ac.kr

Abstract

Segment Anything Model (SAM) has attracted significant attention due to its impressive zero-shot transfer performance and its versatility for numerous computer vision applications that range from image editing to image captioning. However, its large model size and high inference latency hinder its deployment in resource-constrained scenarios. In this paper, we propose MobileSAMv2, a lightweight SAM model that achieves competitive performance with lower computational overheads. MobileSAMv2 leverages ViT architecture and introduces a novel architecture, which is more efficient and faster than SAM. The positive evaluation on multiple scenes with varying image characteristics, various prompts, in-domain segmentation, and semantic object detection, and that our proposed pretraining method, SAMV2, consistently outperforms other masked image pretraining methods. Our proposed method and our proposed SAMV2 pretraining method consistently outperform other methods.

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, China¹

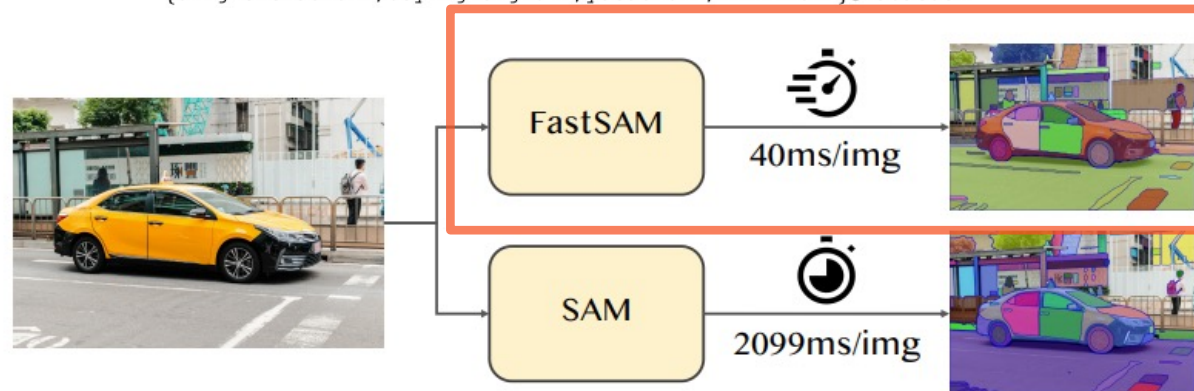
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China²

Objecteye Inc., Beijing, China³

Wuhan AI Research, Wuhan, China⁴

{xu.zhao,yongqi.an,tangm,jqwang}@nlpr.ia.ac.cn

{dingwenchao2021,duyinglong2022,yutao2022,limin2021}@ia.ac.cn



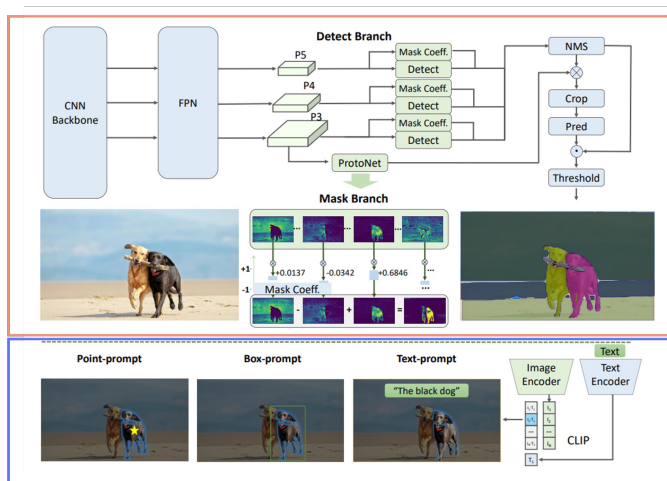
FastSAM

❖ FastSAM vs SAM

- 파라미터 수 비교

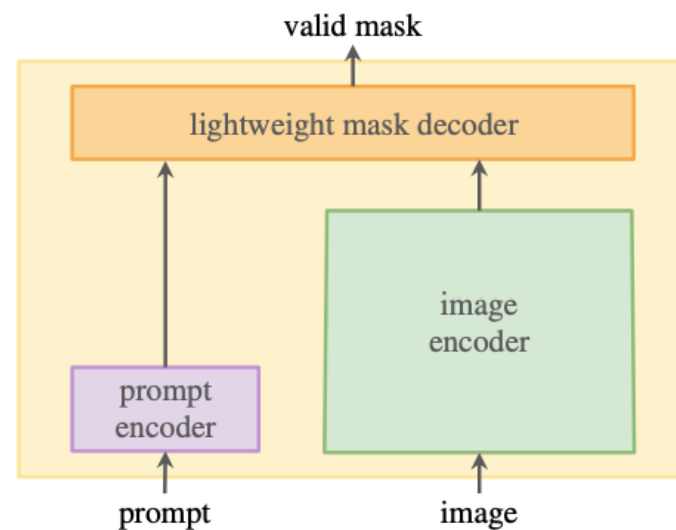
FastSAM

(68M)



SAM

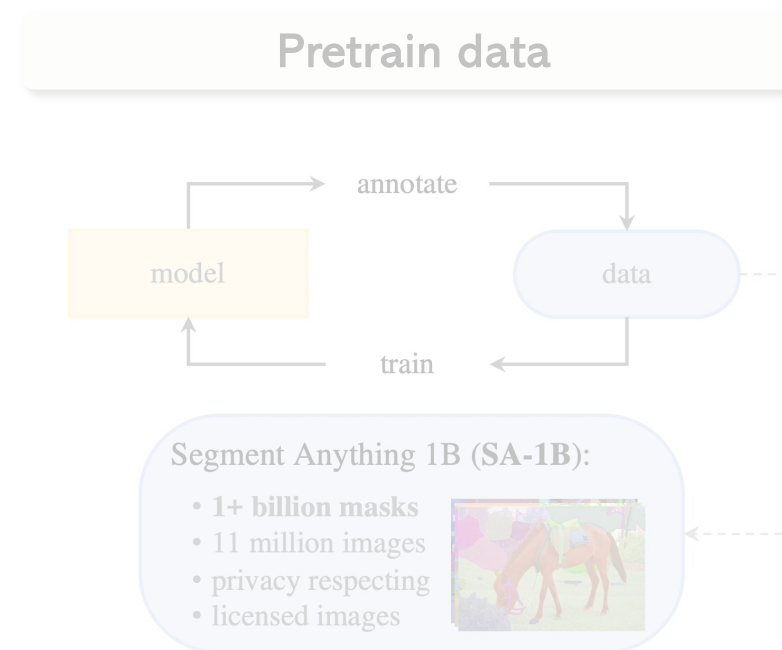
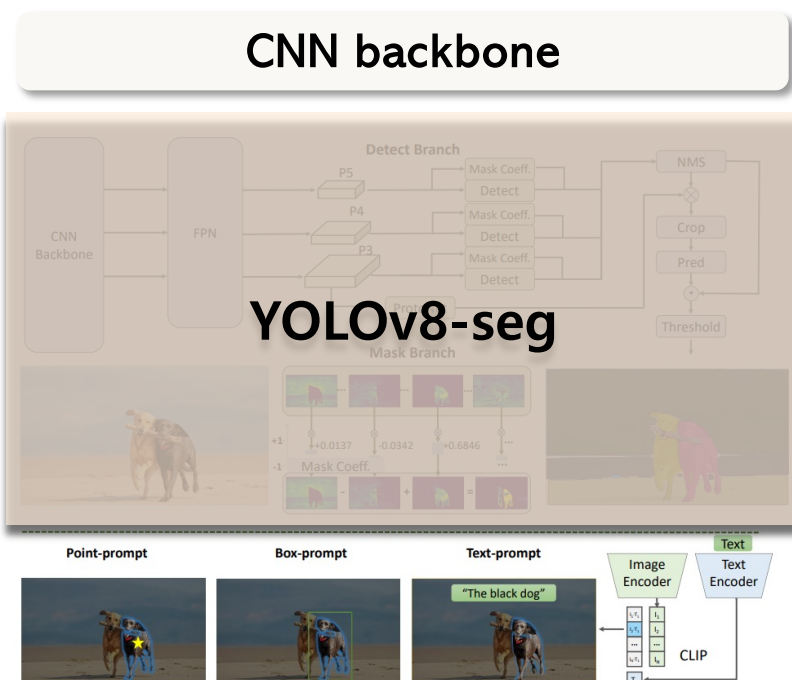
(632M)



FastSAM

❖ Fast Segment Anything

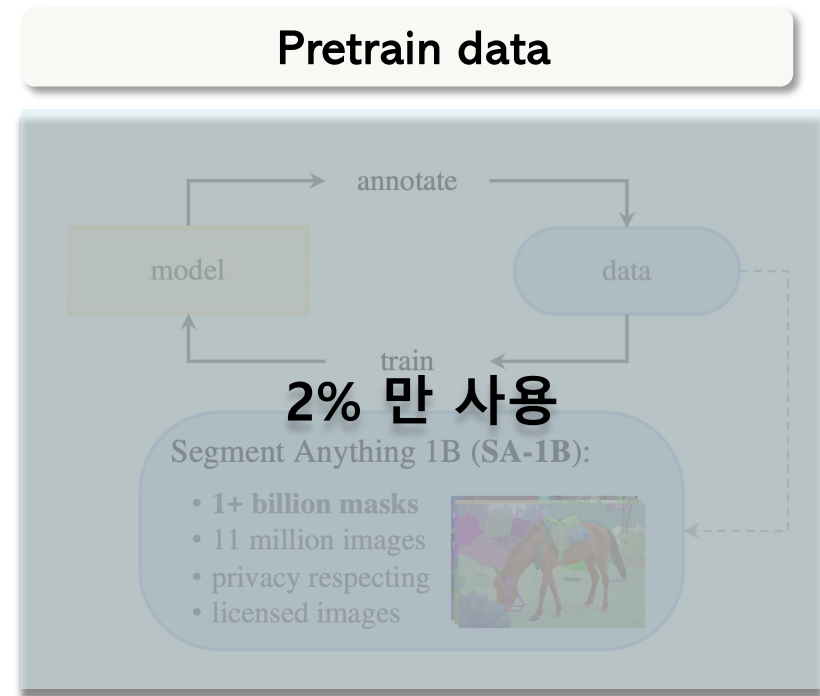
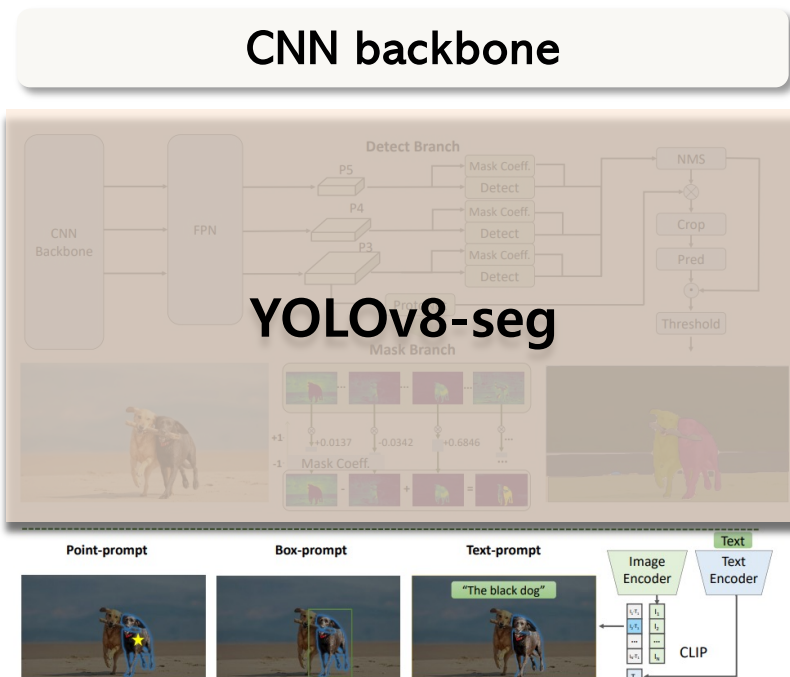
- CNN backbone을 활용하여 image encoder 대체
- SAM에서 게시한 광범위 SA-1B 데이터셋 사용



