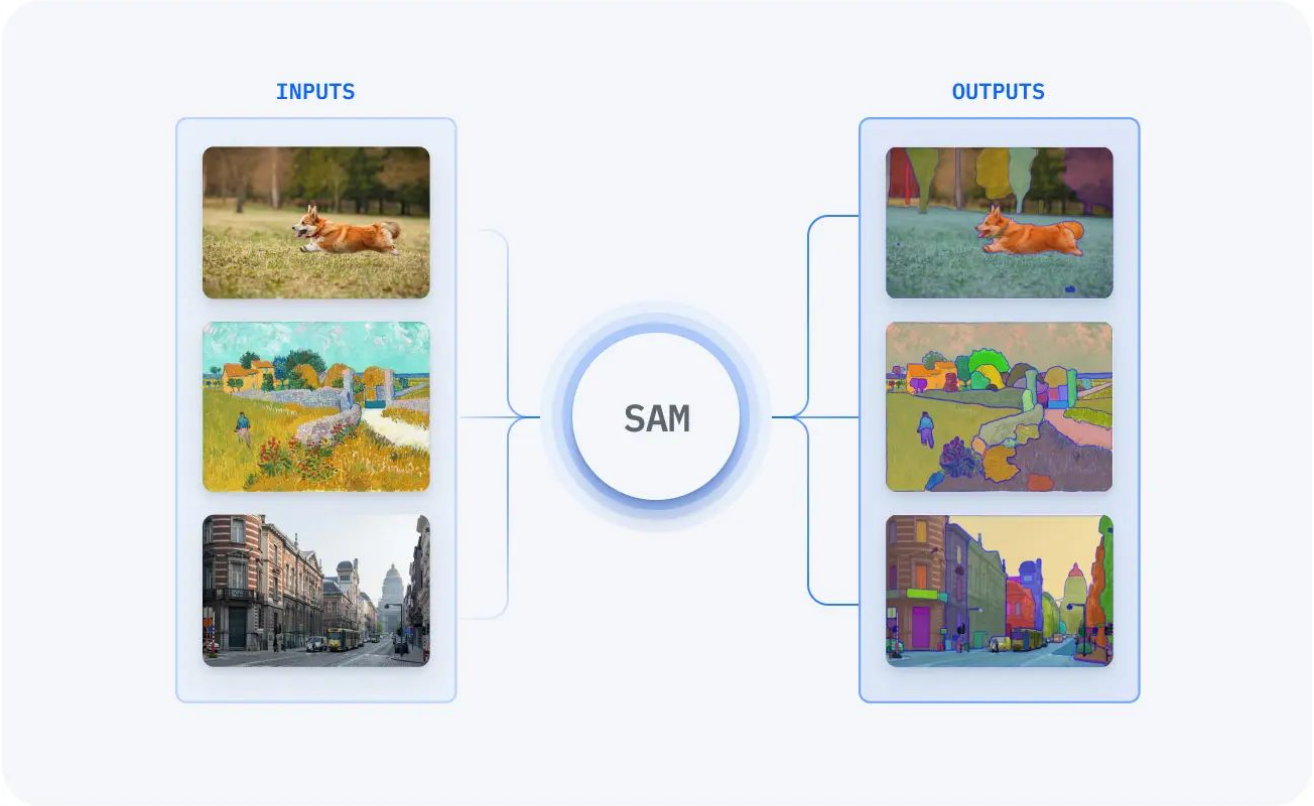
Practices in visual computing 1

Lab10: Image Segmentation 2

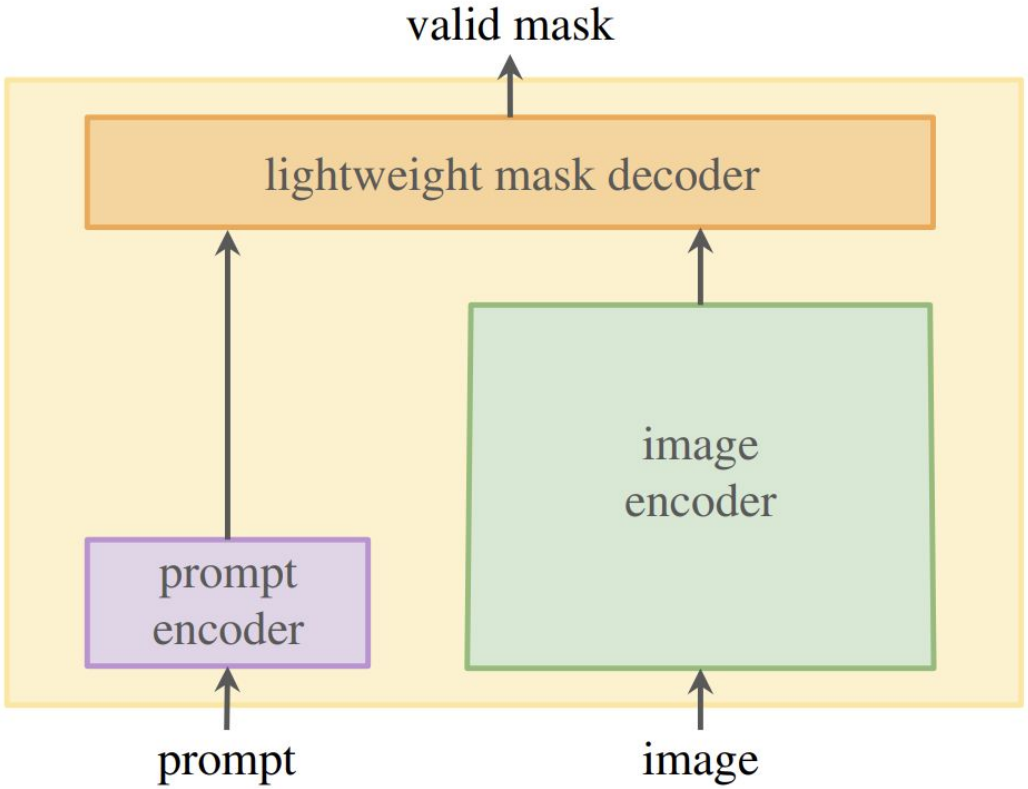
Simon Fraser University
Fall 2025

What is Segment Anything Model (SAM)?

