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

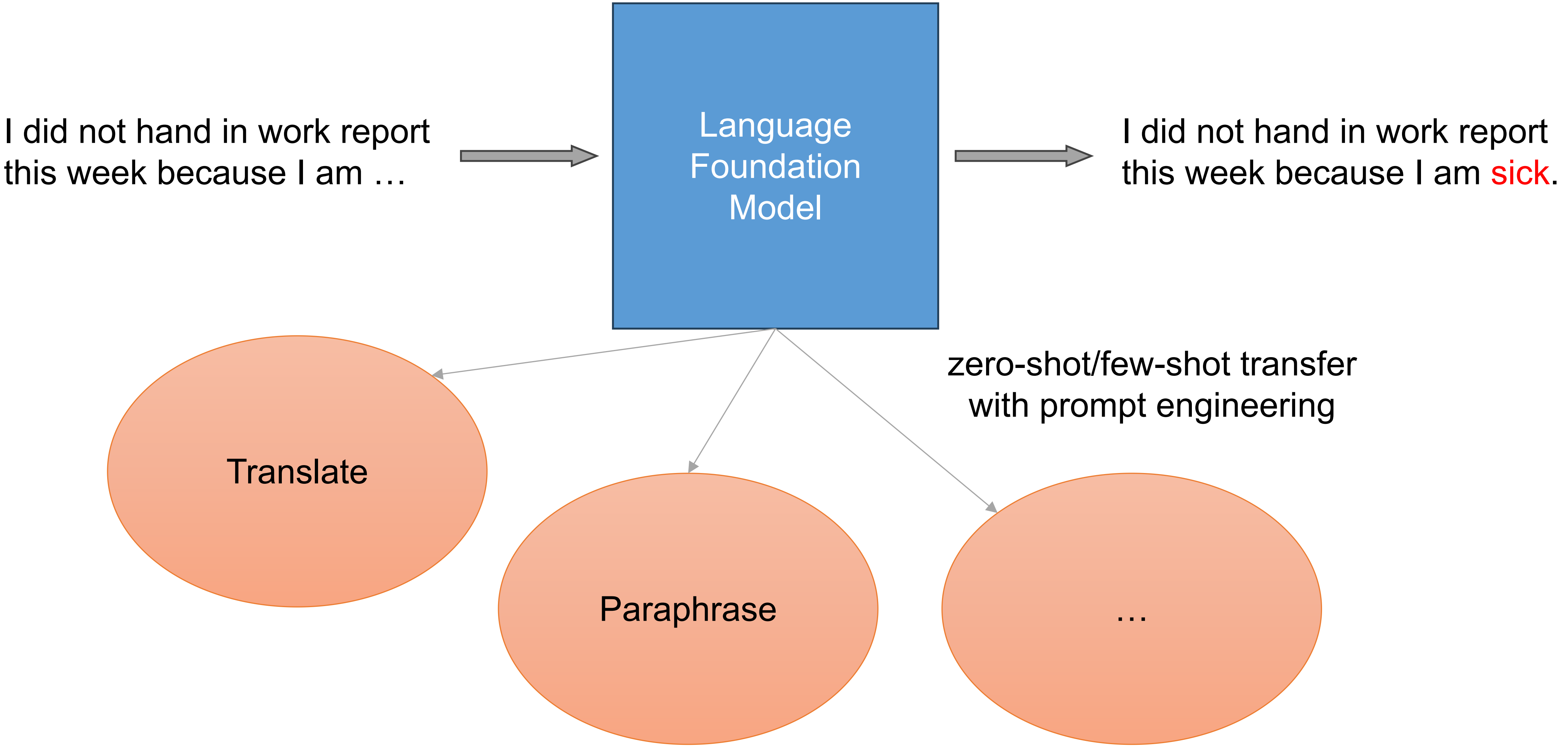
CVPR 2024

Presenter: Chenyu Niu
2025.03.02

Background

PEKING UNIVERSITY

NLP-Prompting and Foundation Model



NLP-Prompting and Foundation Model

I did not hand in work report this week because I am …

Language Foundation Model

I did not hand in work report this week because I am sick.

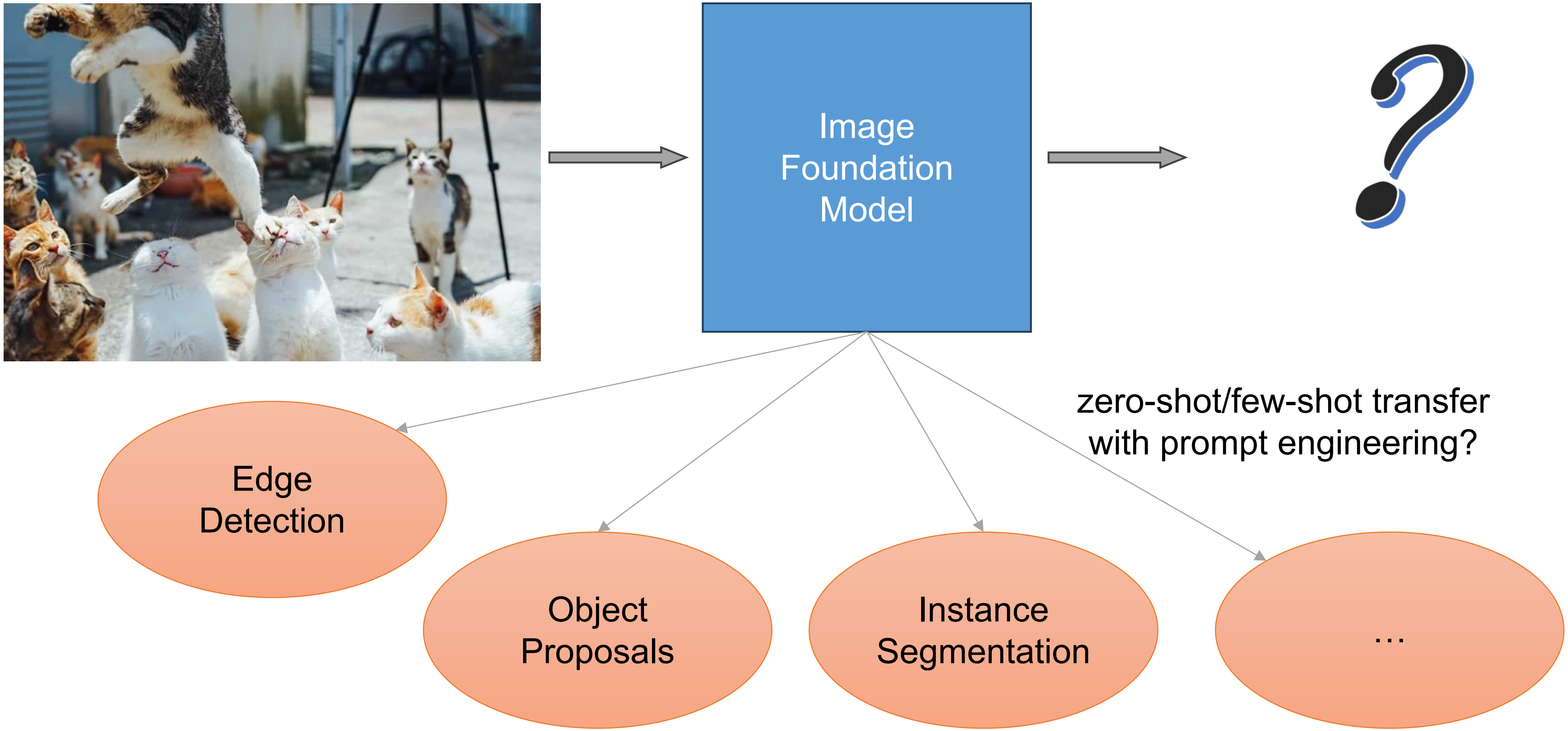
Why is this possible?

- Text data is available at web scale
- No labeling is needed for sequence prediction

Background

PEKING UNIVERSITY

What about Computer Vision Field?



Background

What about Computer Vision Field?

Image Foundation Model

- Image data is available at web scale 😋
- Labeling is NEEDED for many problems 🤔

Background

What about Computer Vision Field?

Image Foundation Model

zero-shot/few-shot transfer
with prompt engineering

Edge Detection

Object Proposals

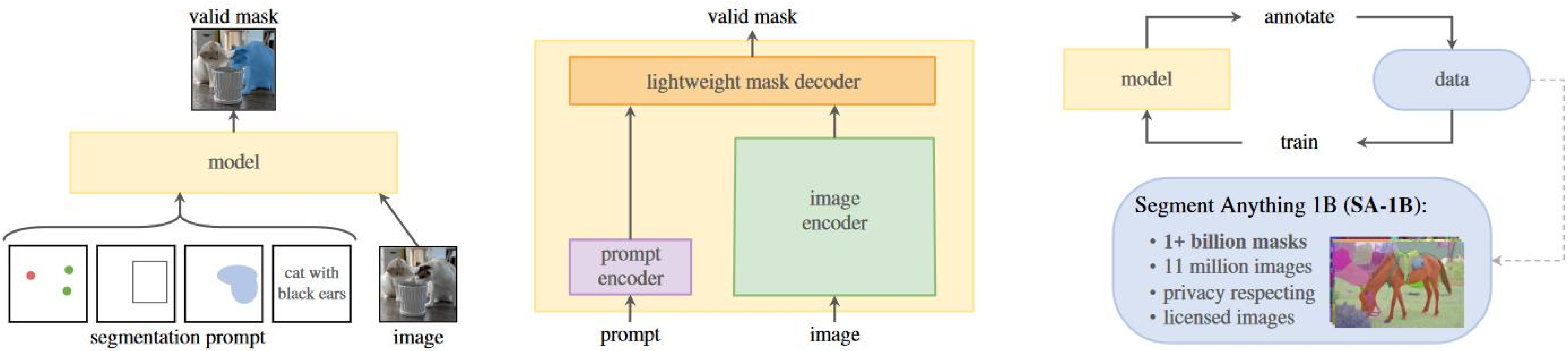
Instance Segmentation

…

Background

Introducing Segment Anything Model (SAM)

Develop a promptable model and pre-train it on a broad dataset using a task that enables powerful generalization.