FastSAM

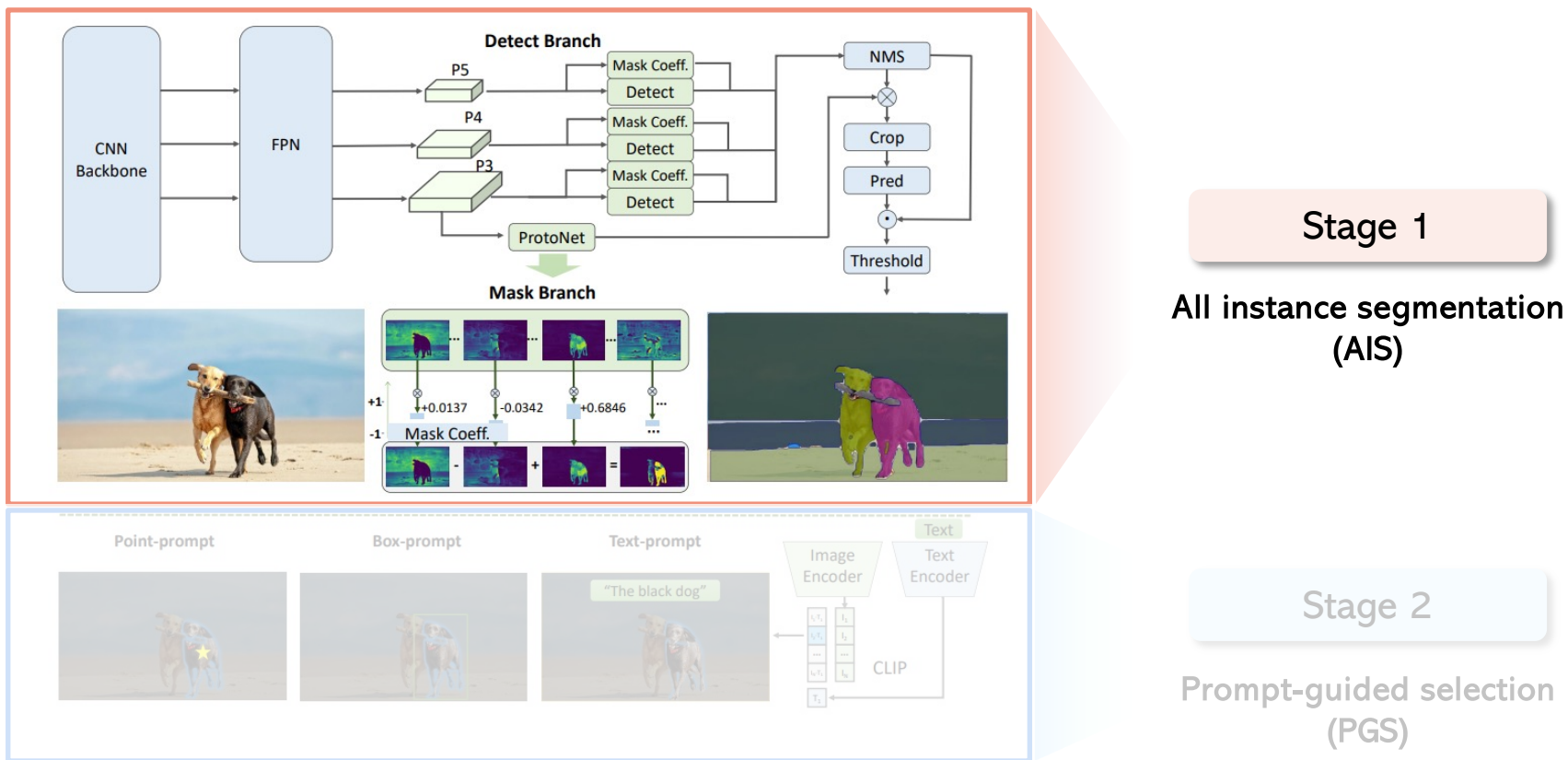
❖ Fast Segment Anything

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



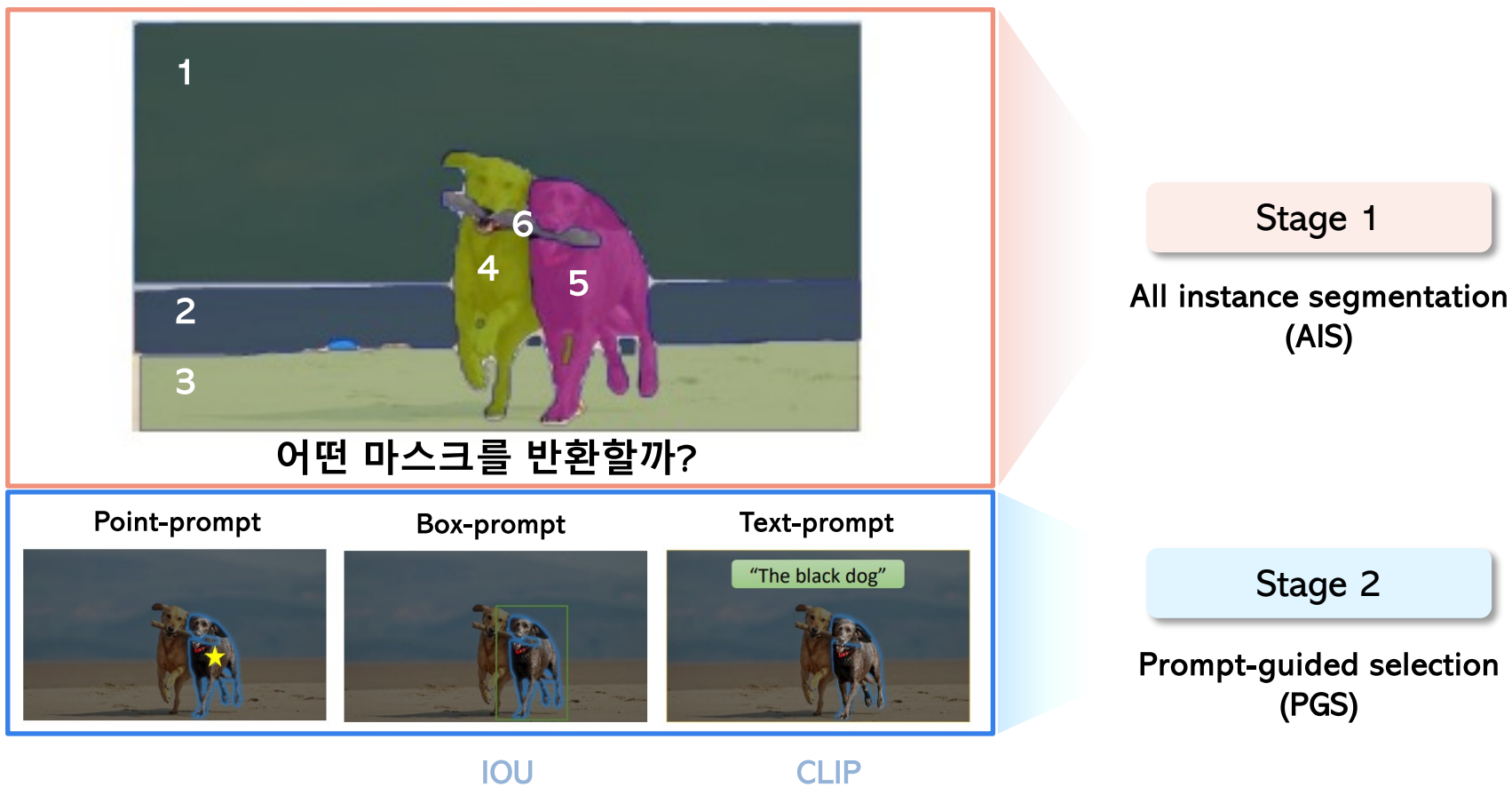
FastSAM

❖ 2-Stage Framework



FastSAM

❖ 2-Stage Framework



Stage 1
All instance segmentation
(AIS)

Stage 2
Prompt-guided selection
(PGS)

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 0.04초					

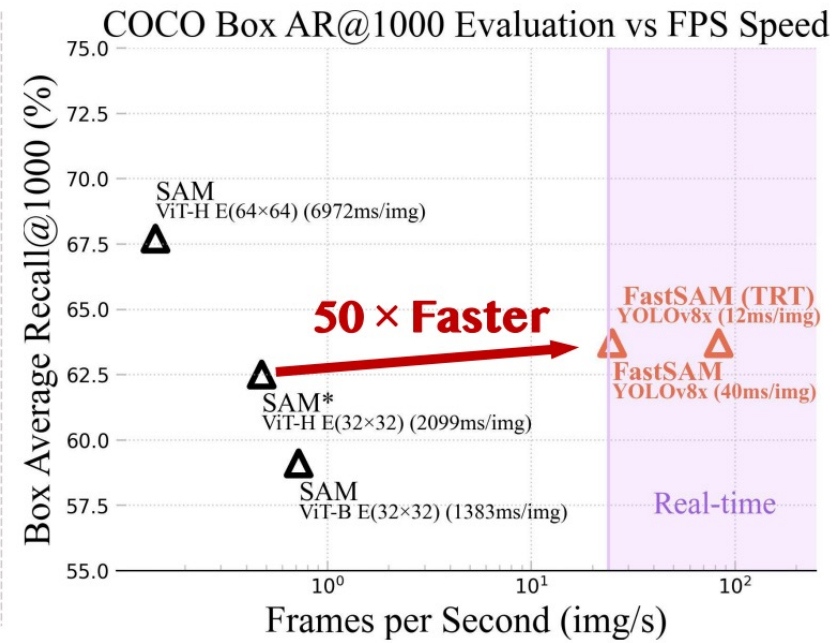
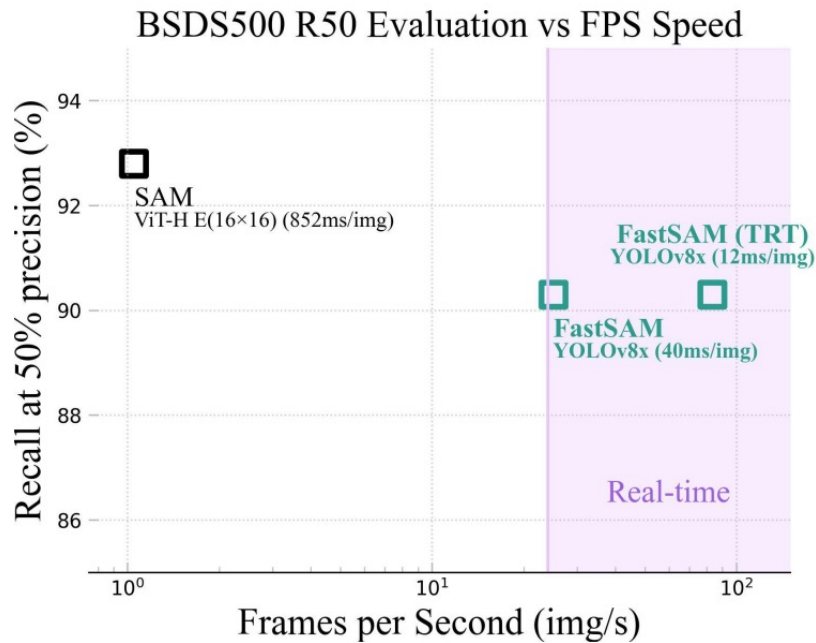
약 7초

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

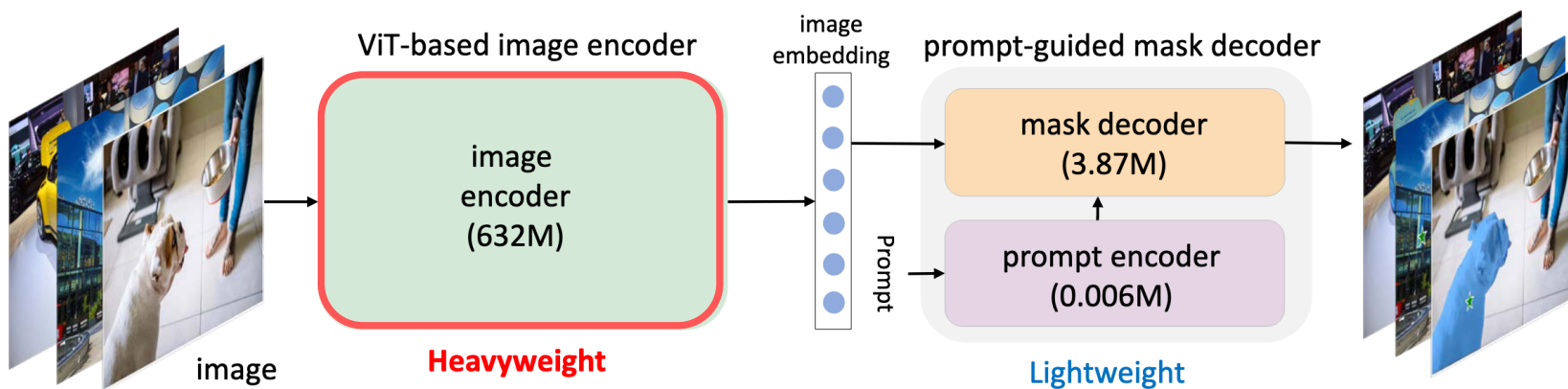
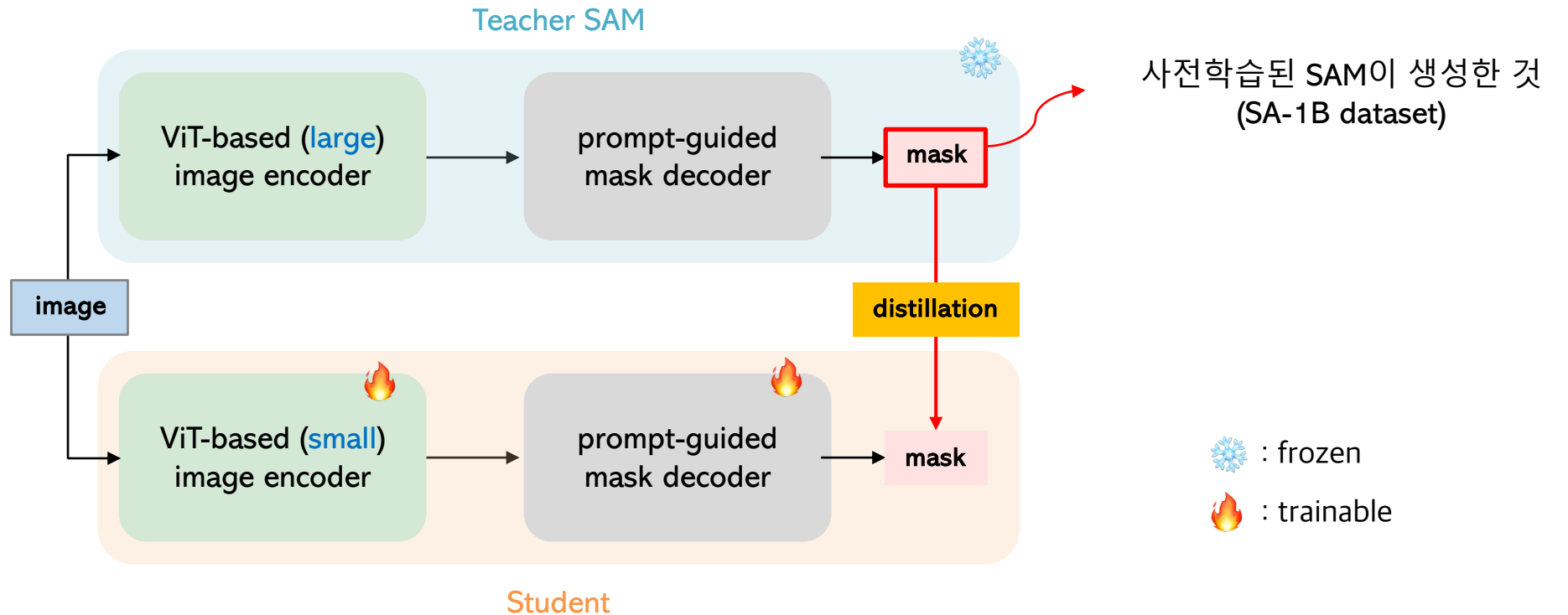


Figure 1: The overview of Segment Anything Model.

MobileSAM

❖ Knowledge distillation (지식 증류)

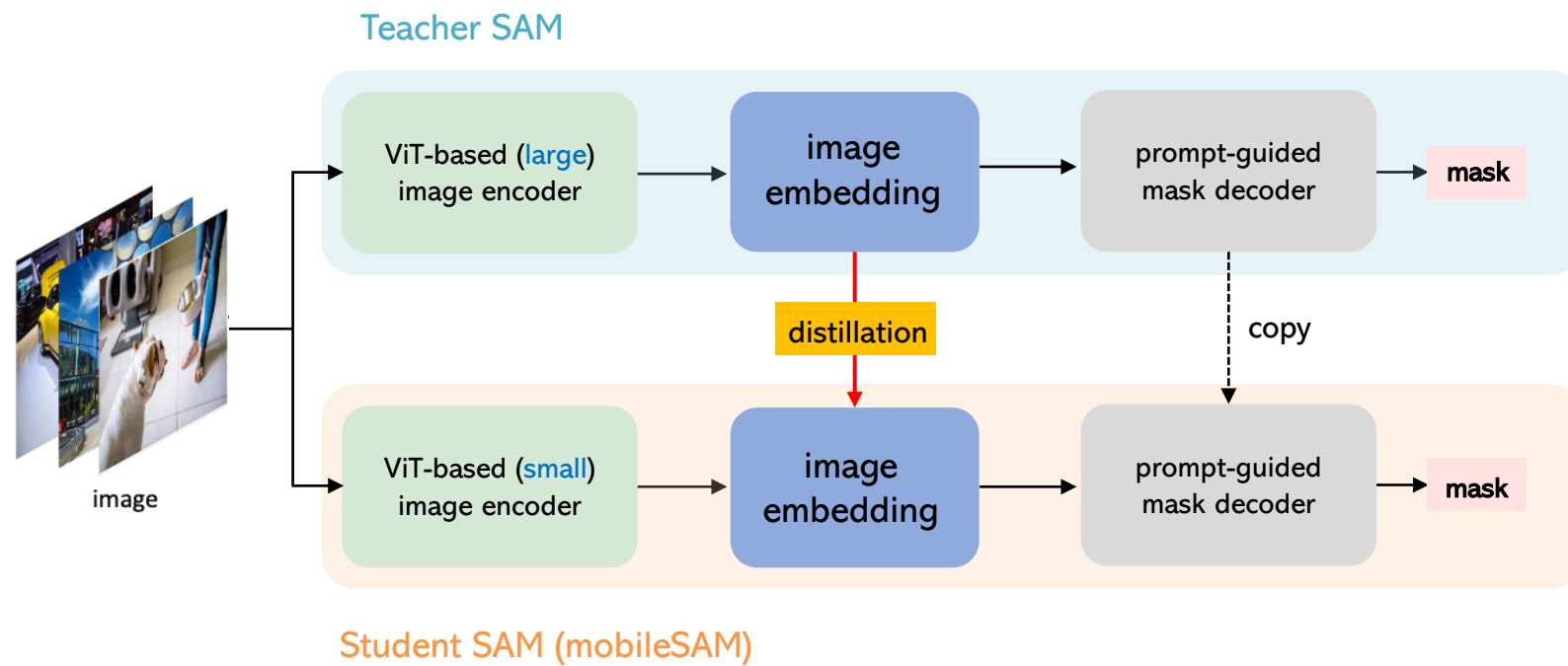
- ViT-H 기반 SAM의 **지식**을 더 작은 이미지 인코더를 사용하는 SAM으로 **전이**
- 인코더 파라미터를 100배, 전체 파라미터를 60배 줄임