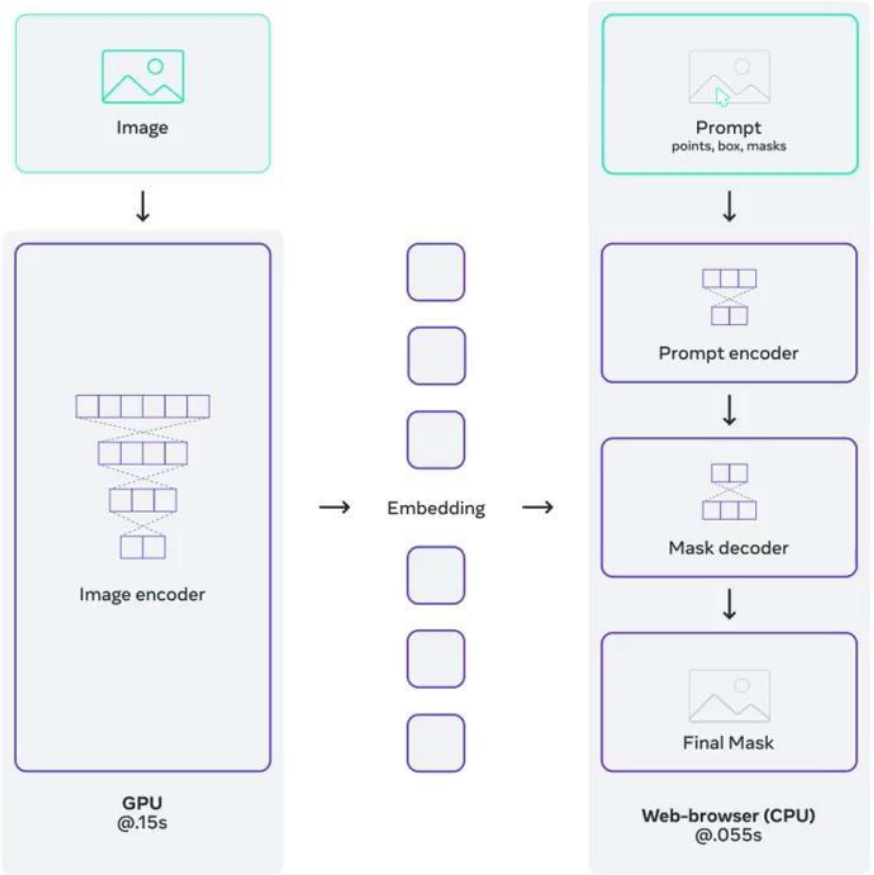


Processing Multiple Prompts

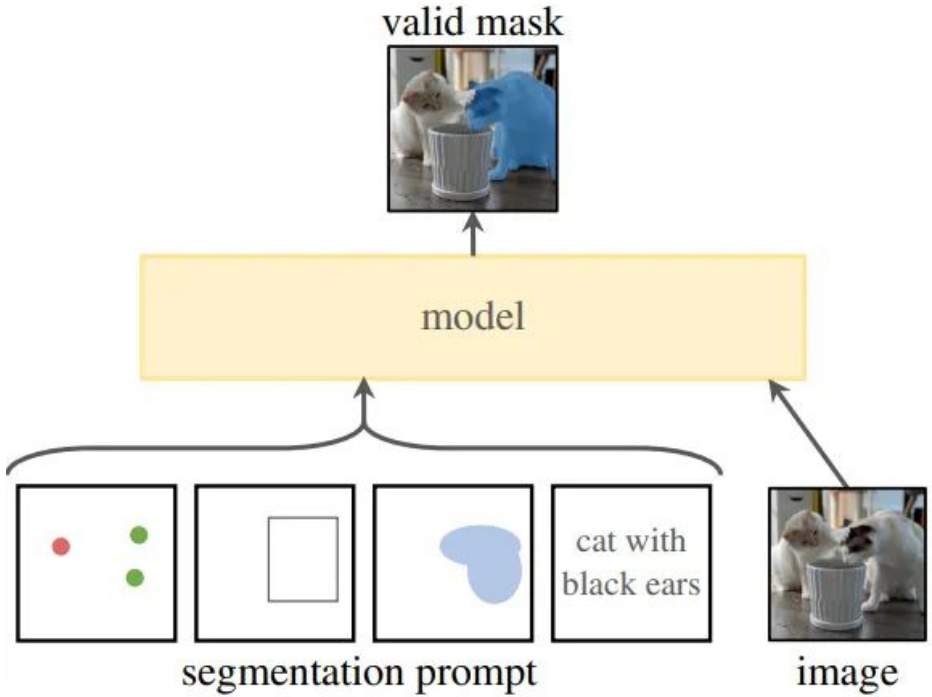
Architecture



Architecture



Prompt Encoder

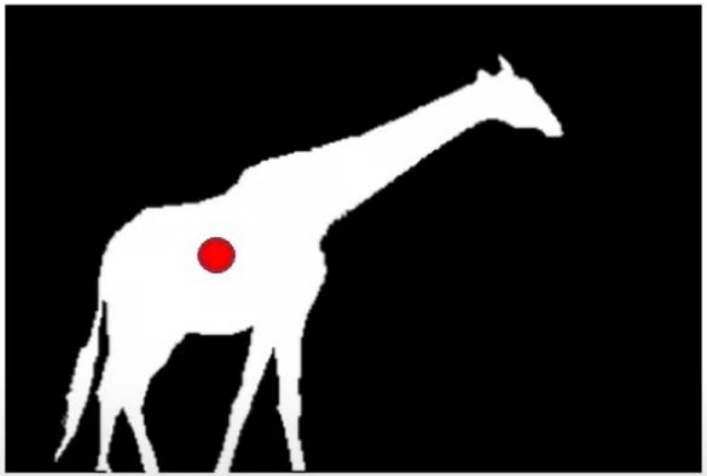