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

What **task** will enable zero-shot generalization?

What is the corresponding **model** architecture?

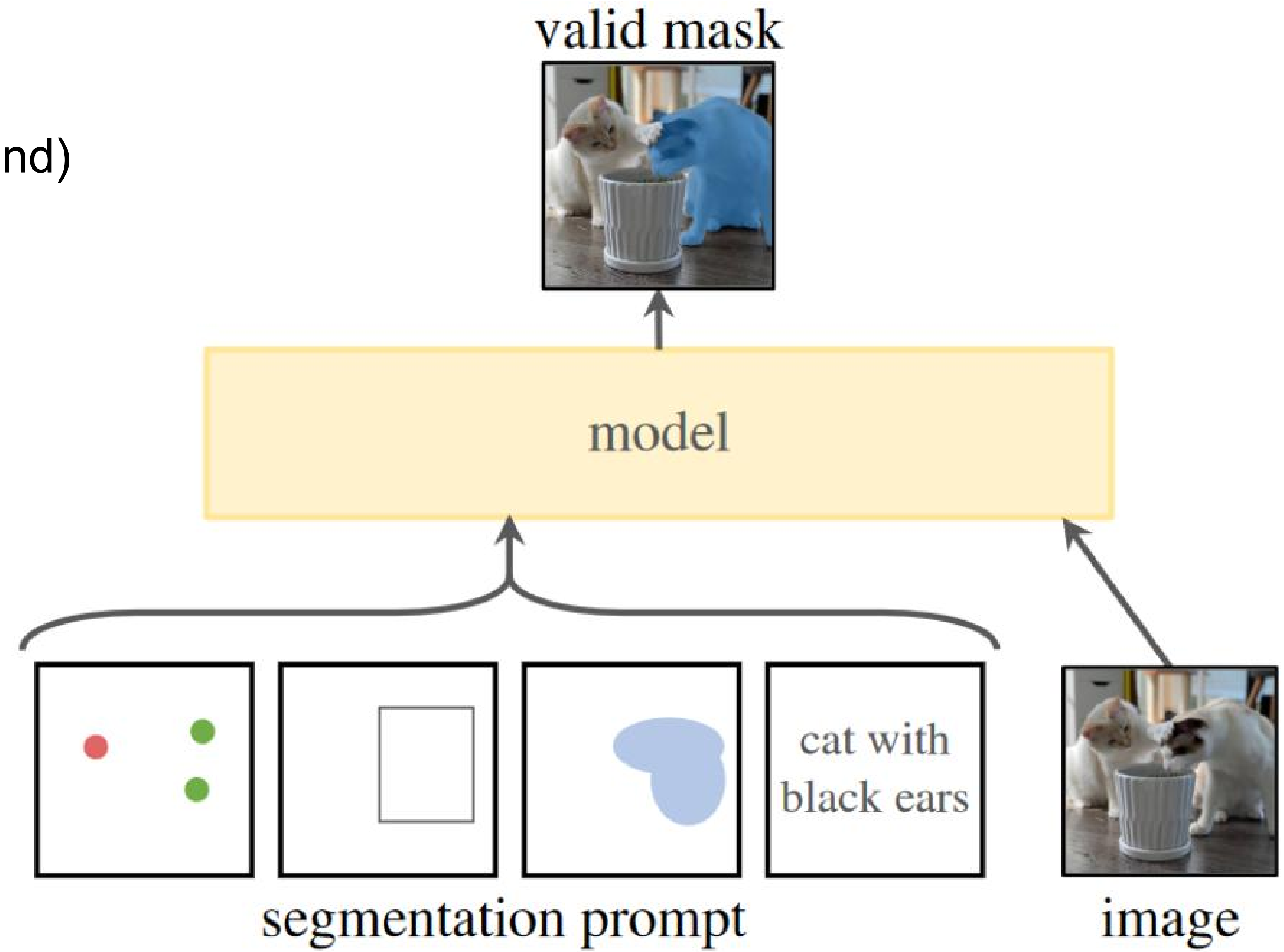
What **data** can power this task and model?

How to "Prompt" a Segmentation Task?

- Sparse Prompt
  - Point (foreground / background)
  - Box
  - Text

- Dense Prompt
  - Mask

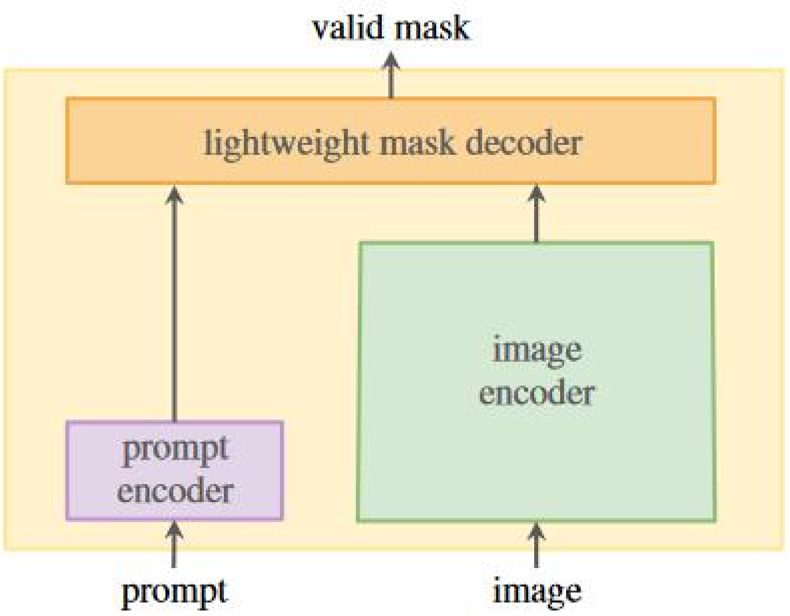
Final Goal:
Given any segmentation *prompt*, return a *valid* segmentation mask

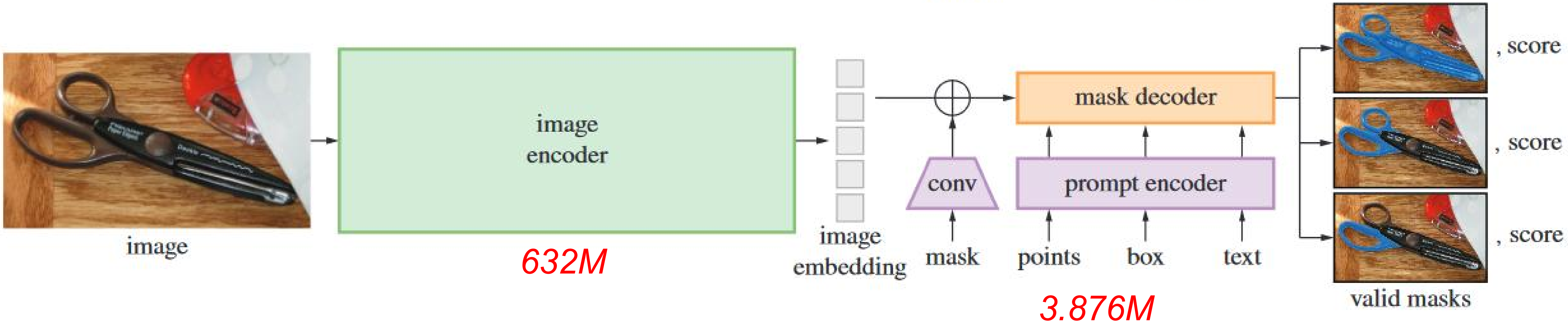


Background (SAM Model)

SAM Architecture

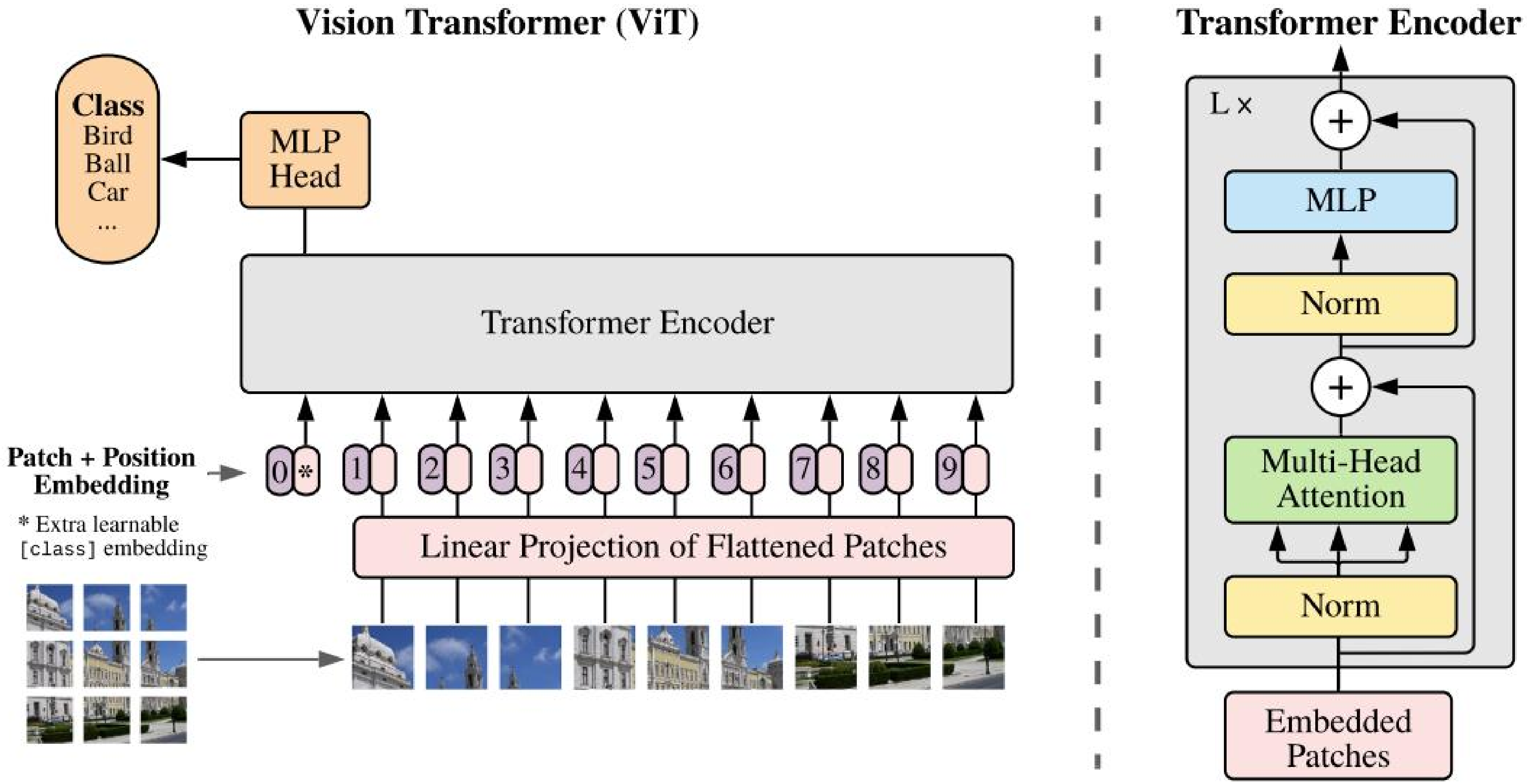
- A *heavy* image encoder
- A prompt encoder
- A *lightweight* mask decoder