MobileSAM

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

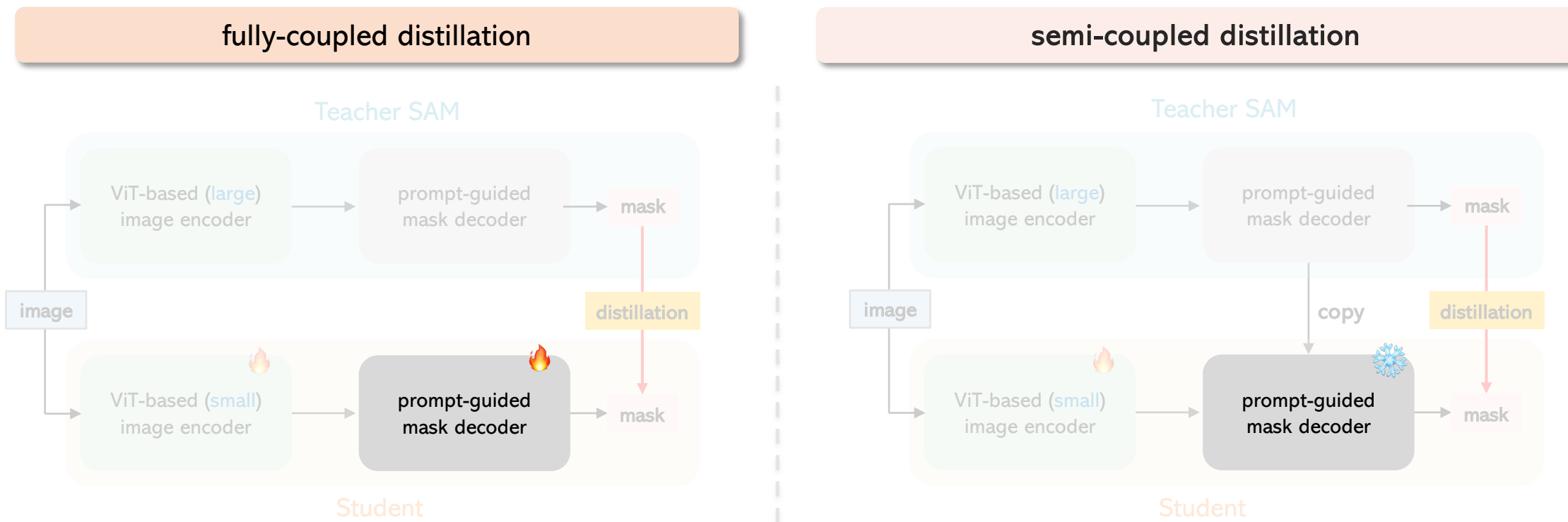
- 1) image encoder distillation
- 2) mask decoder finetuning



MobileSAM

❖ Knowledge distillation (지식 증류)

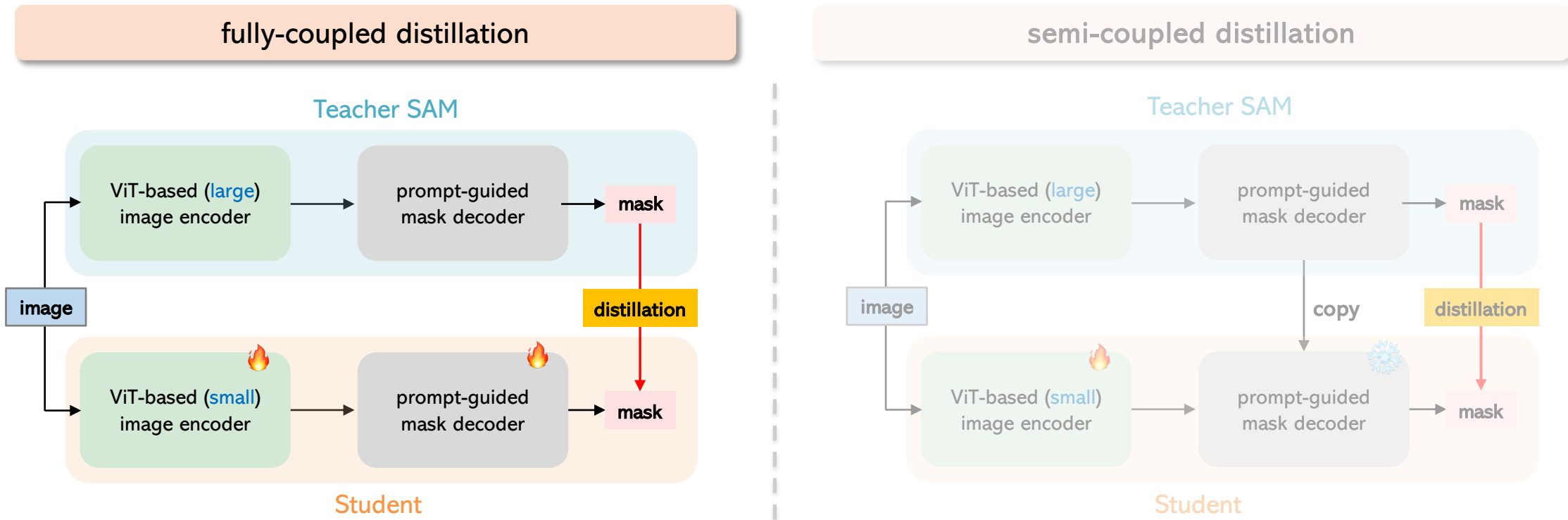
- Coupled optimization of the image encoder and combined decoder



MobileSAM

❖ Fully-coupled distillation

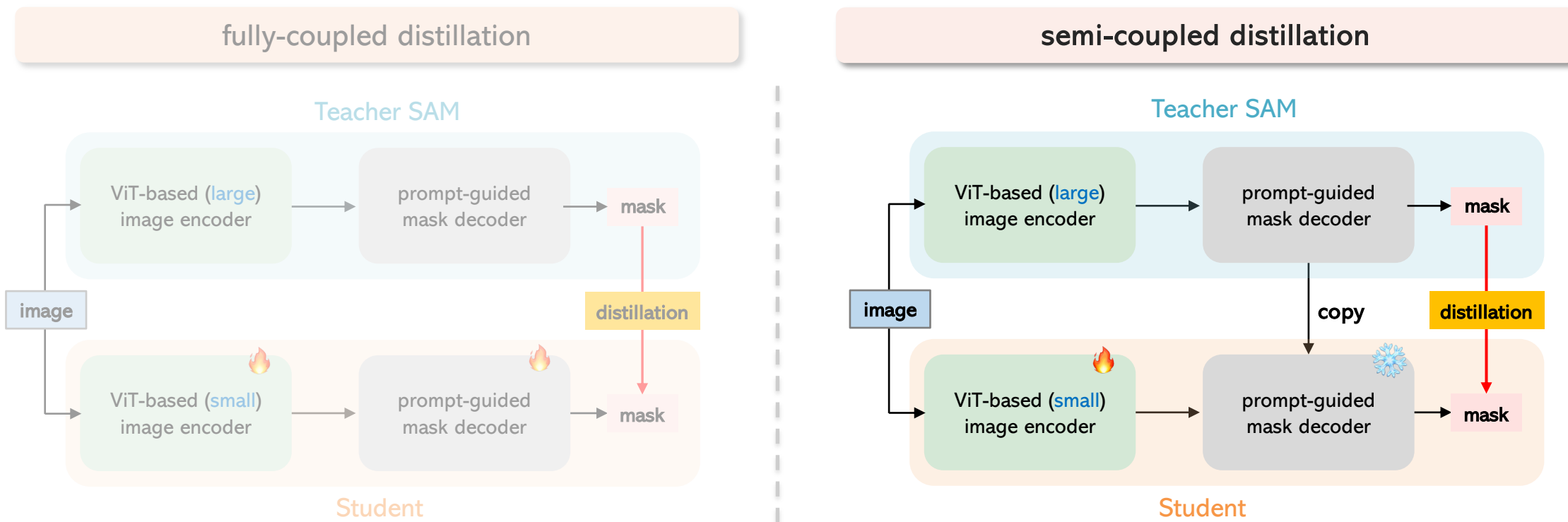
- Encoder와 decoder가 서로 의존적 → 두 모듈이 동시에 학습되기까지 시간이 오래 걸림
- Semi-coupled 방식으로 개선 시도



MobileSAM

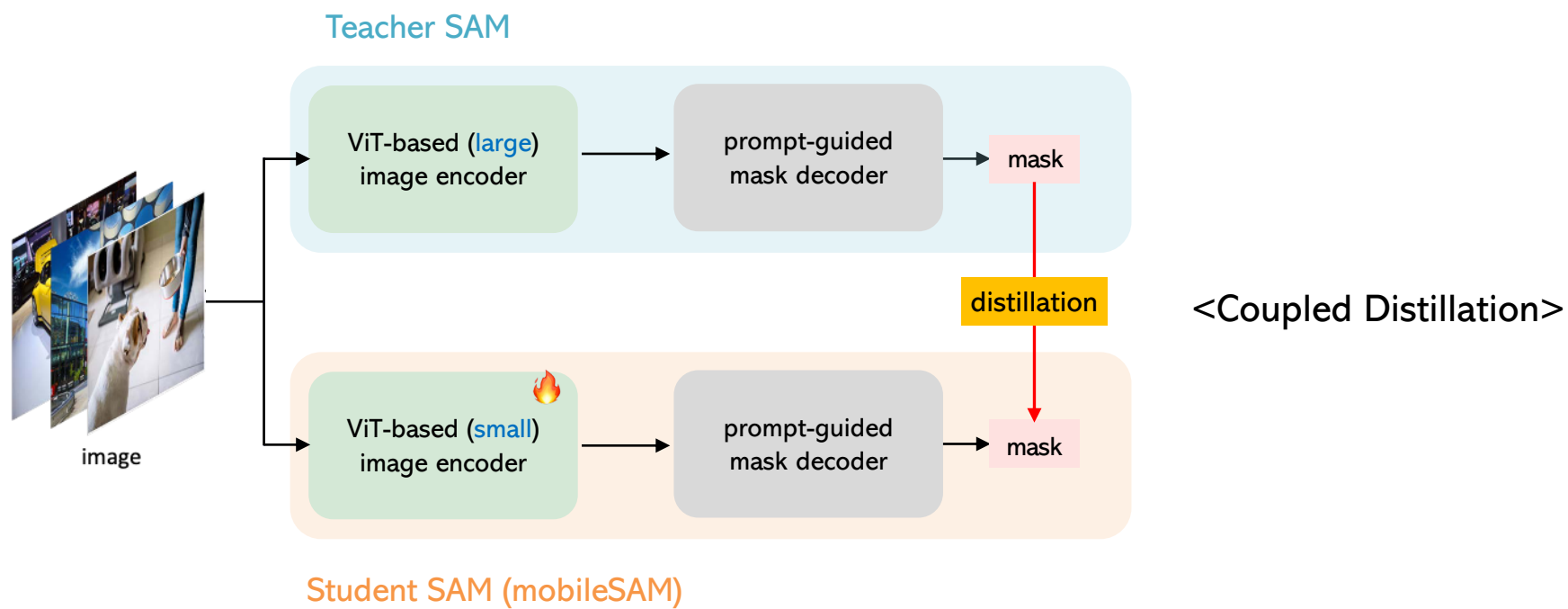
❖ Semi-coupled distillation

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



MobileSAM

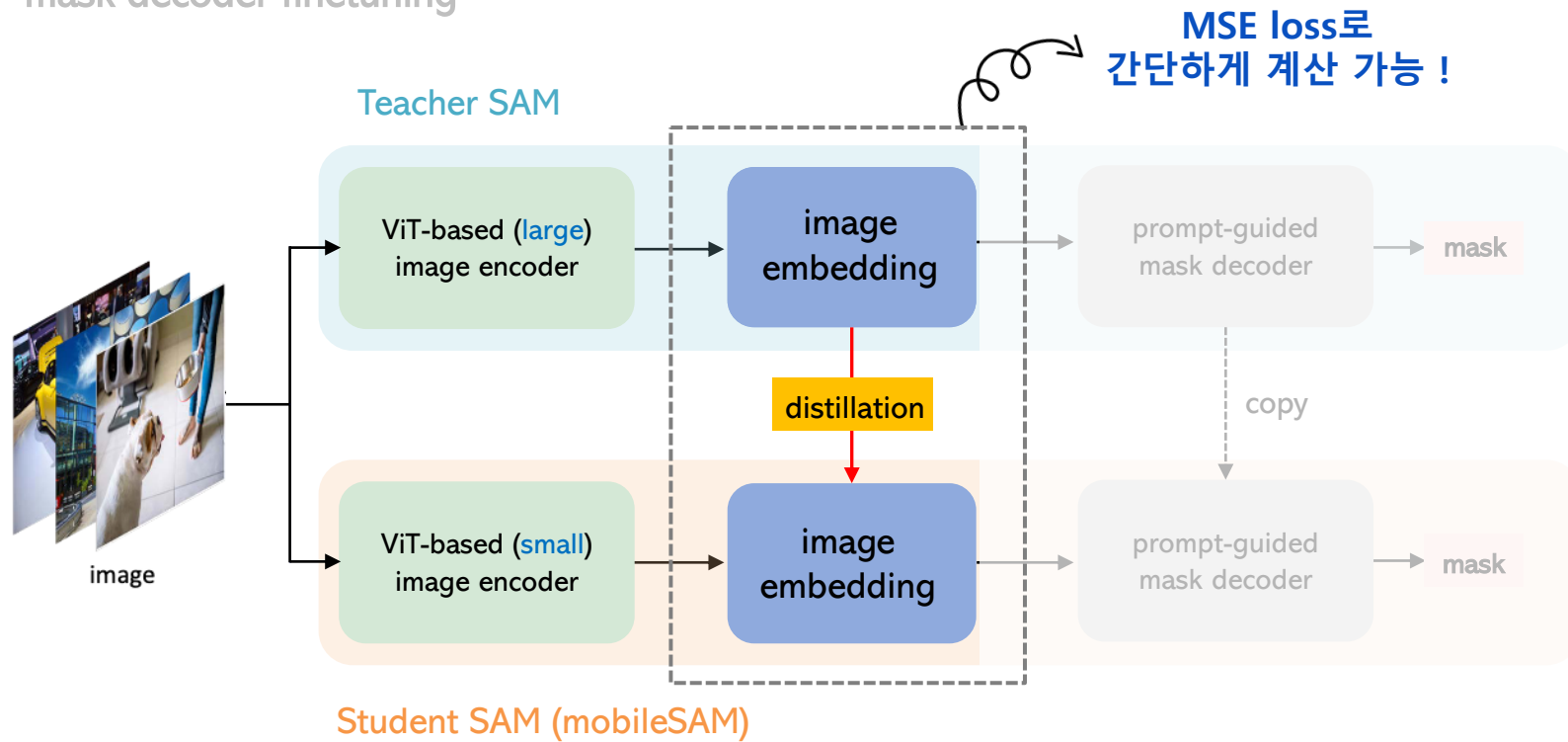
❖ Coupled Distillation의 단점을 보완한 Decoupled Distillation !