Interactive Training

Training Image

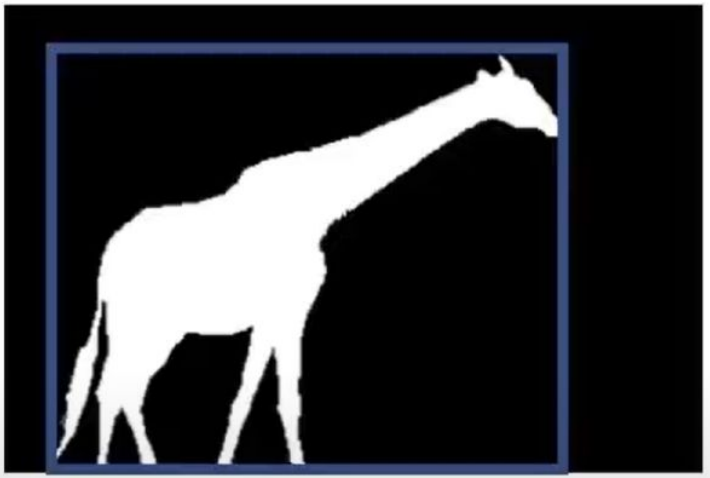
Training Target

Interactive Training - Create Random Prompt



Point Prompt

Pick a random point close to the center of the mask



Bounding Box Prompt

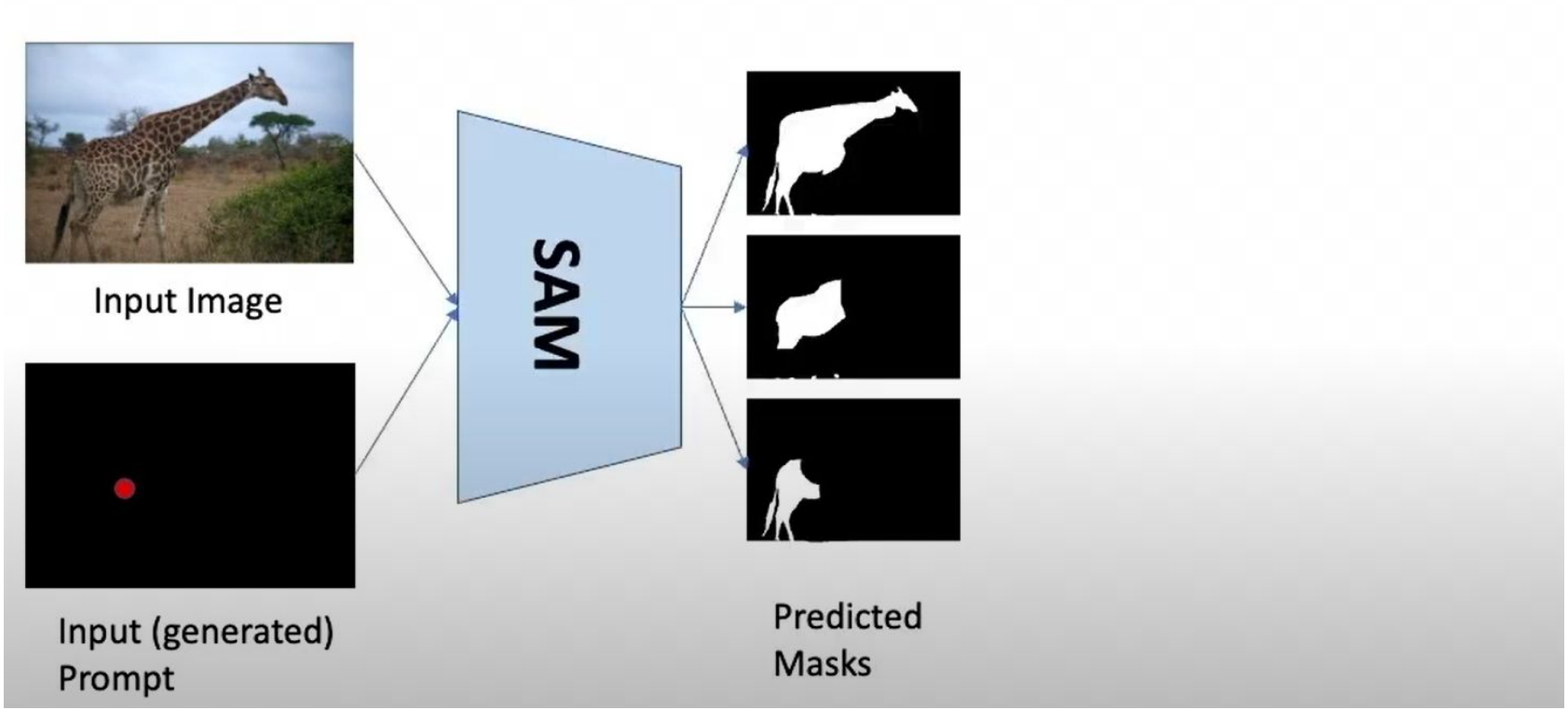
Add jitters to the mask's bounding box

Interactive Training - Predict Mask (Round 1)