Background (SAM Model)

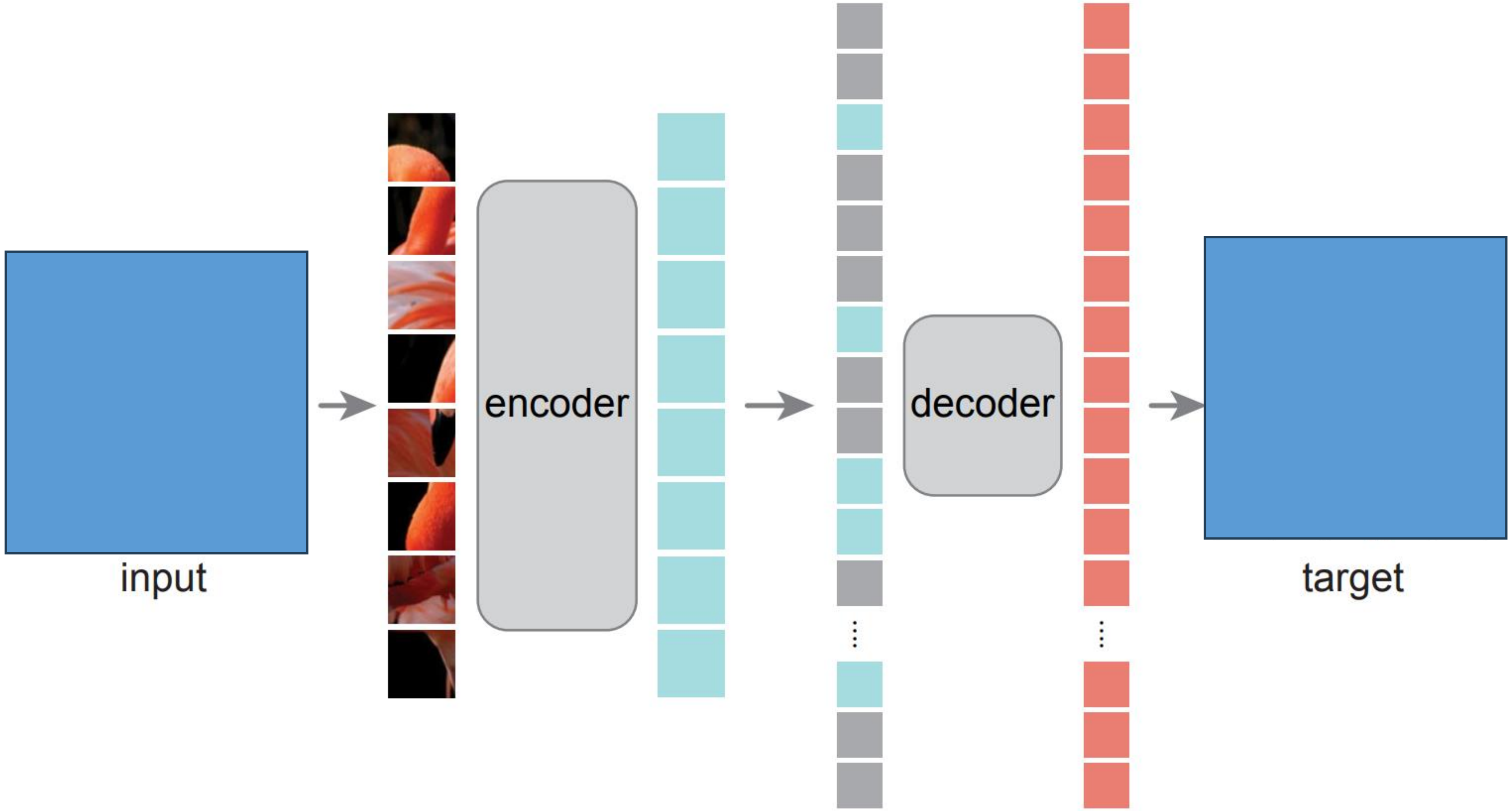
Vision Transformers (ViT)



- Patchify images as token sequences
- Transformer encoder for classification
- Broader spatial correlation

Dosovitskiy et al, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021

Masked Autoencoders (MAE)



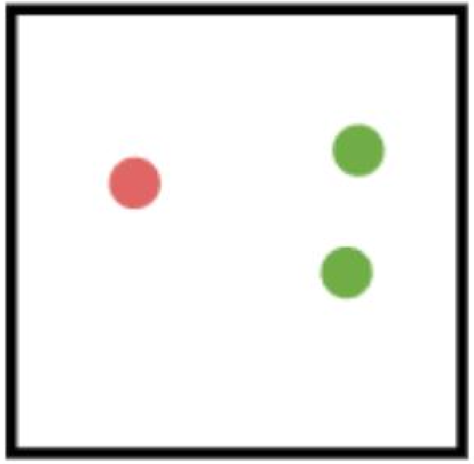
- A large random subset of image is masked out
- Asymmetric encoder-decoder design
- Large models can be trained efficiently and effectively

He et al, Masked Autoencoders Are Scalable Vision Learners, CVPR 2022

Background (SAM Model)