MobileSAM

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

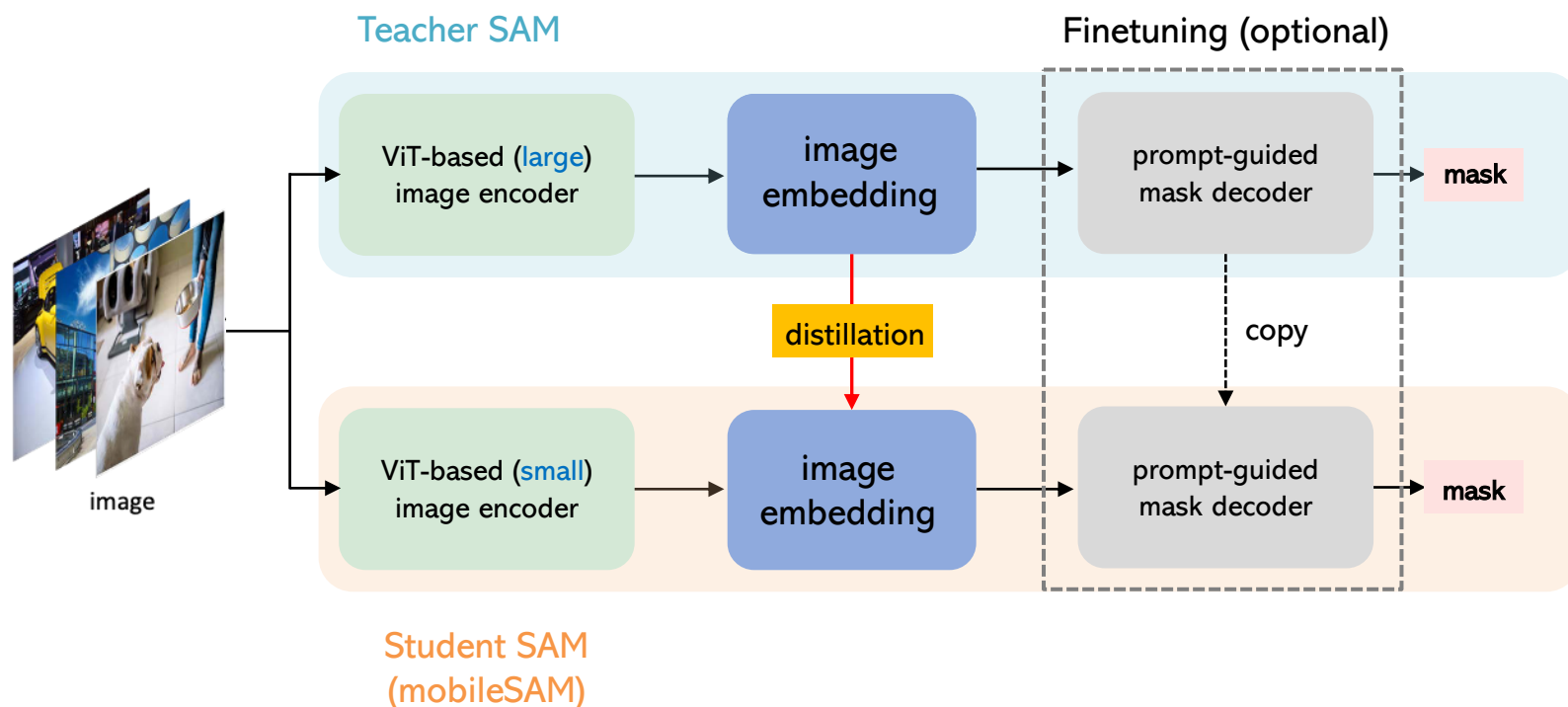
- 1) image encoder distillation
- 2) mask decoder finetuning



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 original SAM



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 smaller
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

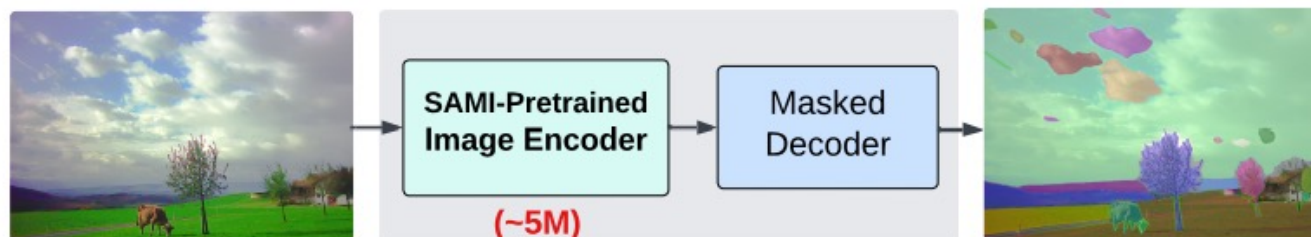
EfficientSAM

❖ EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything

- [Meta AI Research \(2024, CVPR\)](#)
- 인용수: 153회

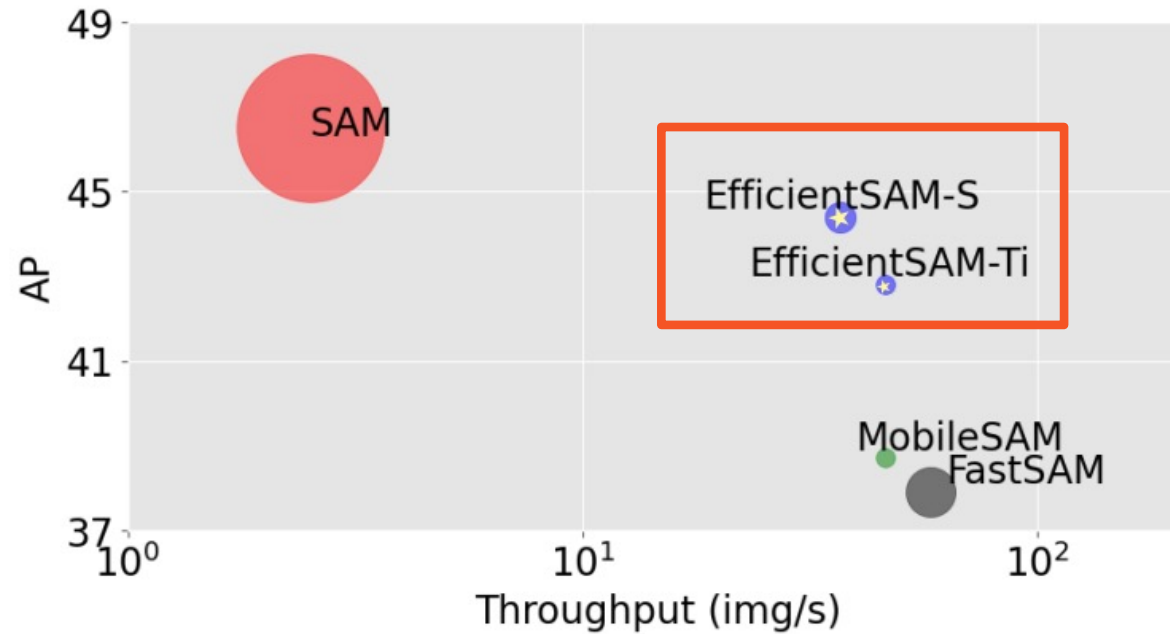
EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything

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

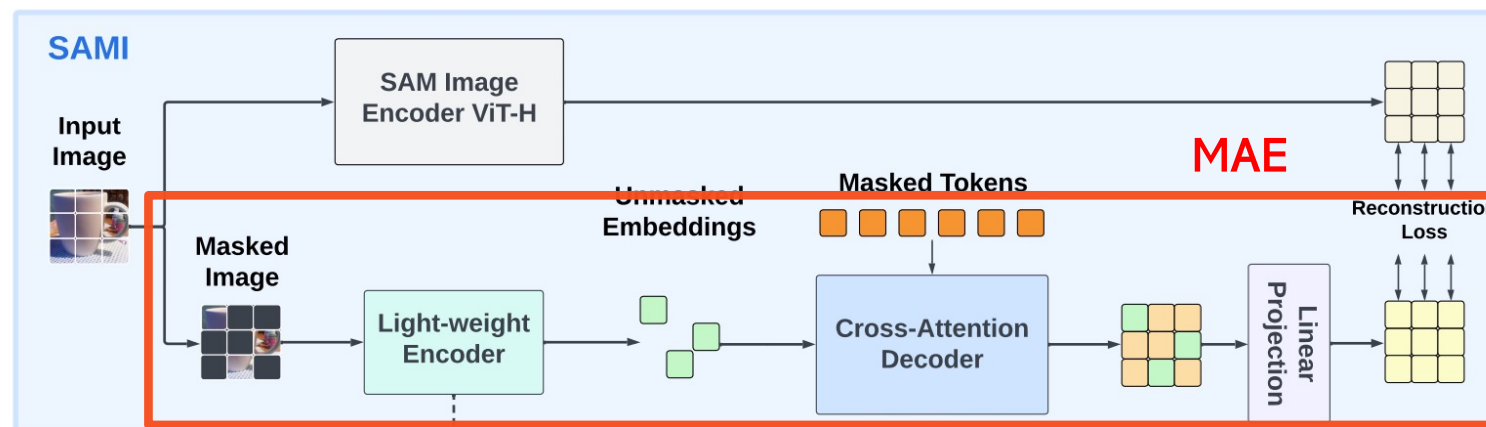
❖ EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything



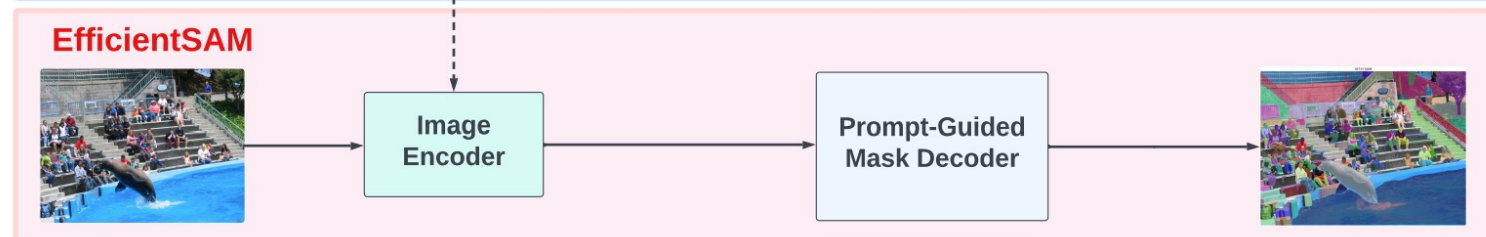
EfficientSAM

❖ Two stage framework

Stage 1
SAMI pretraining



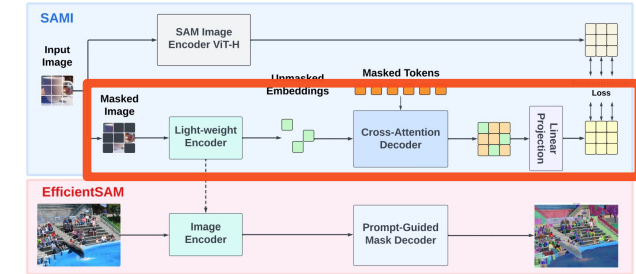
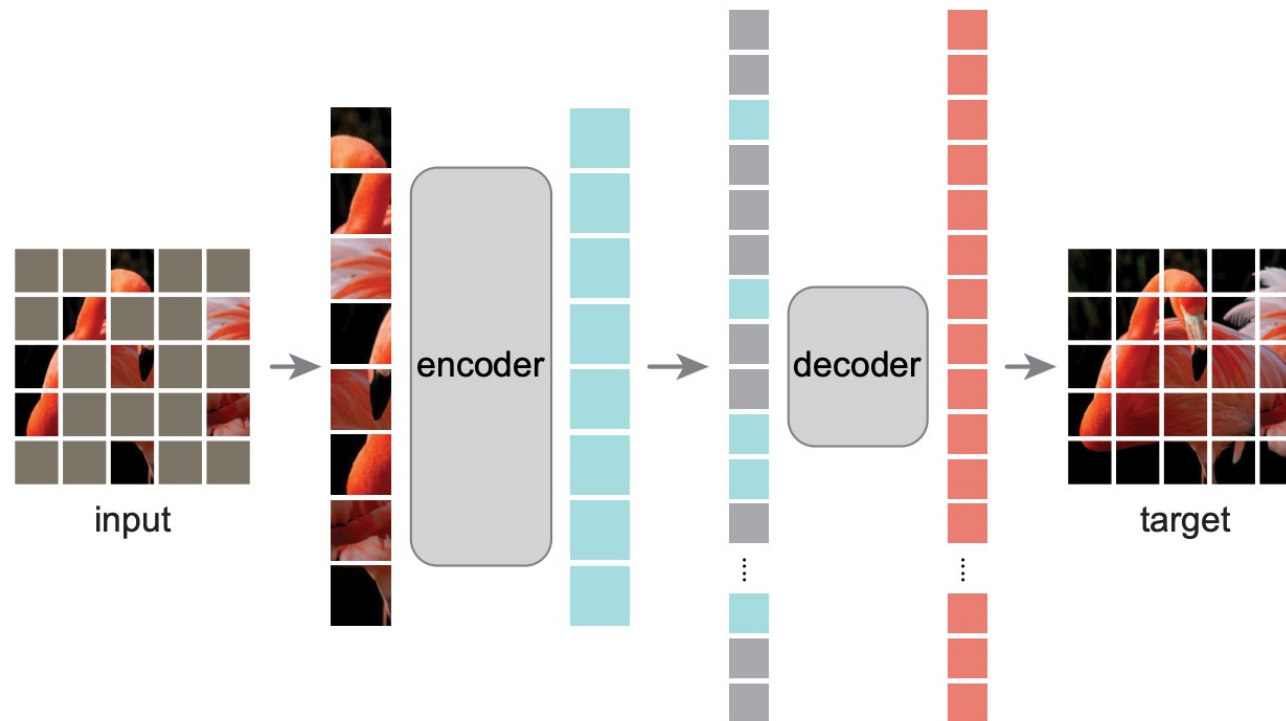
Stage 2
SAM finetuning