Input Image

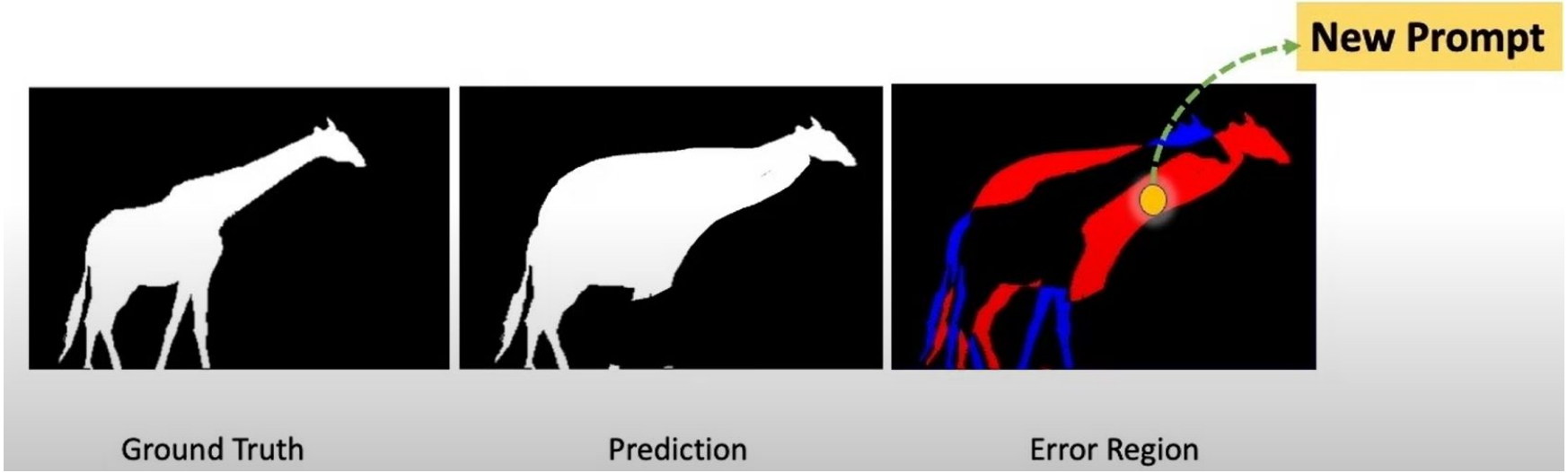
Input (generated)
Prompt

Interactive Training - Predict Mask (Round 1)

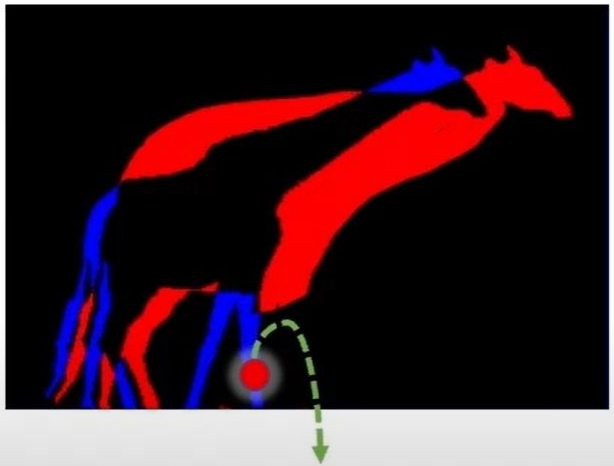


Interactive Training - Predict Mask (Round 1)

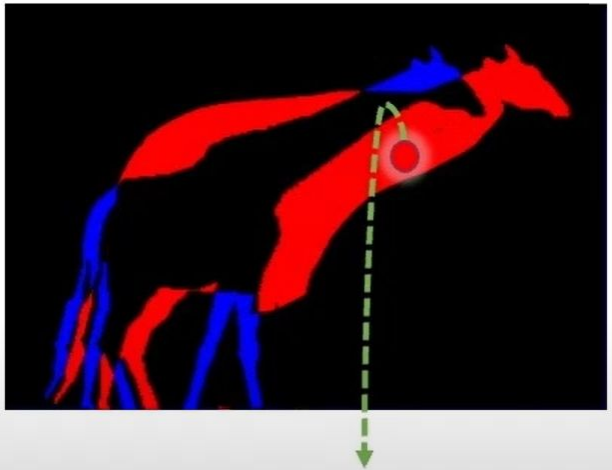
Error region refers to the difference between the target and the predicted mask.