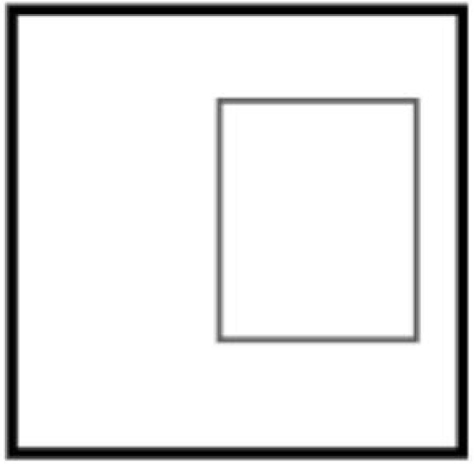
Prompt Encoder

Point



- Positional encoding of the point
- Foreground or background

Box



- Embedding pair
- Top-left corner
- Bottom-right corner

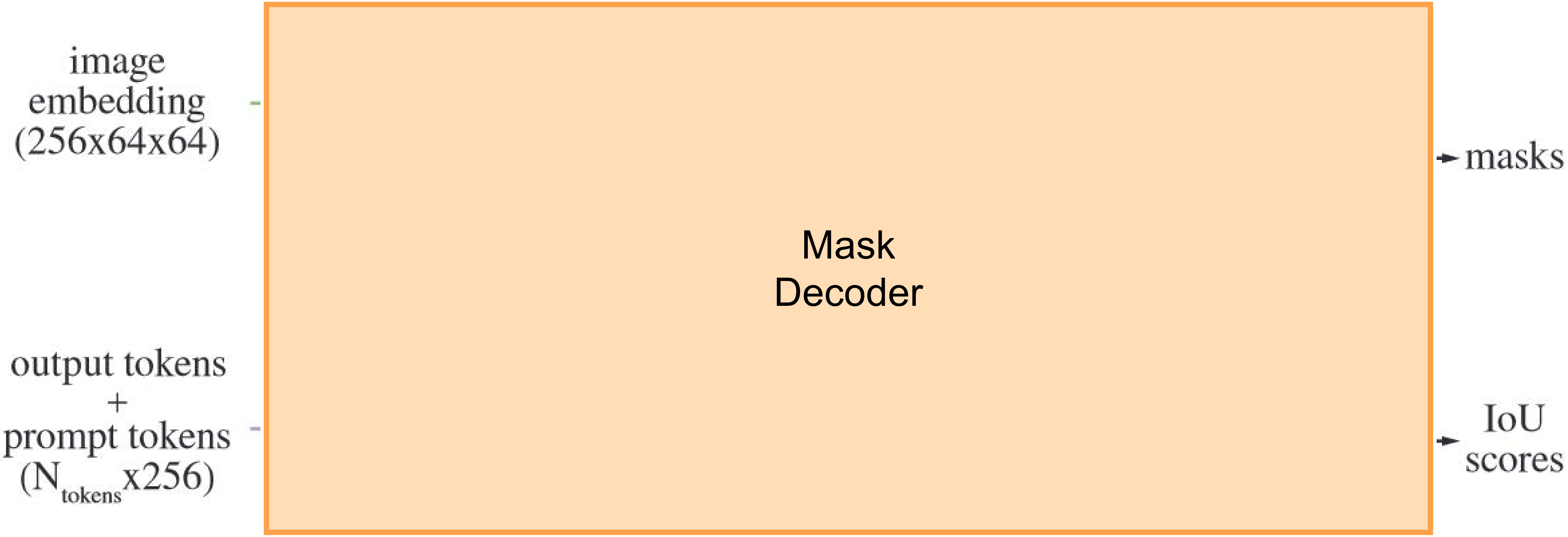
Text

CLIP

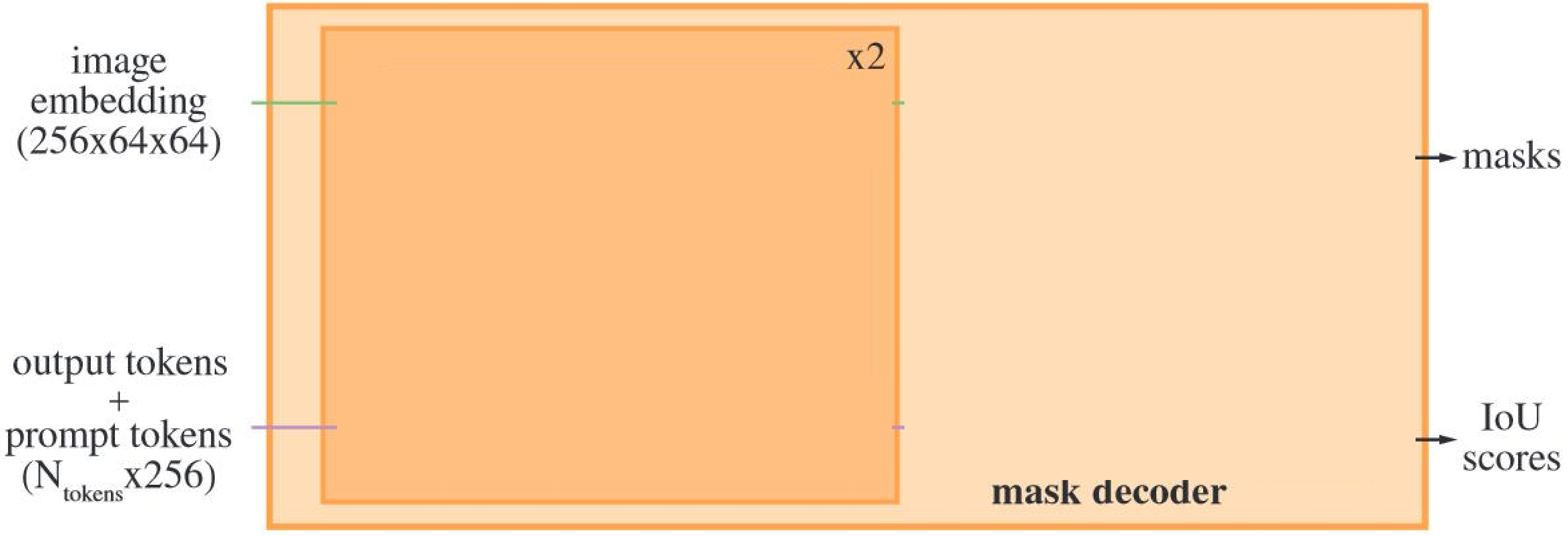
Mask

- Downscaled using CNN
- Added to image-embedding element-wise
- “No mask” embedding

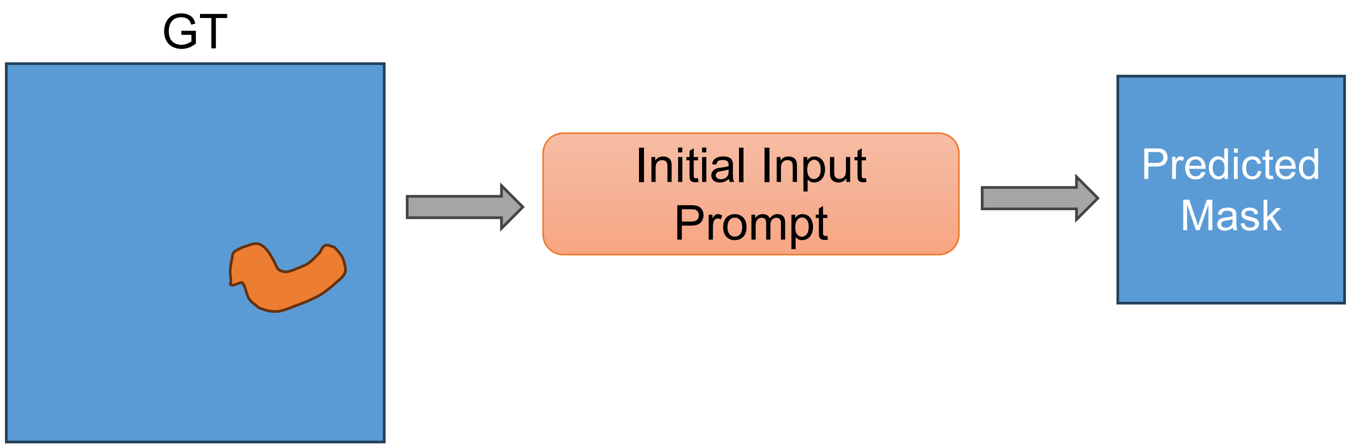
Mask Decoder



Mask Decoder



Training Algo: an Interactive Segmentation Setup



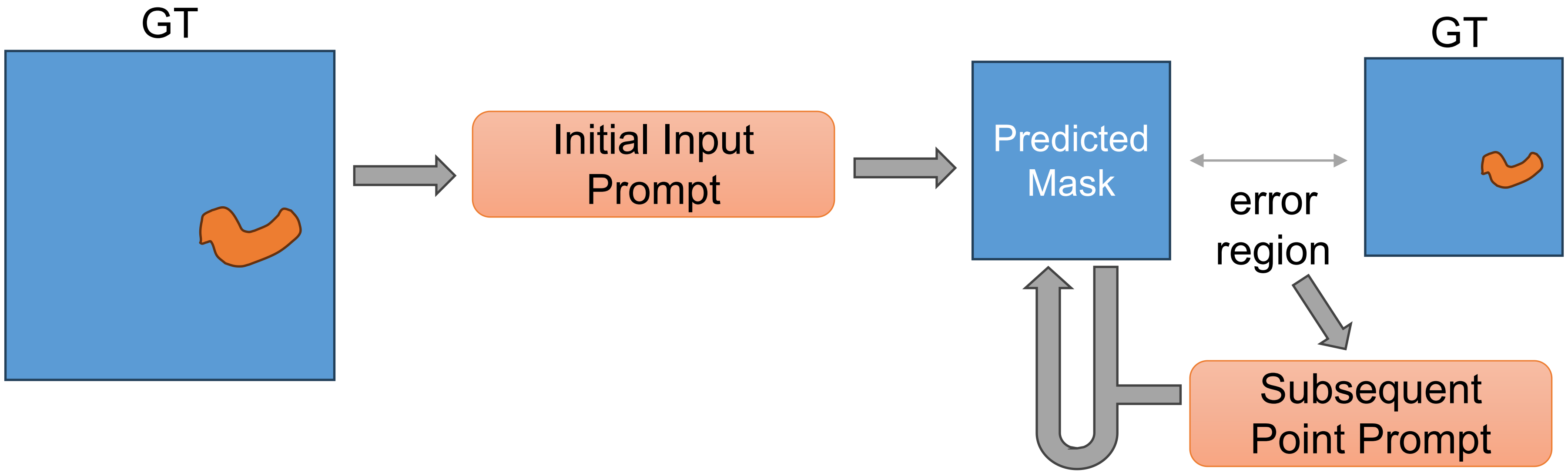
Sample initial prompt

- Randomly select a foreground point/box from ground truth mask
- Make the prediction

Background

北京大學 PEKING UNIVERSITY

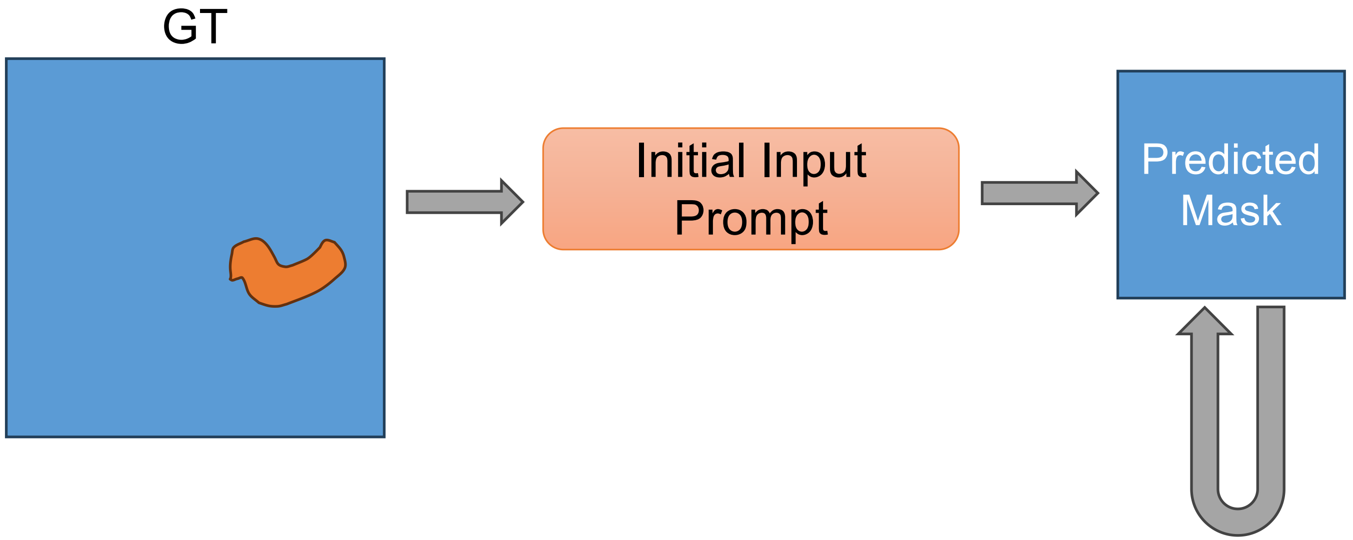
Training Algo: an Interactive Segmentation Setup