EfficientSAM

❖ MAE Architecture

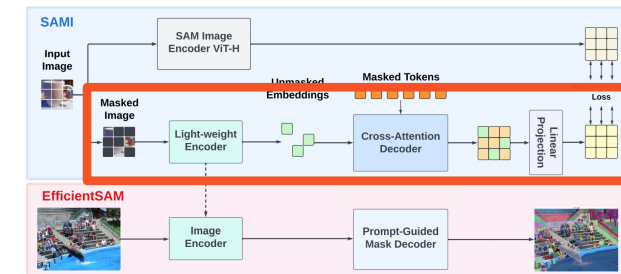
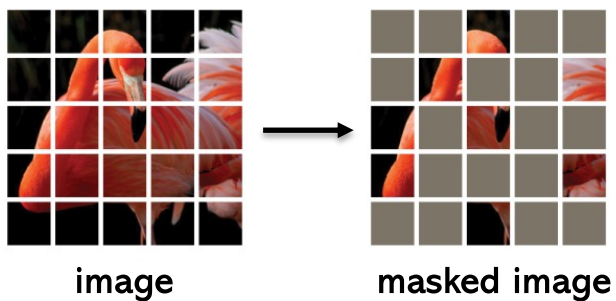
- Masked Autoencoders Are Scalable Vision Learners



EfficientSAM

❖ MAE Architecture

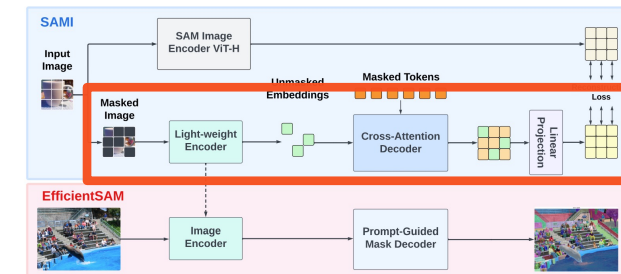
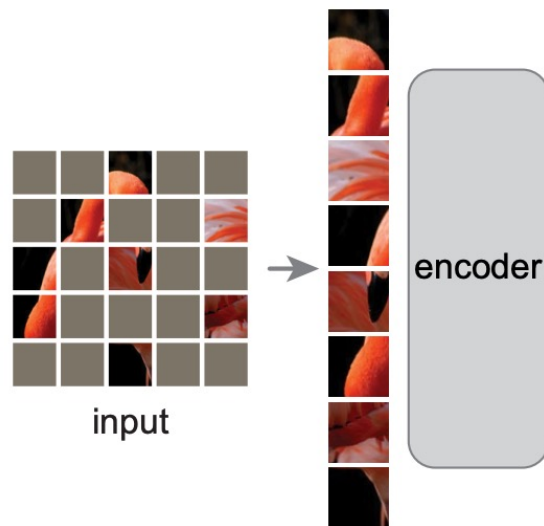
- Masked Autoencoders Are Scalable Vision Learners



EfficientSAM

❖ MAE Architecture

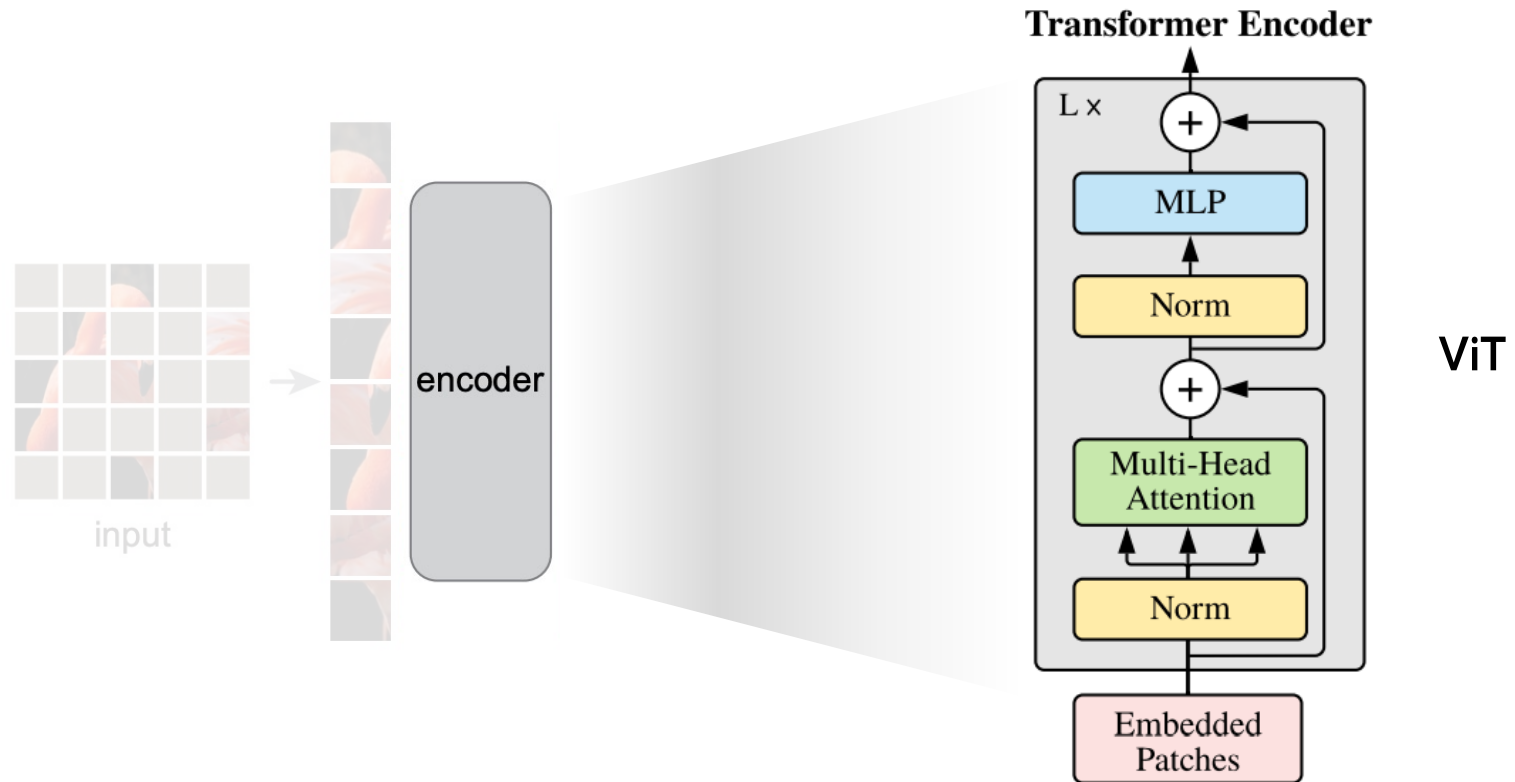
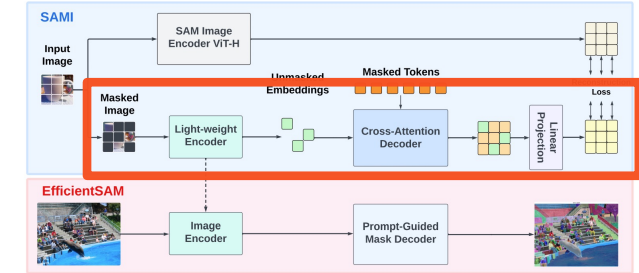
- Masked Autoencoders Are Scalable Vision Learners



EfficientSAM

❖ MAE Architecture

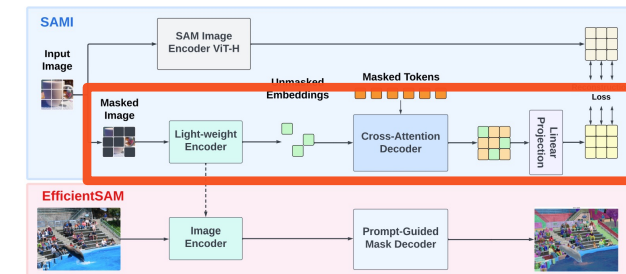
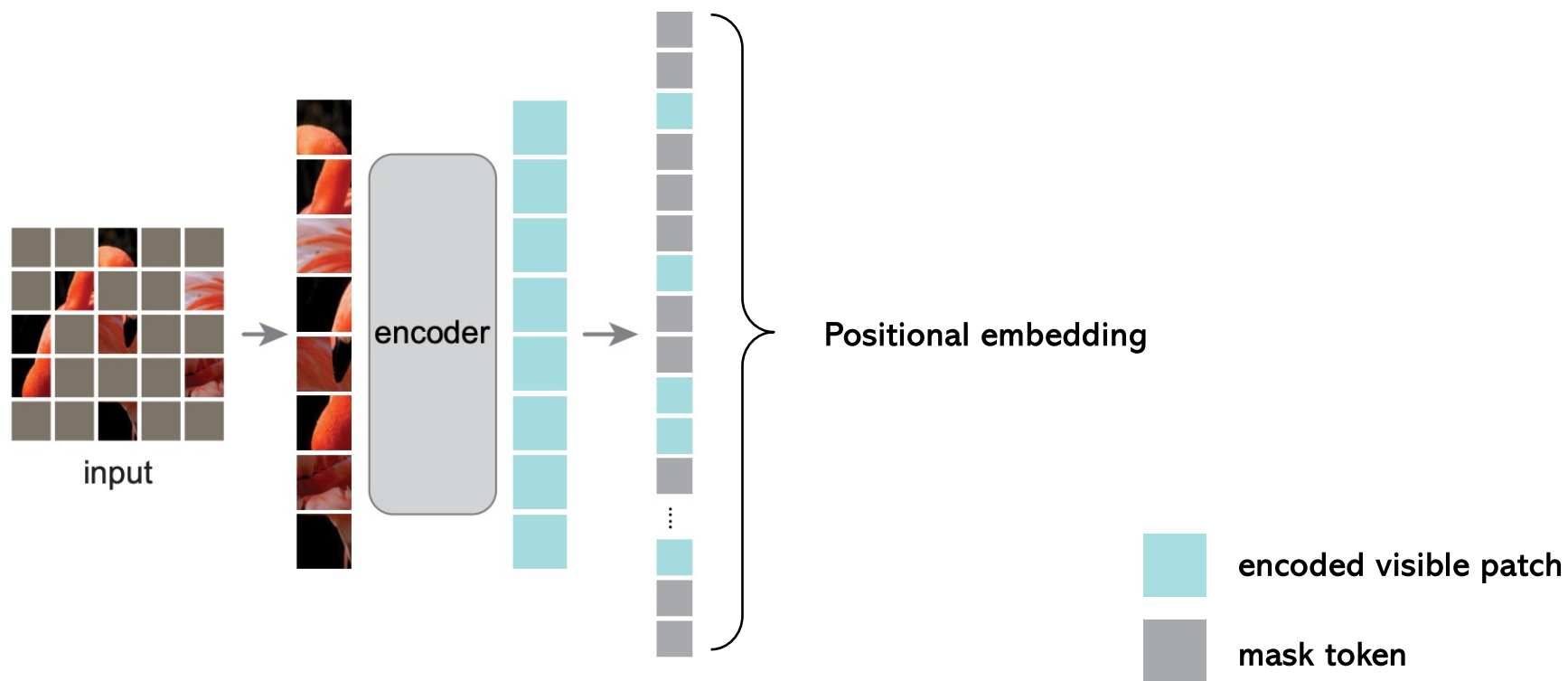
- Masked Autoencoders Are Scalable Vision Learners



EfficientSAM

❖ MAE Architecture

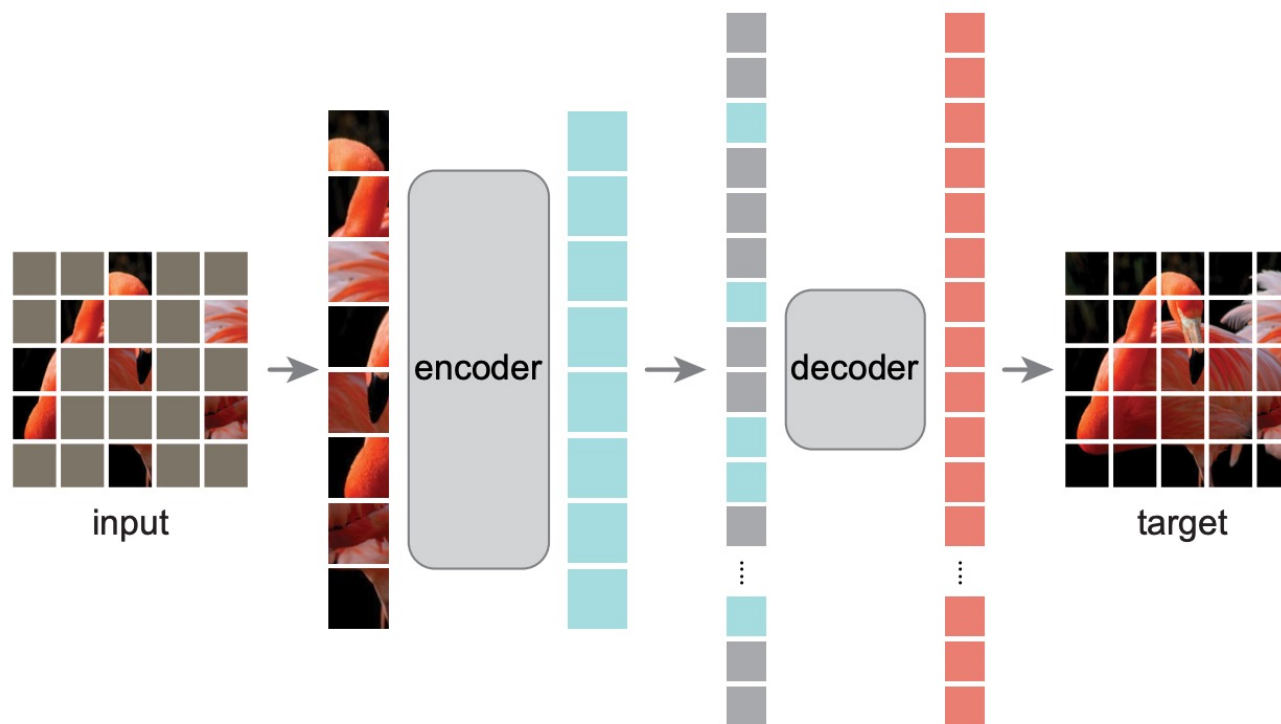
- Masked Autoencoders Are Scalable Vision Learners