Interactive Training

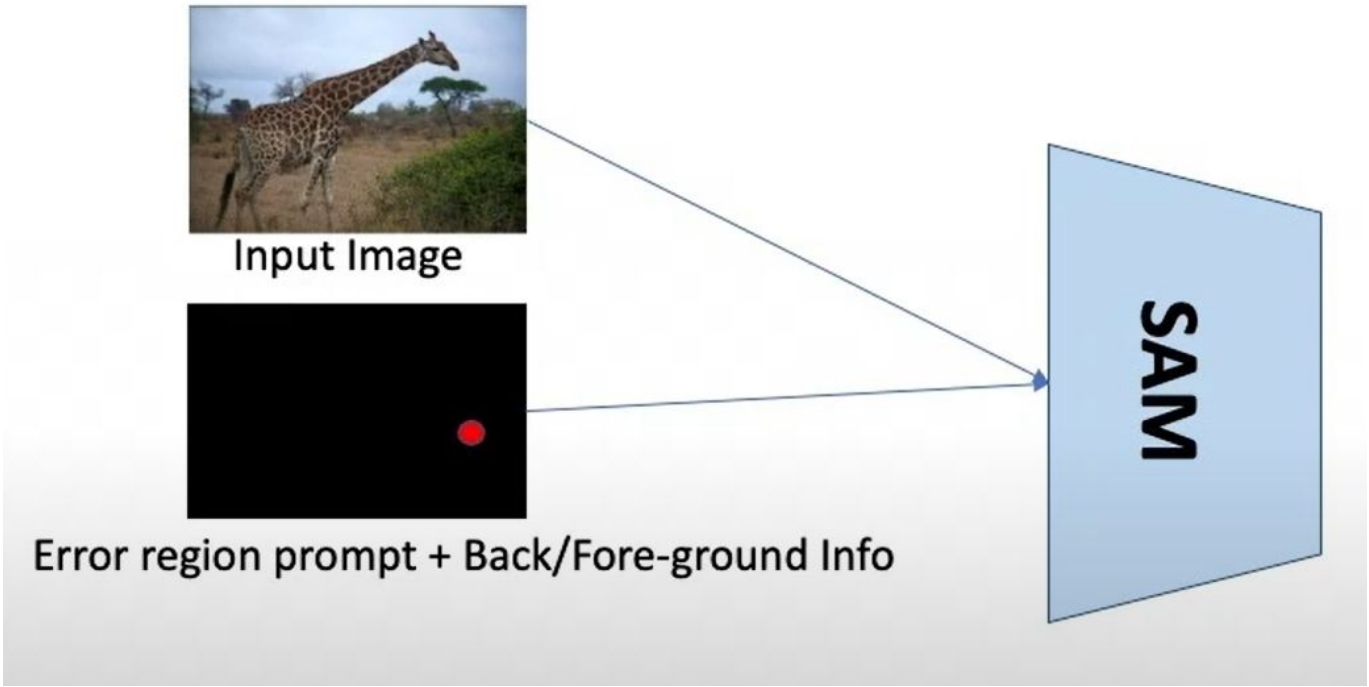


False negative region
Present only in Ground Truth
Foreground point



False Positive region
Present only in Prediction
Background point

Interactive Training - Round 2



Interactive Training - Round 2

Input Image

Error region prompt + Back/Fore-ground Info

Dense mask

Note that the dense mask insterted are unthreshold logits and not a binary mask.