Iteratively provide subsequent points (8 iterations)

- Given the predicted result of last iteration (unthreshold mask logits for maximal information)
- Subsequent points selected from error region
  - **Foreground** point for false negative
  - **Background** point for false positive
- Make the prediction

Training Algo: an Interactive Segmentation Setup

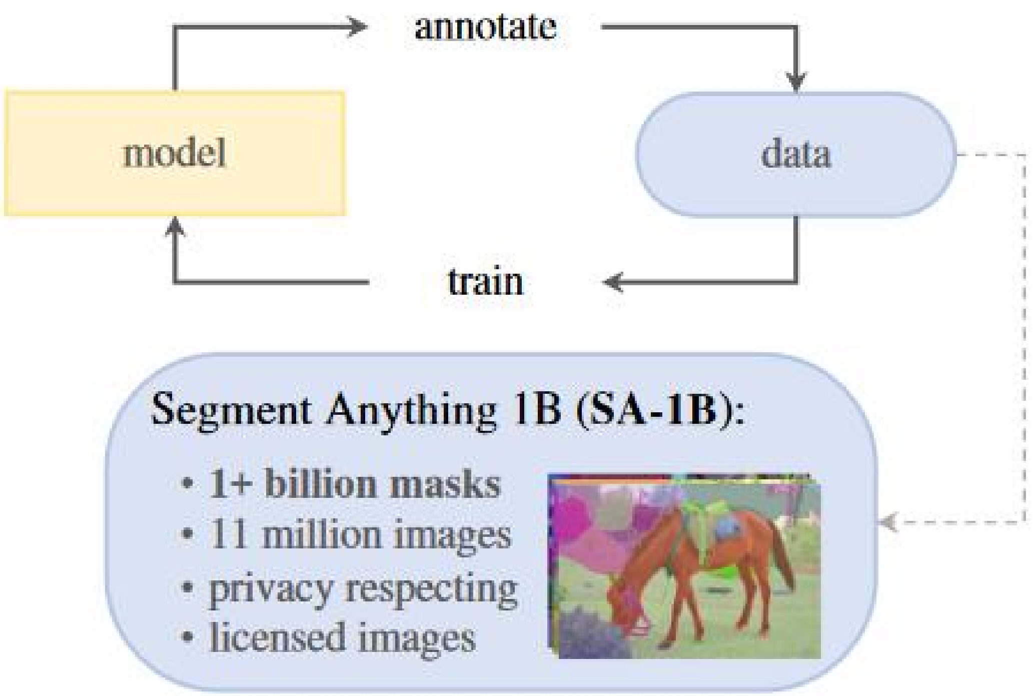


No new information is supplied (2 iterations)

- Given the predicted result of last iteration
- Make the prediction
- One in the middle, one at last

Background (SAM Data)

Data Engine



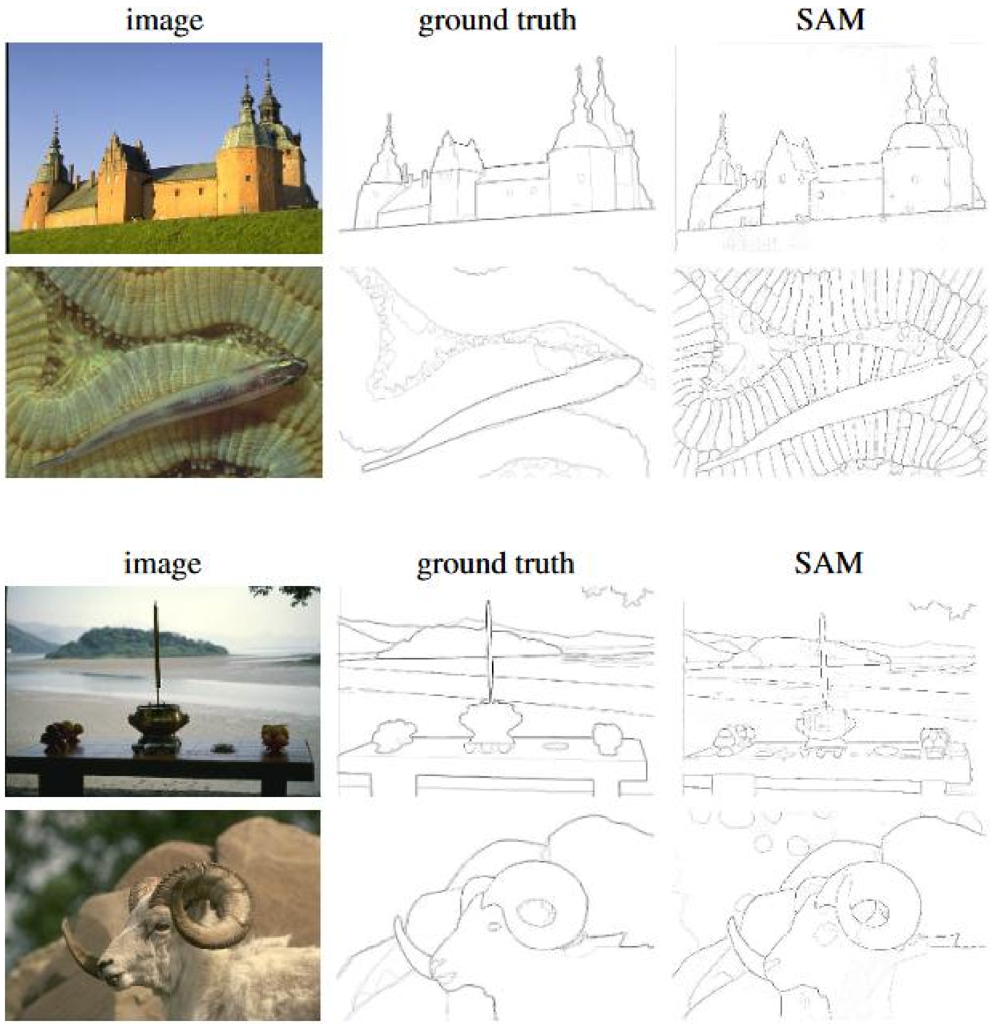
- Assisted-manual stage
- Semi-automatic stage
- Fully automatic stage
  - Only used for data generation
  - Using a special version of SAM

Background

Zero-Shot Transfer Examples

Edge Detection

- Prompt SAM with 16×16 grid of foreground points resulting in 768 masks
- Remove redundant masks by NMS
- Use Sobel filter on unthresholded mask probability map
- Standard lightweight postprocessing



Zero-Shot Transfer Examples

Object Proposals

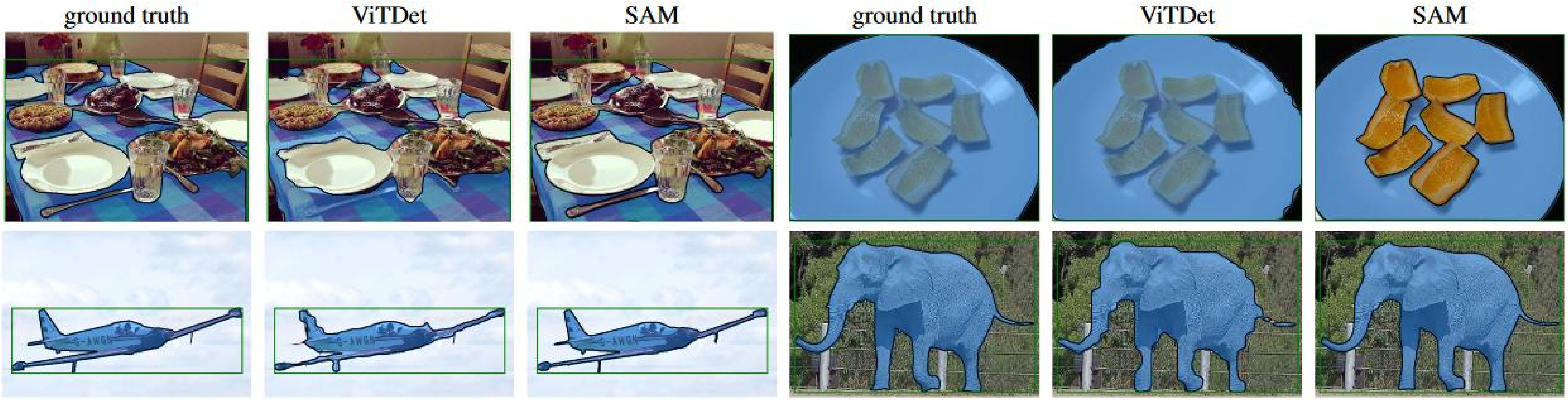
- Prompt SAM with 64×64 grid of foreground points
- Remove redundant masks by NMS
- Rank mask by the average of confidence and stability scores to get top 1000 masks

	mask AR@1000						
method	all	small	med.	large	freq.	com.	rare
ViTDet-H [62]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
zero-shot transfer methods:							
SAM – single out.	54.9	42.8	76.7	74.4	54.7	59.8	62.0
SAM	59.3	45.5	81.6	86.9	59.1	63.9	65.8

Zero-Shot Transfer Examples

Instance Segmentation

- Prompt SAM with boxes output by ViTDet-H
- An additional mask refinement iteration



What Problem Remains?