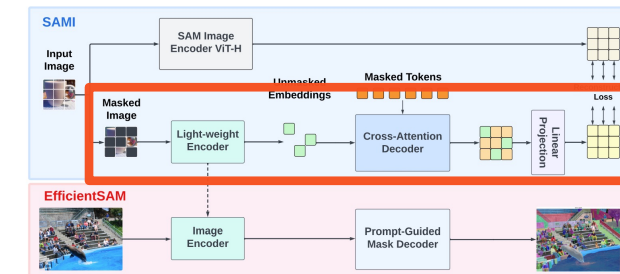
EfficientSAM

❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners



decoder에서만 mask token을 사용하는
비대칭 구조

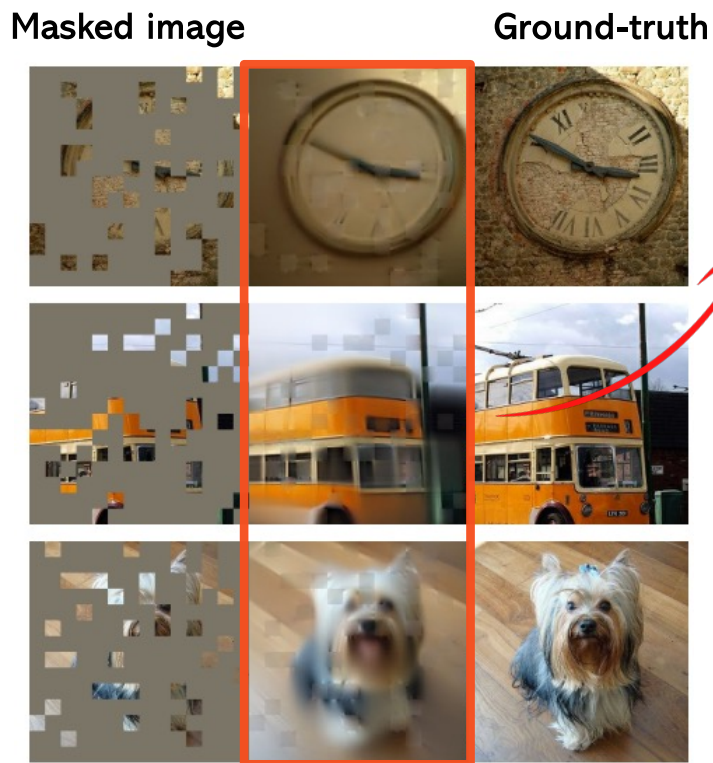


teal square encoded visible patch
grey square mask token

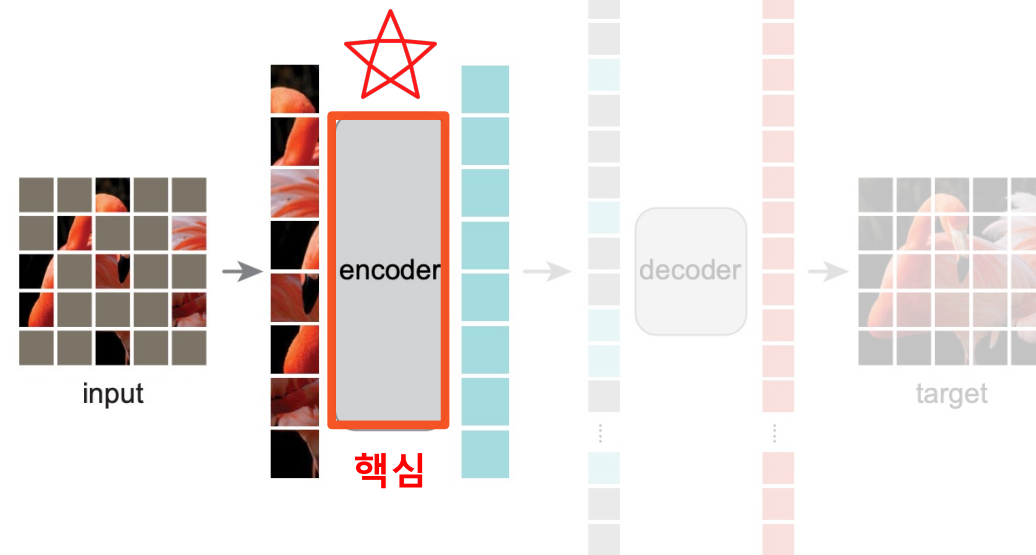
EfficientSAM

❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners



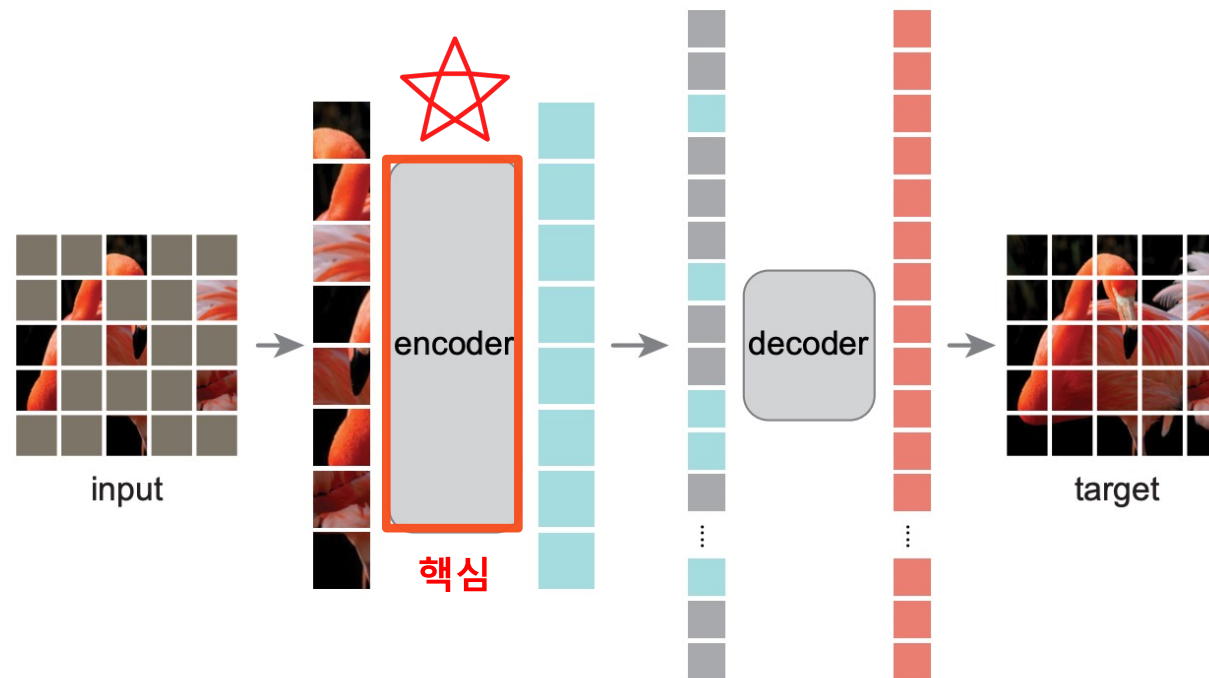
MAE reconstruction
Blurry ..