SAM

Architecture

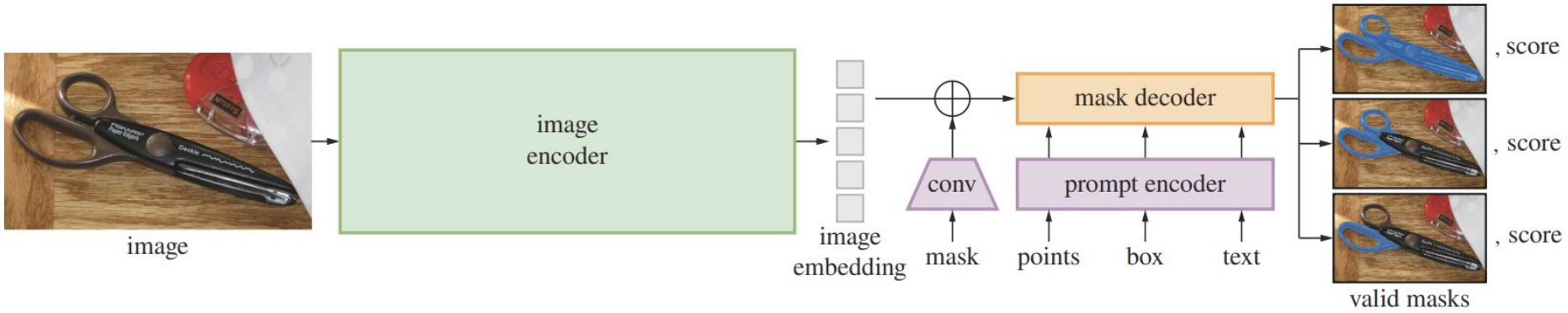
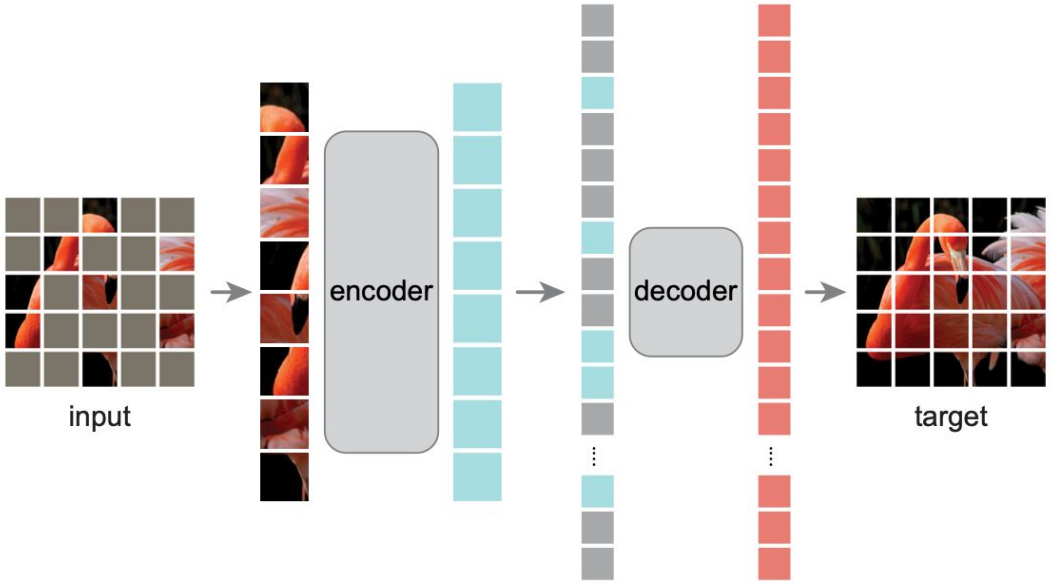
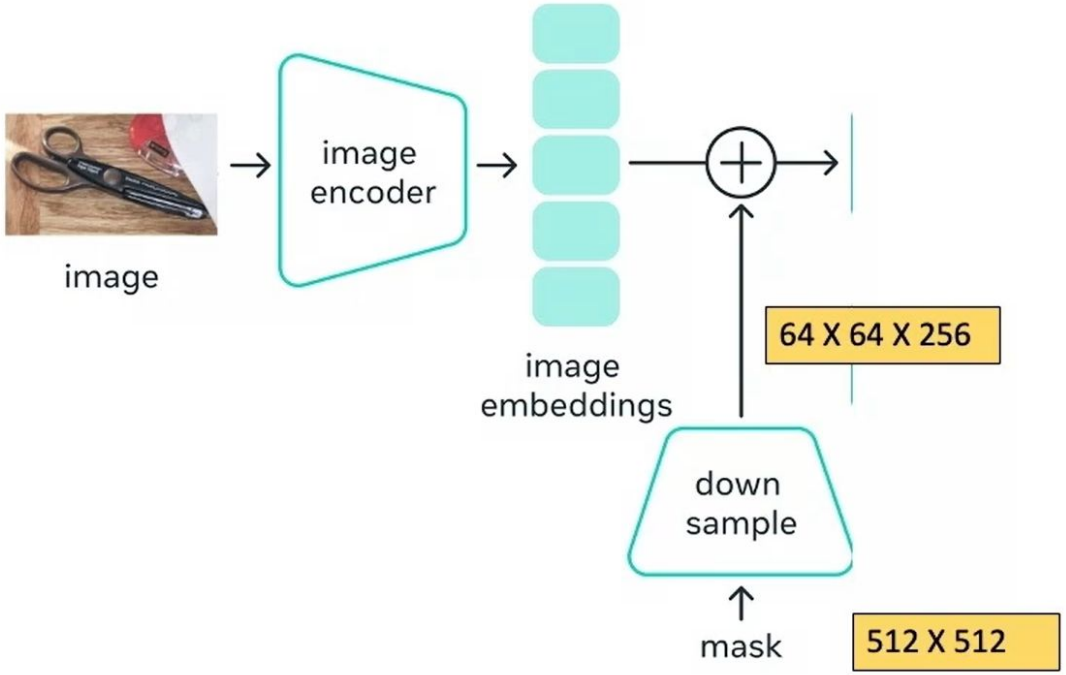