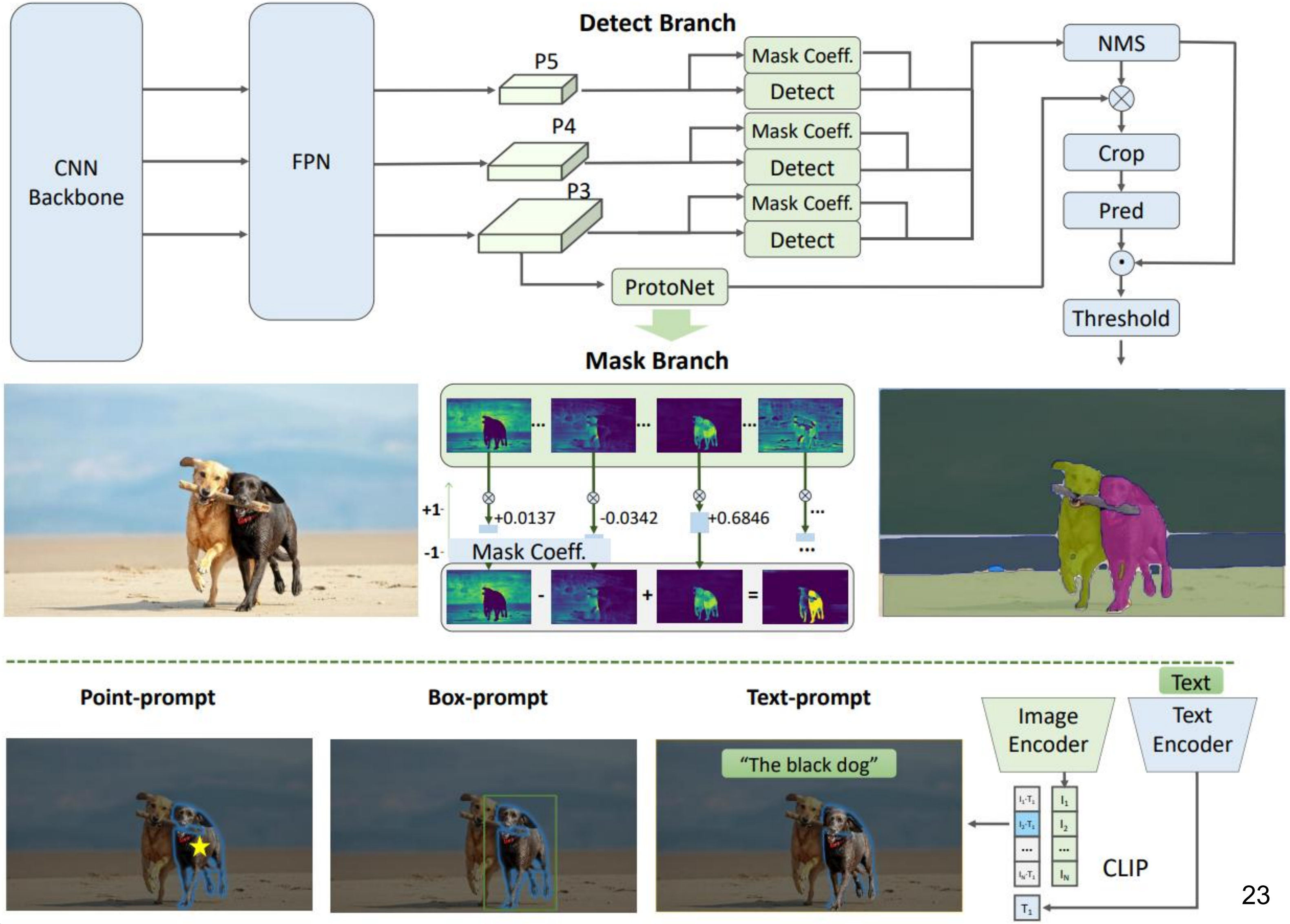
SLOW! 🐌🐌🐌

About 2 images/second on a single NVIDIA A100 with one box prompt…

Parameters	Original SAM
Image encoder	632M
prompt-guided decoder	3.876M
Speed	0.452s

FastSAM

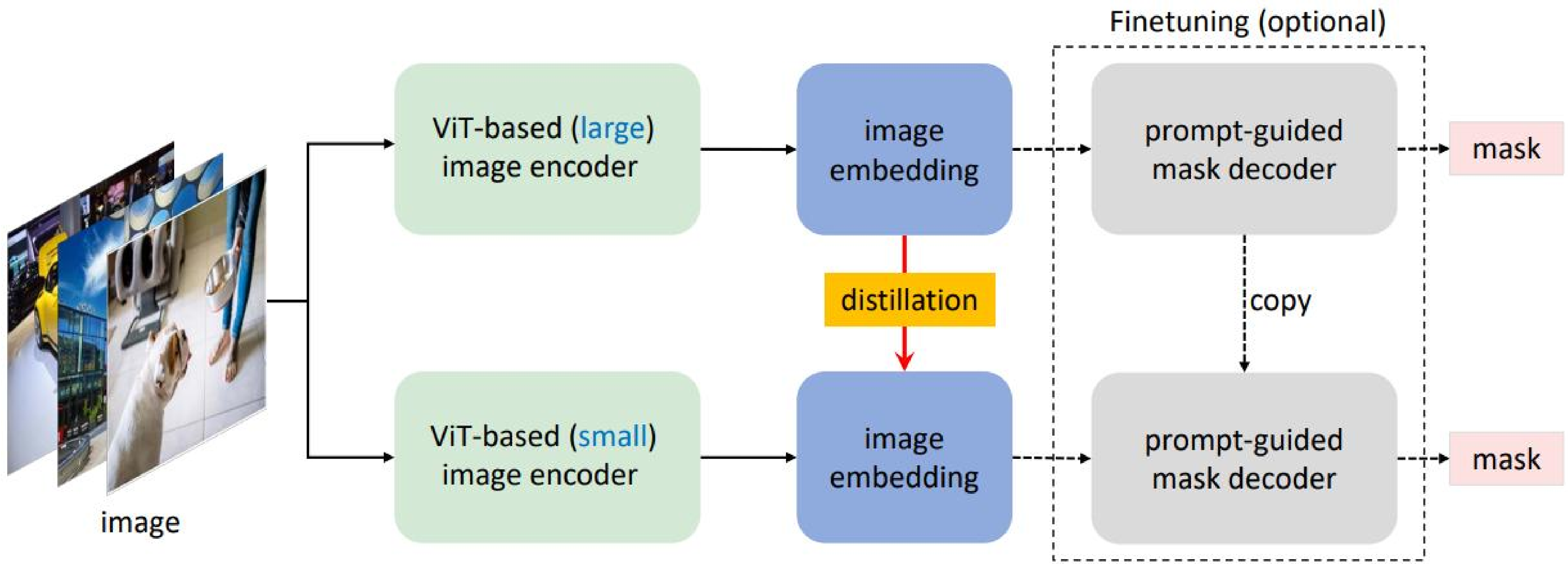
- All-instance segmentation based on YOLOv8-seg
- Prompt-guided selection



Zhao et al, Fast Segment Anything, arXiv:2306.12156

PEKING UNIVERSITY

MobileSAM

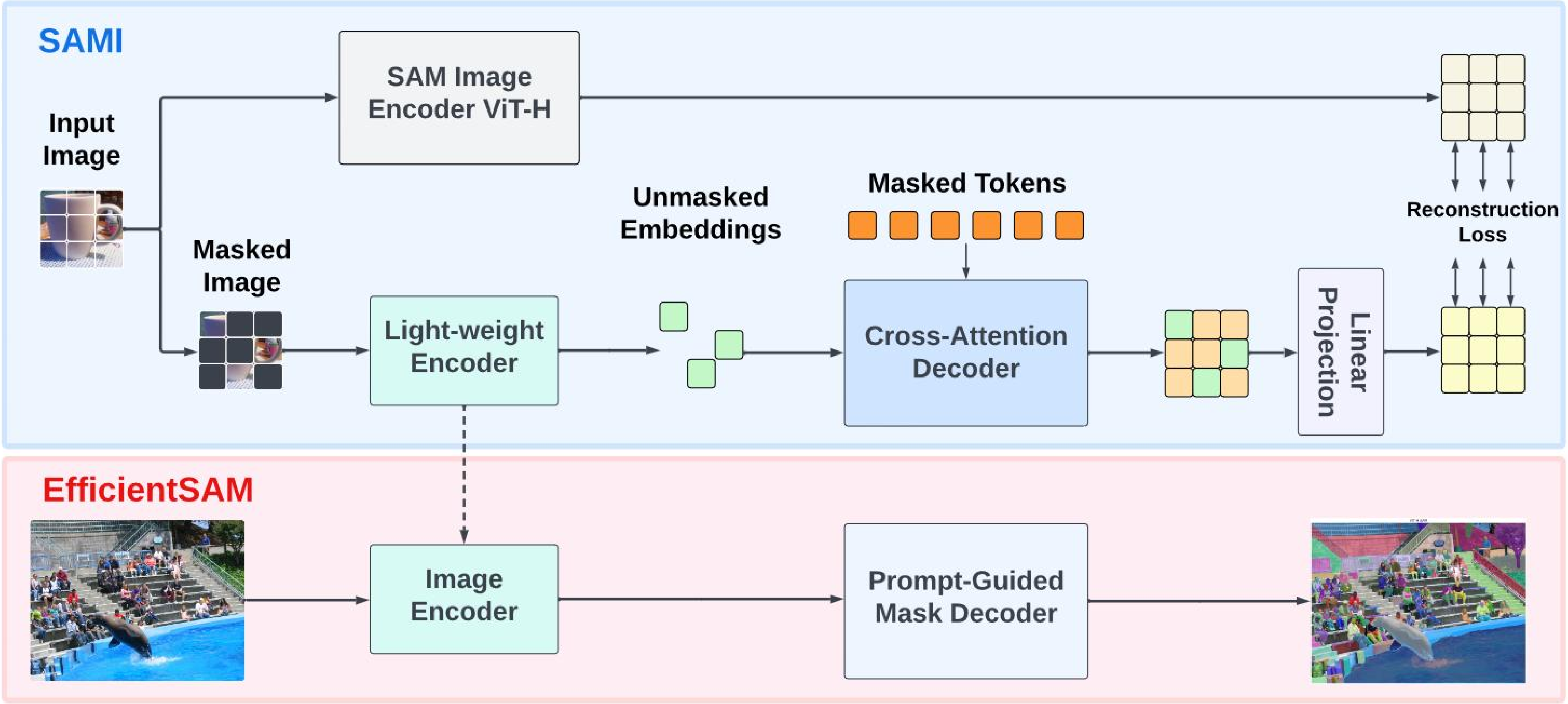


- Distill the knowledge from the default ViT-H encoder to a tiny ViT encoder
- Finetuning on the decoder is optional