EfficientSAM

❖ MAE Architecture

- Masked Autoencoders Are Scalable Vision Learners

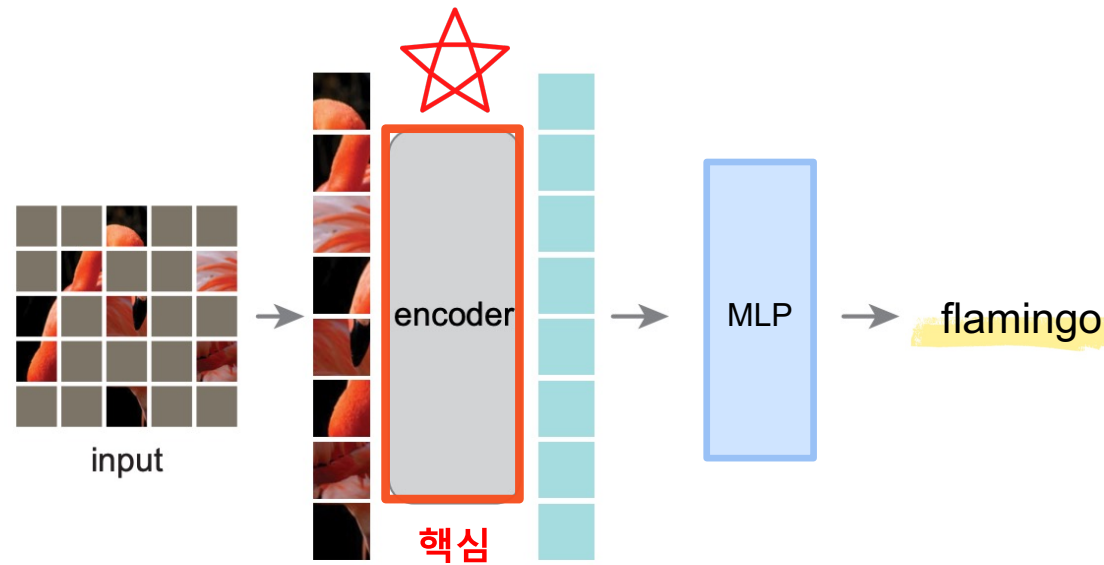


EfficientSAM

❖ MAE Architecture

Downstream Task

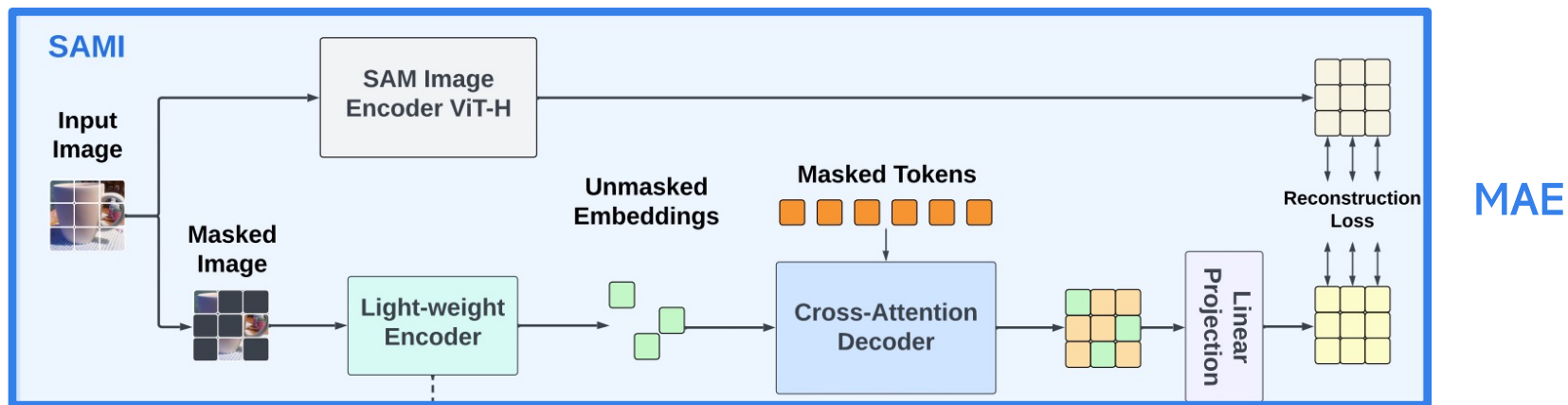
ex) Classification



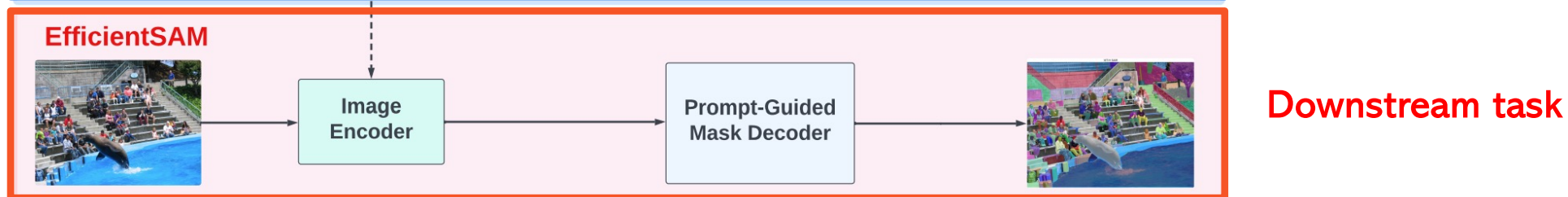
EfficientSAM

❖ 2-Stage

Stage 1
SAMI pretraining



Stage 2
SAM finetuning

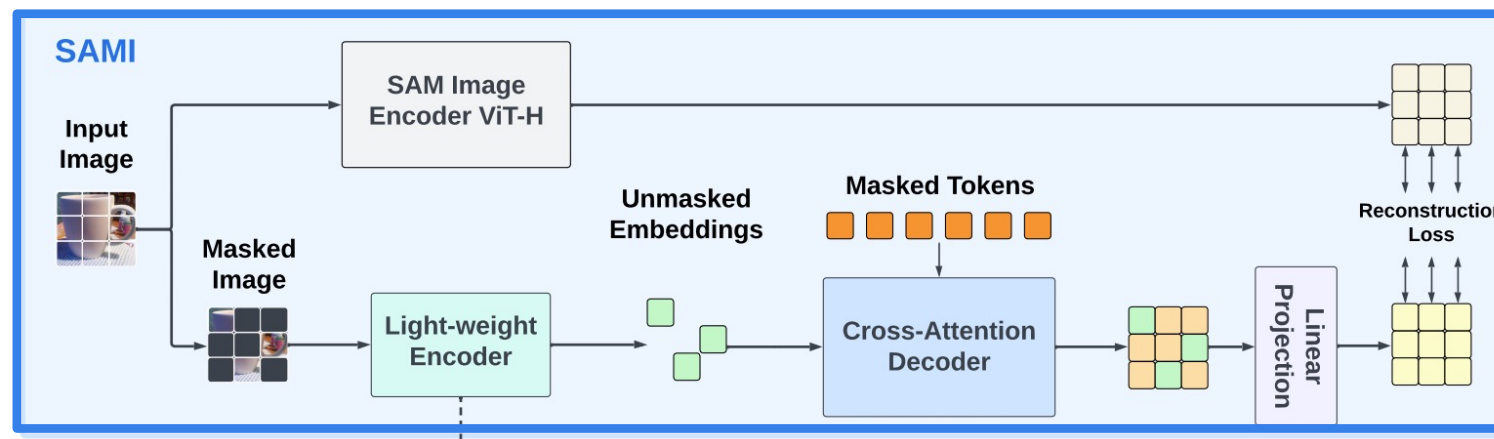


EfficientSAM

❖ 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

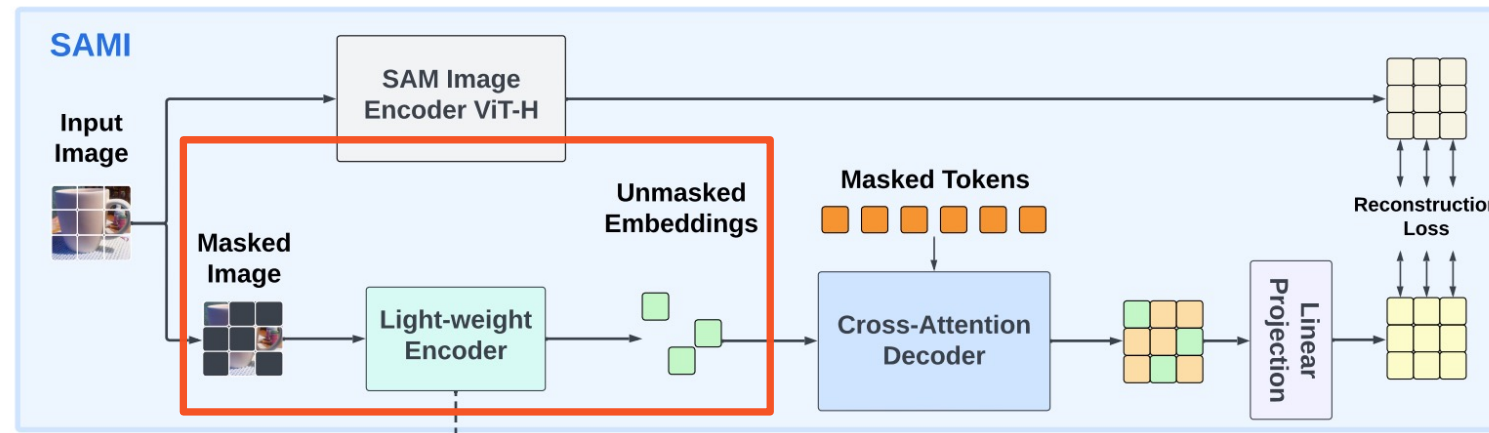


EfficientSAM

❖ 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



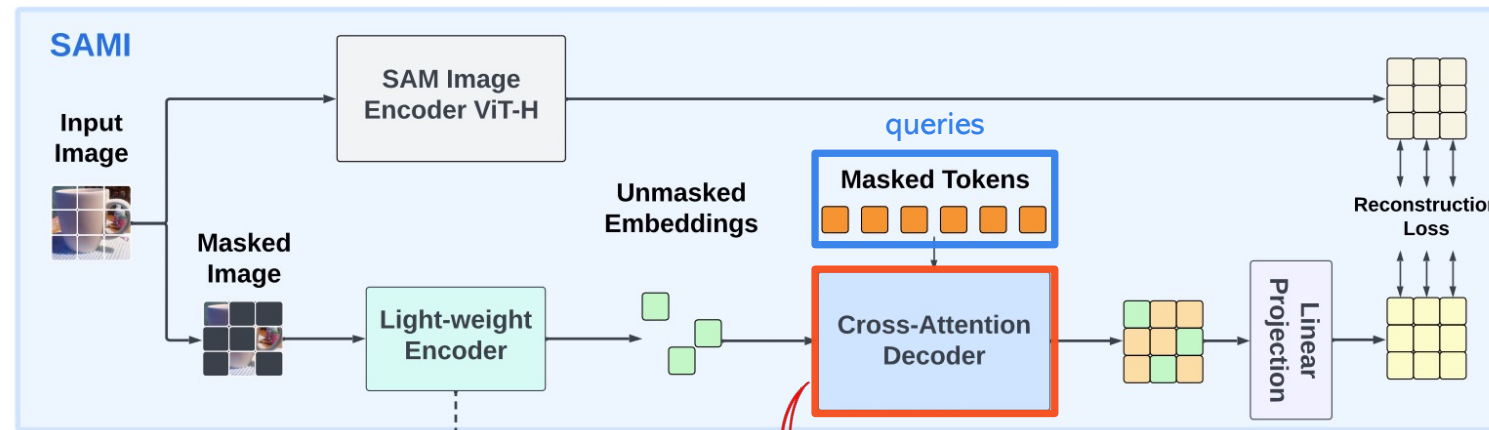
Encoder의 input:
unmasked patch

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

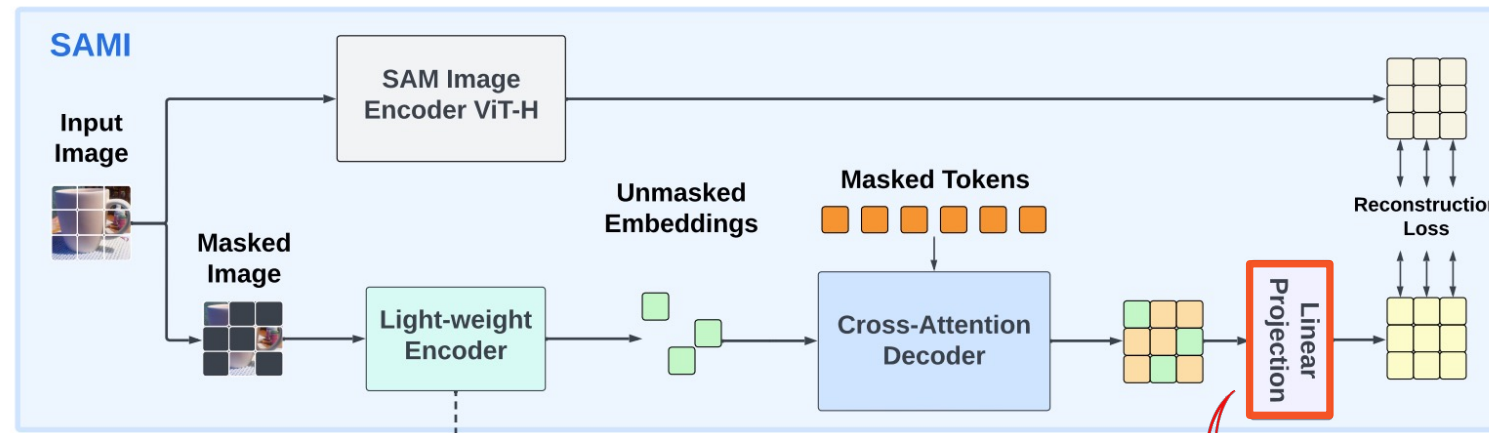
Keys & values: unmasked features from encoder & masked features

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



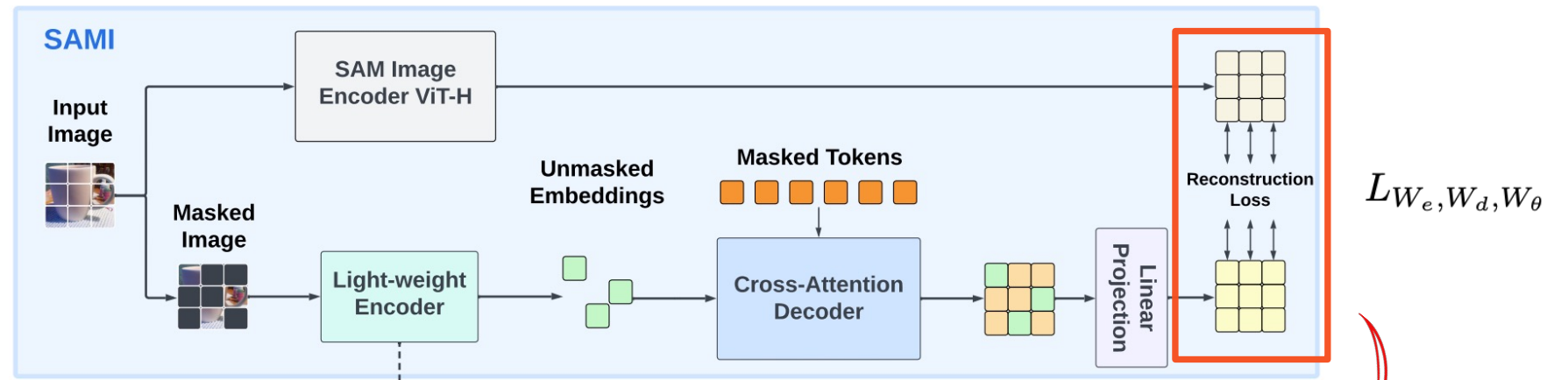
Linear Projection:
SAM image encoder & MAE output
Feature dimension mismatch 해결

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



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

where N is the number of input tokens

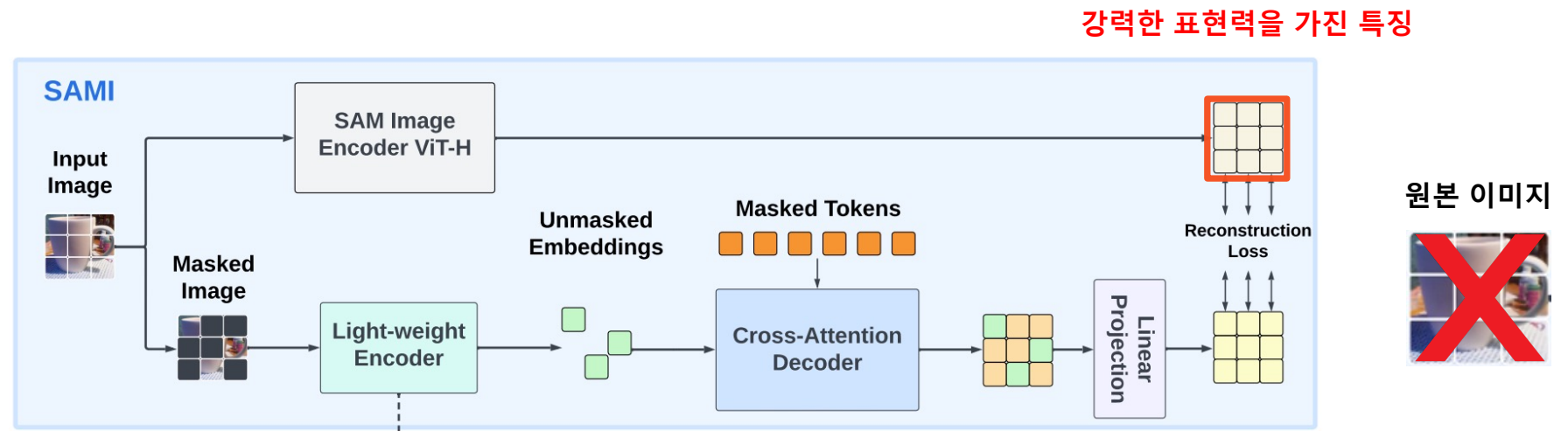


EfficientSAM

❖ SAMI pretraining (Stage 1)

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

Stage 1
SAMI pretraining

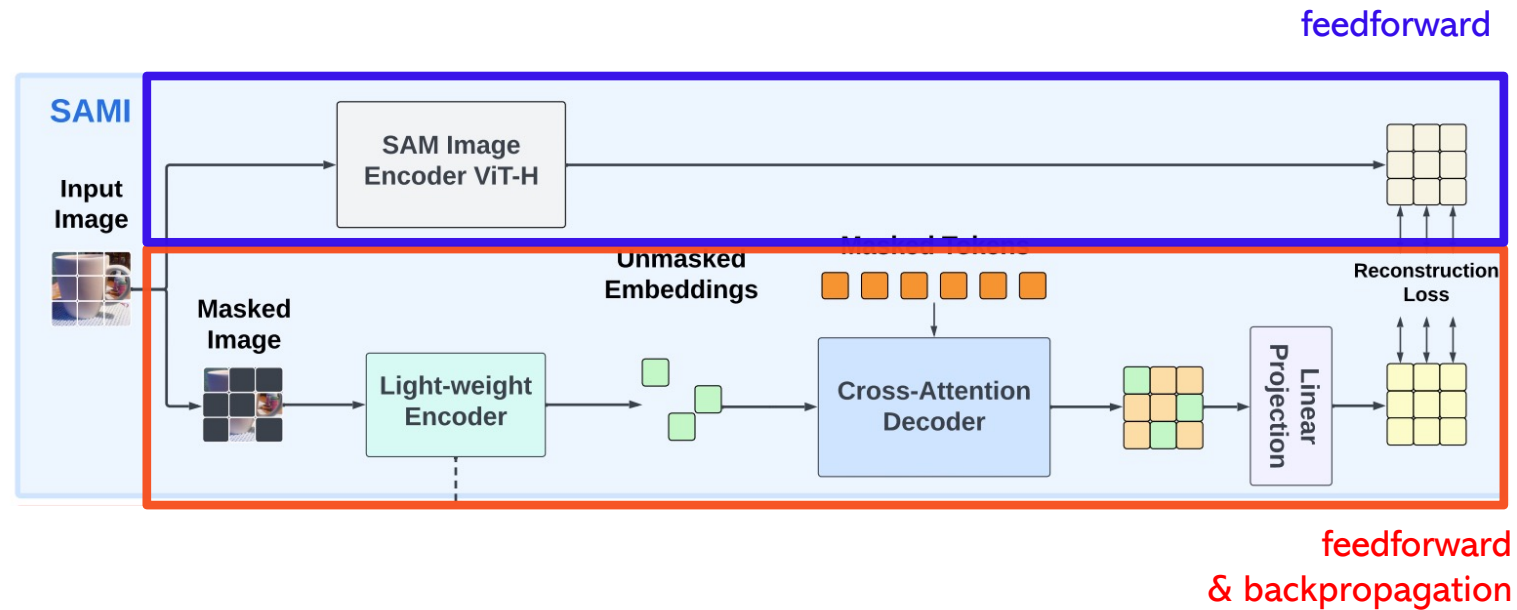


EfficientSAM

❖ SAMI pretraining (Stage 1)

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

Stage 1 SAMI pretraining

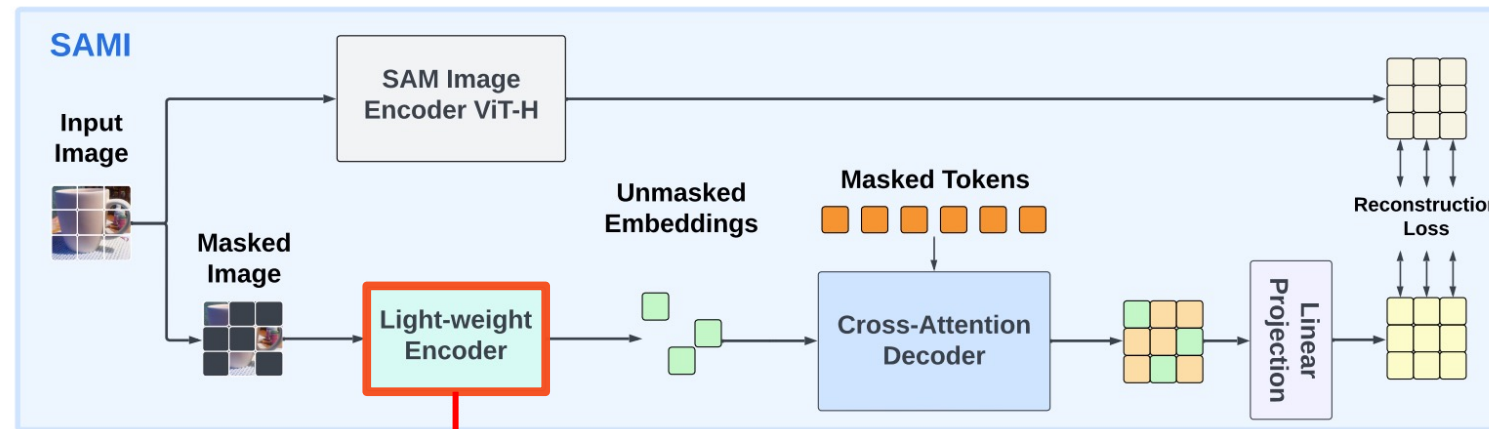


EfficientSAM

❖ SAMI pretraining (Stage 1)