Zhang et al, Faster Segment Anything: Towards Lightweight SAM for Mobile Applications, arXiv:2306.14289

PEKING UNIVERSITY

EfficientSAM Framework



Experimental Settings

- Pretraining datasets: ImageNet-1K training set with 1.2M images
- Finetune on various downstream tasks
  - Image classification
  - Object detection and instance segmentation
  - Semantic segmentation
  - Segment anything

Experiments

北京大学 PEKING UNIVERSITY

Results for *SAMI*

Image Classification

Method	Backbone	Training Data	Acc.(%)
DeiT-Ti[53]	ViT-Tiny	IN1K	74.5
SSTA-Ti[60]	ViT-Tiny	IN1K	75.2
DMAE-Ti[2]	ViT-Tiny	IN1K	70.0
MAE-Ti[26]	ViT-Tiny	IN1K	75.2
SAMI-Ti (ours)	ViT-Tiny	SA1B (11M) + IN1K	**76.8**
DeiT-S[53]	ViT-Small	IN1K	81.2
SSTA-S[60]	ViT-Small	IN1K	81.4
DMAE-S[2]	ViT-Small	IN1K	79.3
MAE-S[26]	ViT-Small	IN1K	81.5
BEiT-S[3]	ViT-Small	D250M+IN22K+IN1K	81.7
CAE-S[12]	ViT-Small	D250M+IN1K	82.0
DINO-S[6]	ViT-Small	IN1K	82.0
iBOT-S[73]	ViT-Small	IN22K+IN1K	82.3
SAMI-S (ours)	ViT-Small	SA1B (11M) + IN1K	**82.7**
DeiT-B[53]	ViT-Base	IN1K	83.8
DMAE-B[2]	ViT-Base	IN1K	84.0
BootMAE[18]	ViT-Base	IN1K	84.2
MAE-B[26]	ViT-Base	IN1K	83.6
BEiT-B[3]	ViT-Base	D250M+IN22K+IN1K	83.7
CAE-B[12]	ViT-Base	D250M+IN1K	83.9
DINO-B[6]	ViT-Base	IN1K	82.8
iBOT-B[73]	ViT-Base	IN22K+IN1K	84.4
SAMI-B (ours)	ViT-Base	SA1B (11M) + IN1K	**84.8**

Object Detection and Instance Segmentation

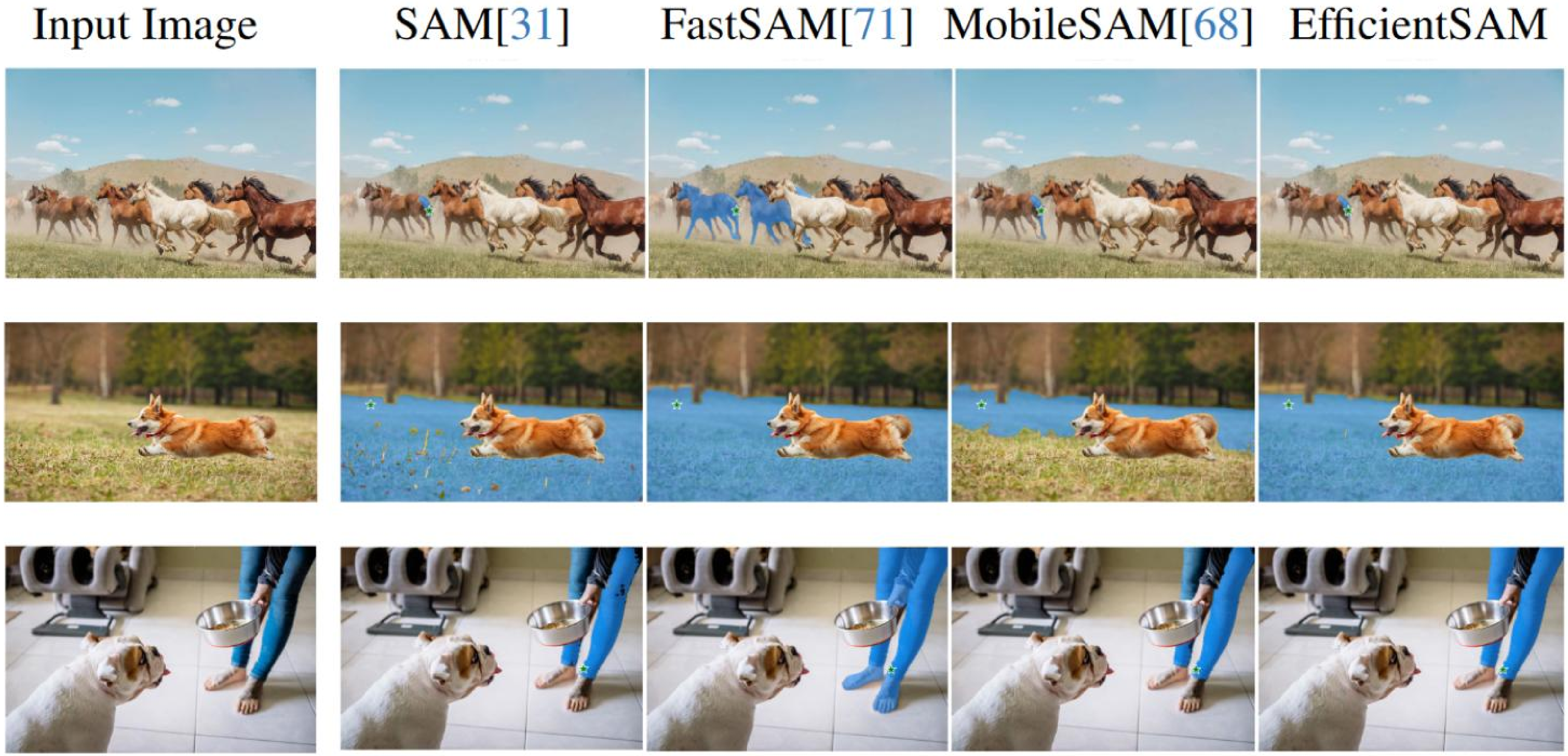
Method	Backbone	AP^{bbox}	AP^{mask}
MAE-Ti[26]	ViT-Tiny	37.9	34.9
SAMI-Ti(ours)	ViT-Tiny	**44.7**	**40.0**
MAE-S[26]	ViT-Small	45.3	40.8
DeiT-S[53]	ViT-Small	47.2	41.9
DINO-S[6]	ViT-Small	49.1	43.3
iBOT-S[73]	ViT-Small	49.7	44.0
SAMI-S (ours)	ViT-Small	**49.8**	**44.2**
MAE-B[26]	ViT-Base	51.6	45.9
SAMI-B (ours)	ViT-Base	**52.5**	**46.5**

Semantic Segmentation

Method	Backbone	mIOU
MAE-Ti[26]	ViT-Tiny	39.0
SAMI-Ti(ours)	ViT-Tiny	**42.7**
MAE-S[26]	ViT-Small	44.1
SAMI-S (ours)	ViT-Small	**48.8**
MAE-B[26]	ViT-Base	49.3
SAMI-B (ours)	ViT-Base	**51.8**

Experiments

Point-Prompt Input

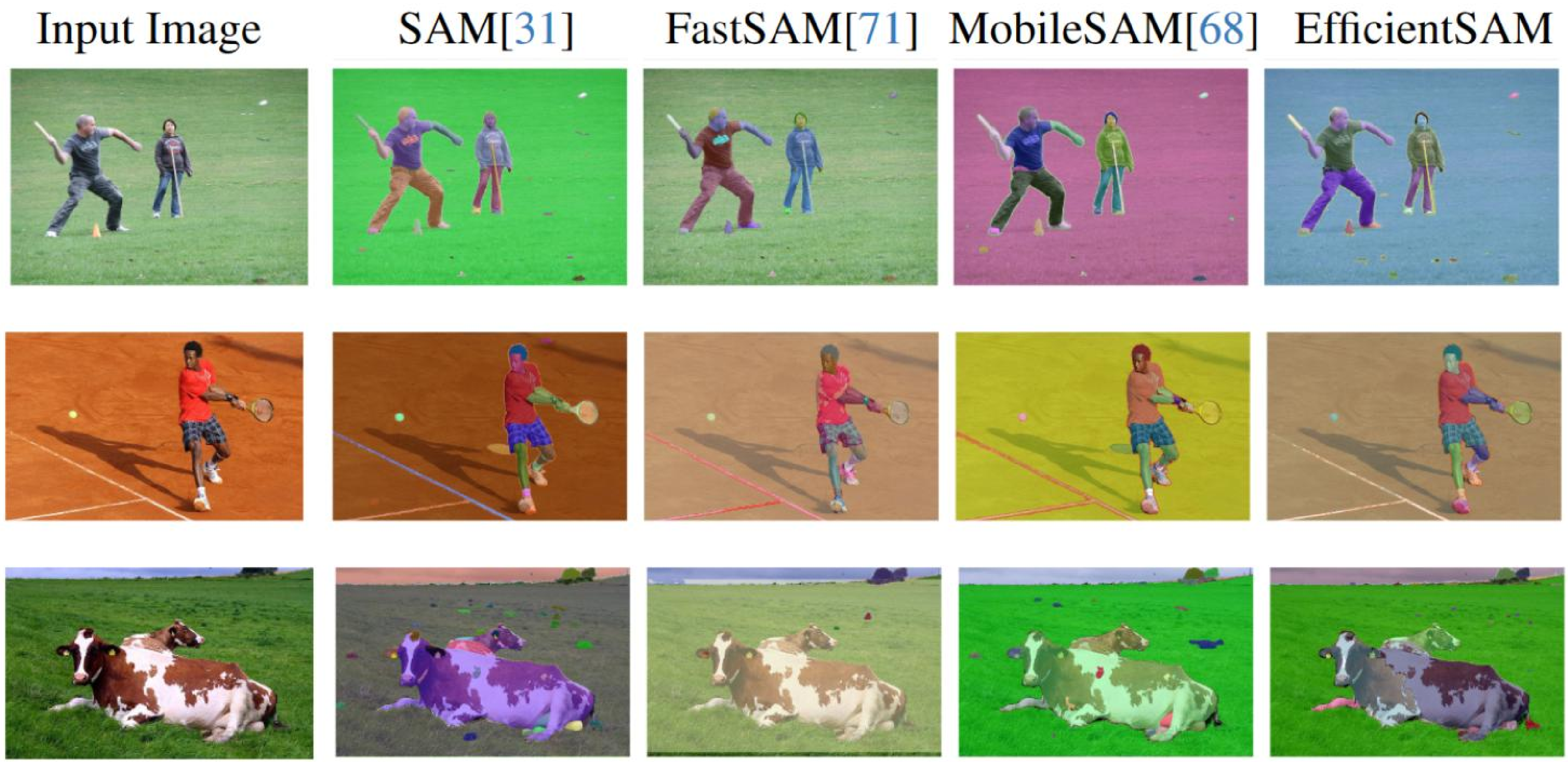


PEKING UNIVERSITY 北京大學

Box-Prompt Input

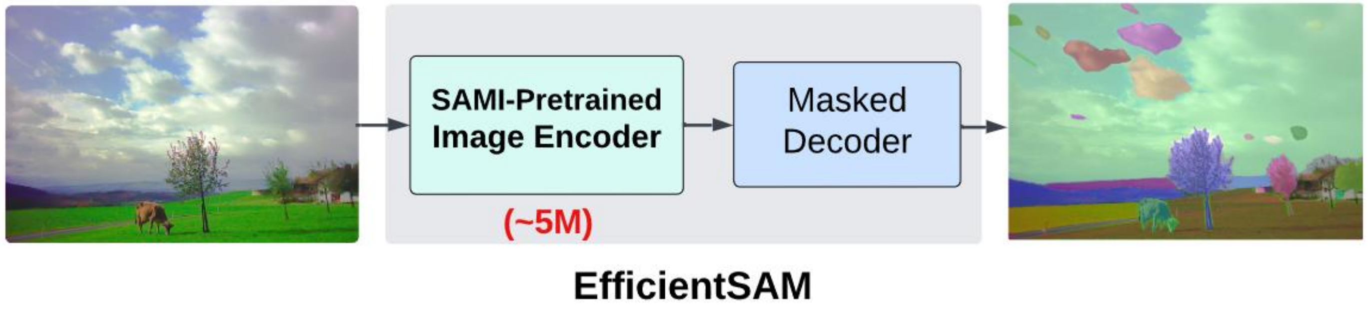
Experiments

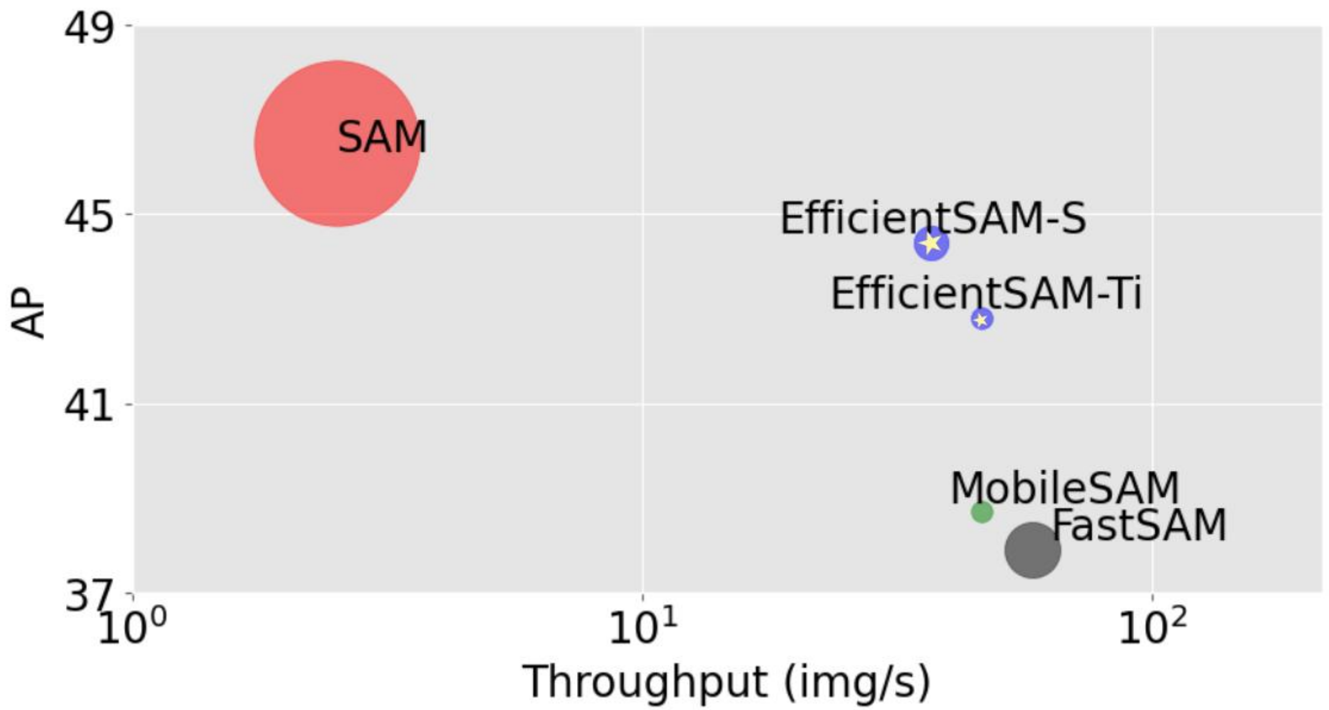
Salient Instance Segmentation



Conclusion

- Proposed a SAM-leveraged masked image pretraining framework *SAMI*
- Delivered **EfficientSAM,** 20x fewer parameters & 20x faster runtime
- “A smaller, faster, and almost as good version of SAM. ”

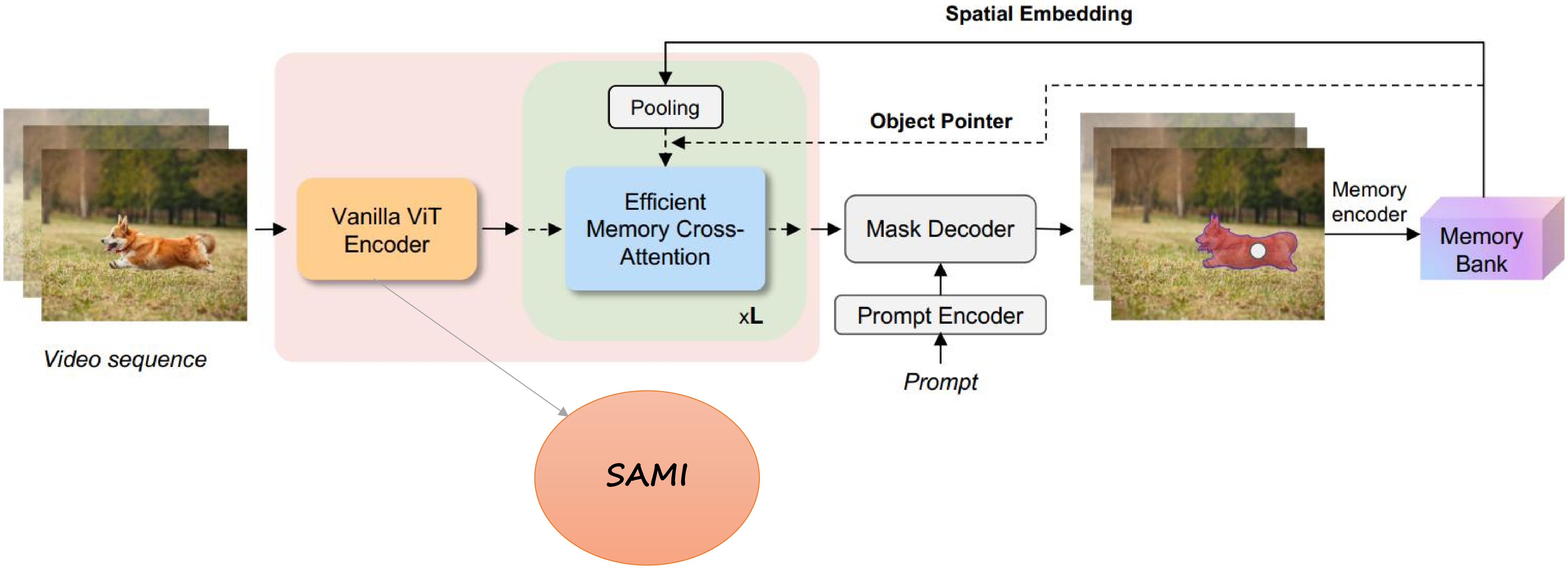




Tested on a single NVIDIA A100 with one box prompt

Future Work

Efficient Track Anything Model (EfficientTAM)



Xiong et al, Efficient Track Anything, arXiv:2411.18933

Thanks for listening!

Presenter: Chenyu Niu
2025.03.02