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

Stage 1 SAMI pretraining

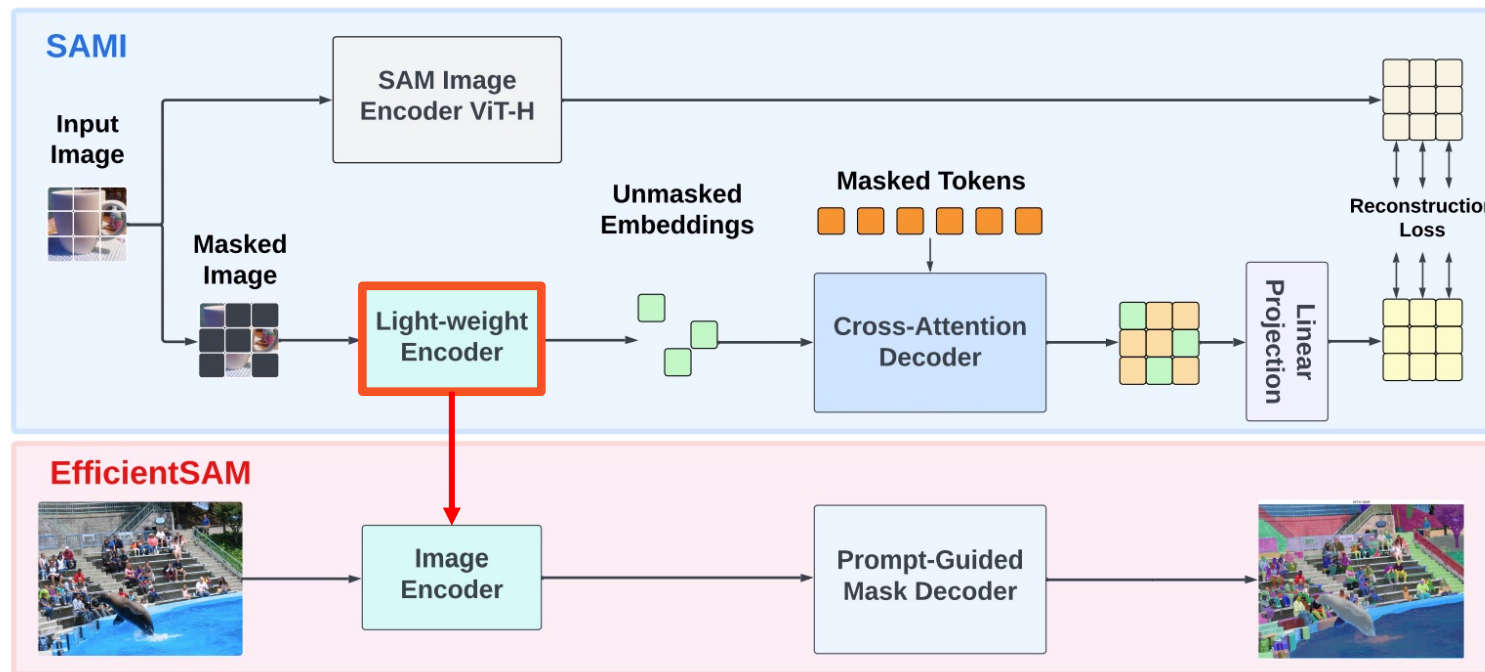


extract feature representations
for various vision task

EfficientSAM

❖ SAM finetuning (Stage 2)

- 경량화된 인코더를 SAM의 mask decoder와 결합하여 segmentation 작업 수행하도록 finetuning

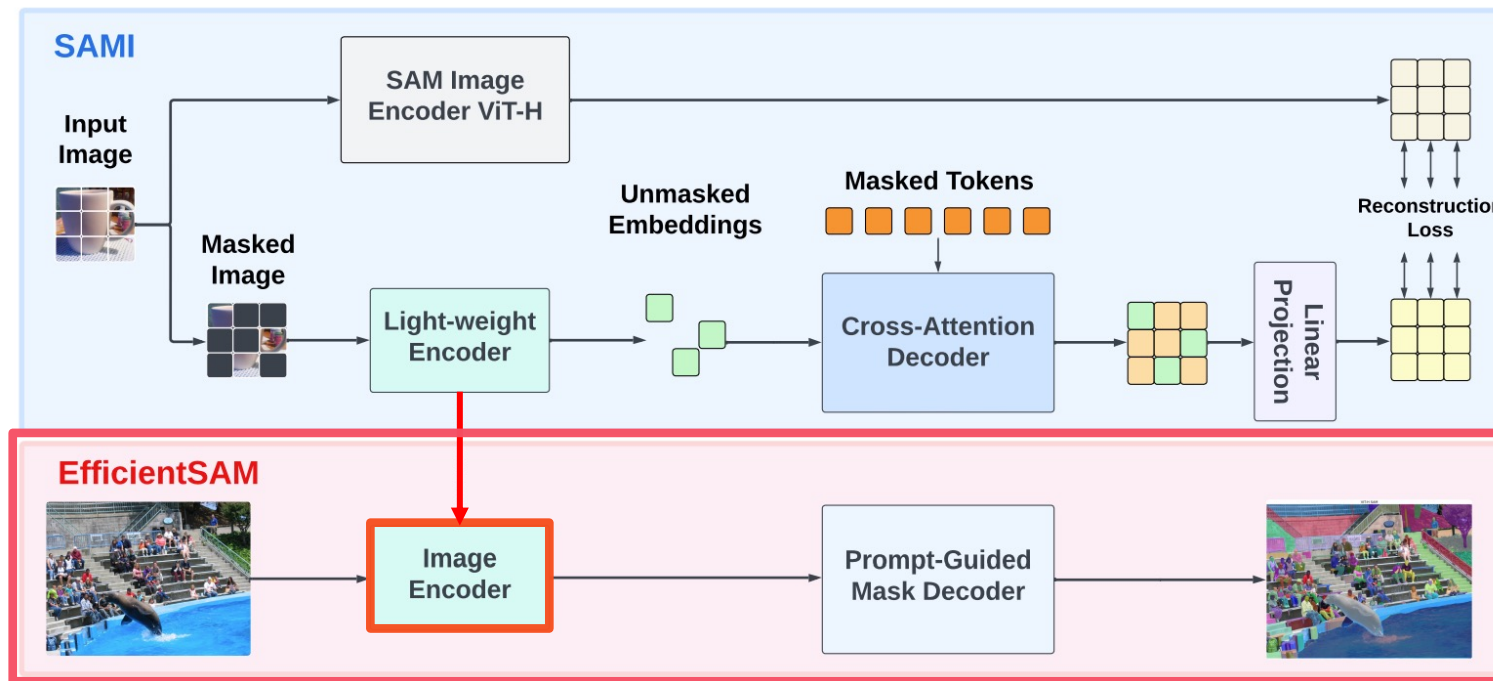


Stage 2
SAM finetuning

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

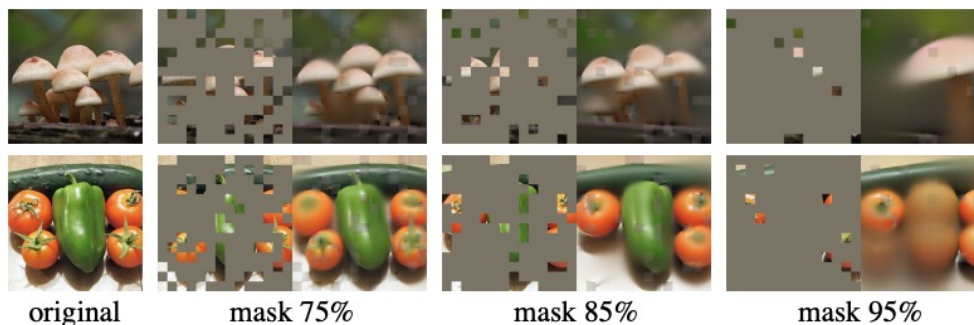
- Only underperforms SAM by 1.5 mIOU

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.



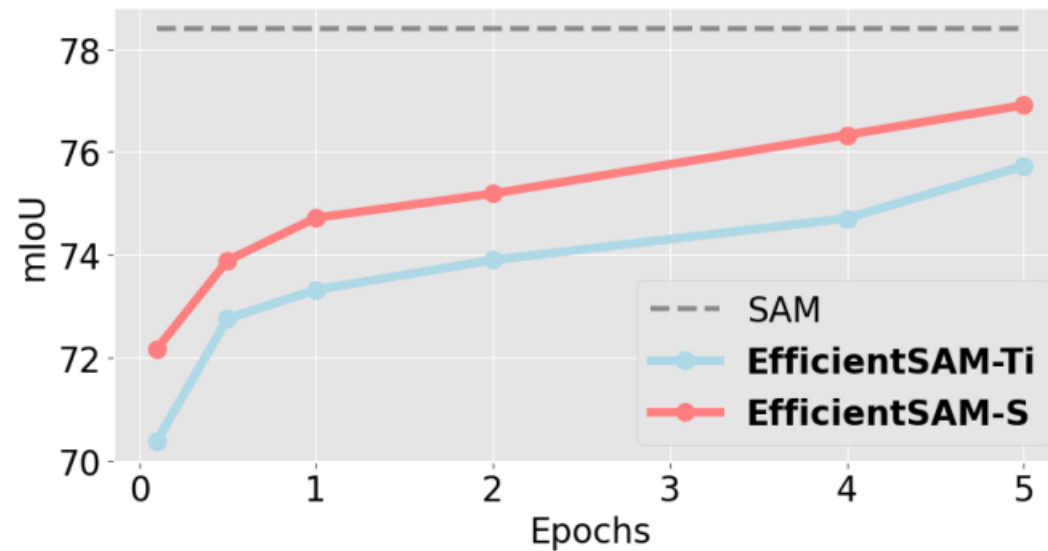
적은 정보만을 가지고도 더 강력한 일반화 능력을 갖도록 훈련되며, 과적합을 방지하고 효율적인 특징 표현을 학습

Main Results

❖ Ablation Studies

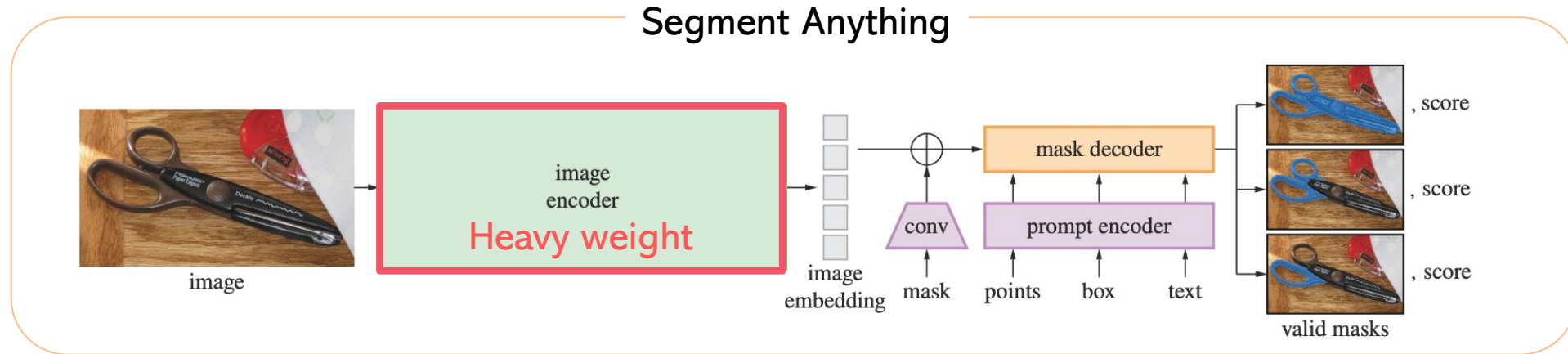
빠른 성능 향상

Advantages of SAMI-pretrained image encoders

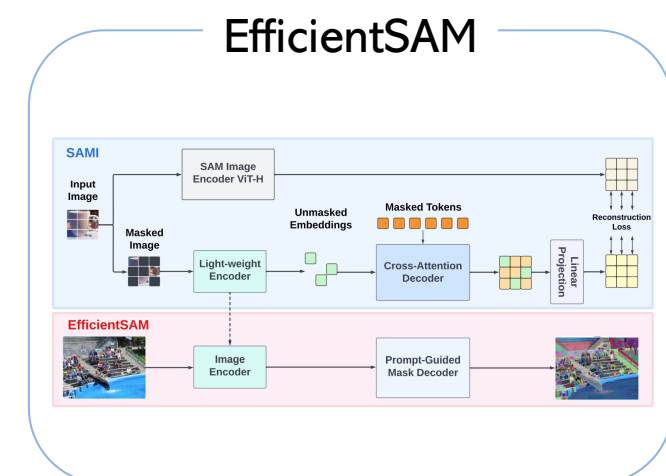
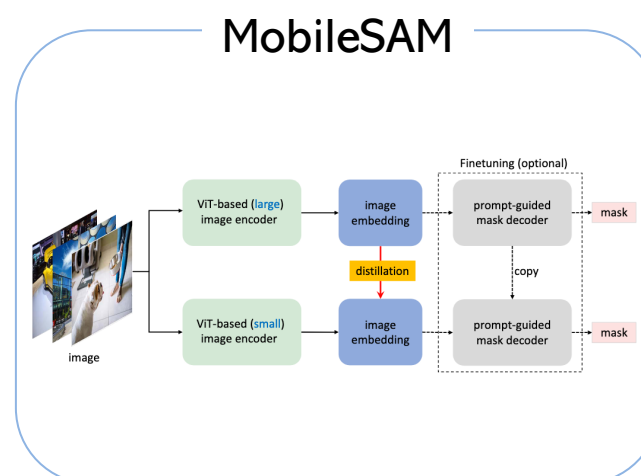
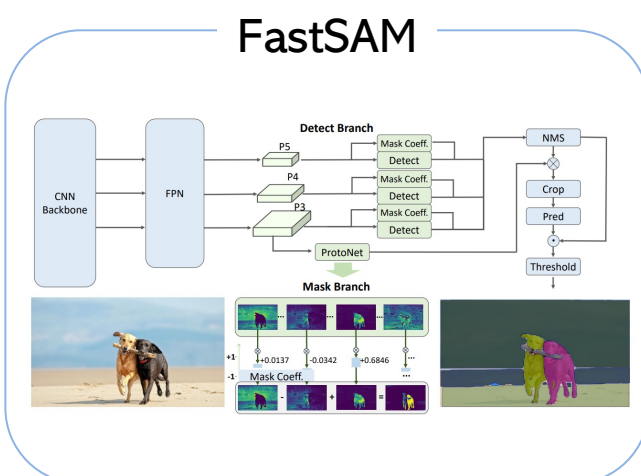
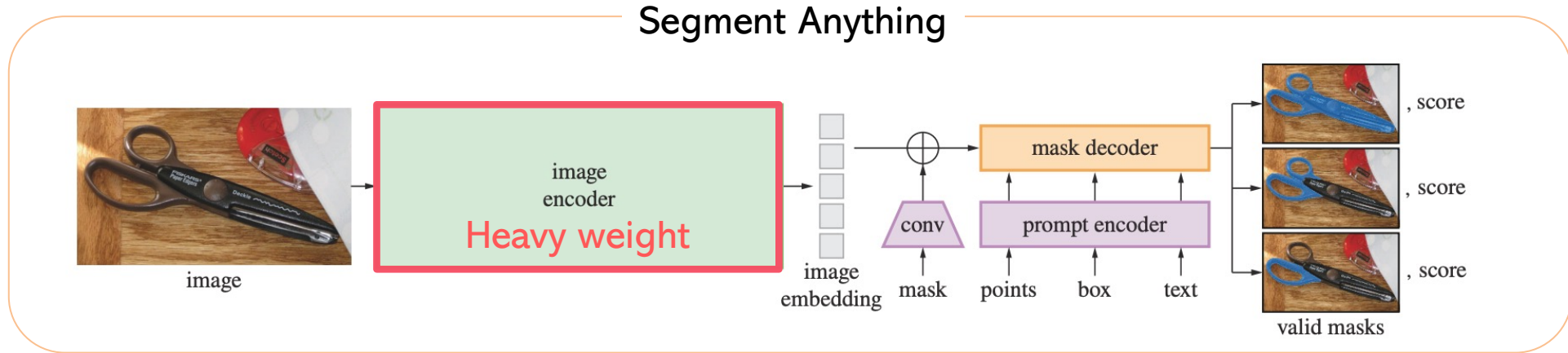


EfficientSAM 모델이 SAMI를 활용한 사전 학습된 이미지 인코더 덕분에 빠르게 좋은 성능을 발휘할 수 있음

Conclusion



Conclusion



Thank you