Image Encoder



Masked Autoencoders Are Scalable Vision Learners, He. et al - 2021

Dense Mask Encoder



Sparse Prompt Encoding

Encoding points

- Positional Encodings of point (x, y)
- Trained embedding indicating "foreground" or "background" point

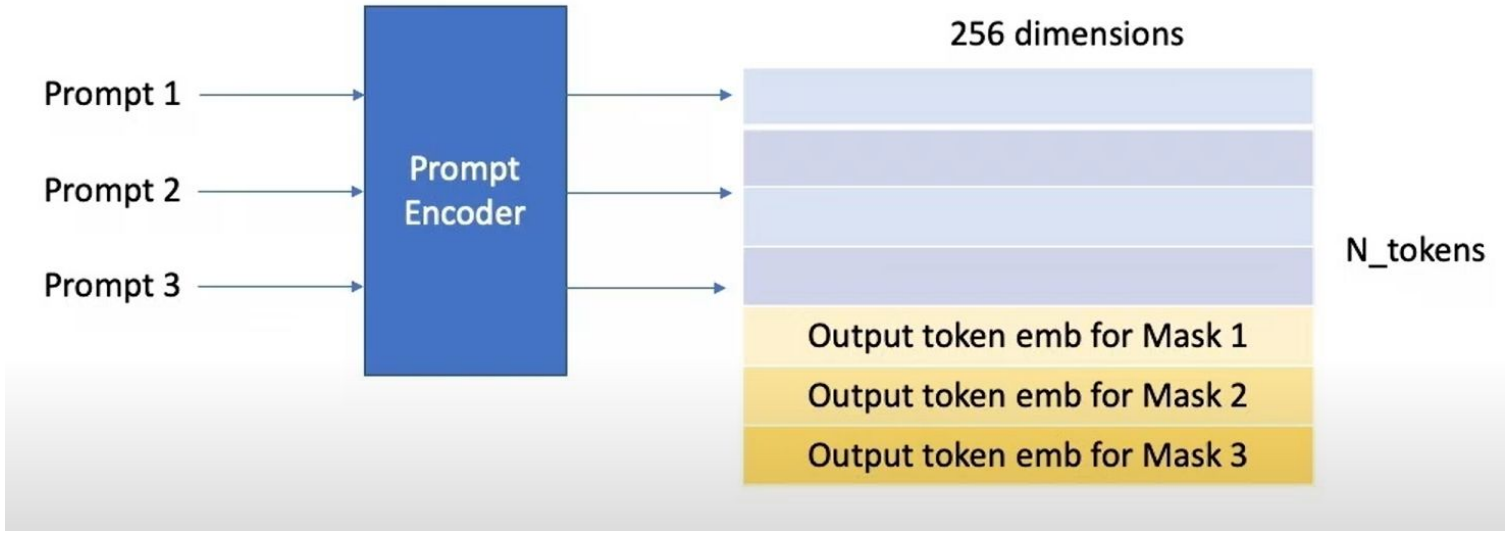
Encoding bounding boxes

- Positional encoding for top left point + embedding for "Top Left"
- Positional encoding for bottom right point + embedding for "Bottom Right"

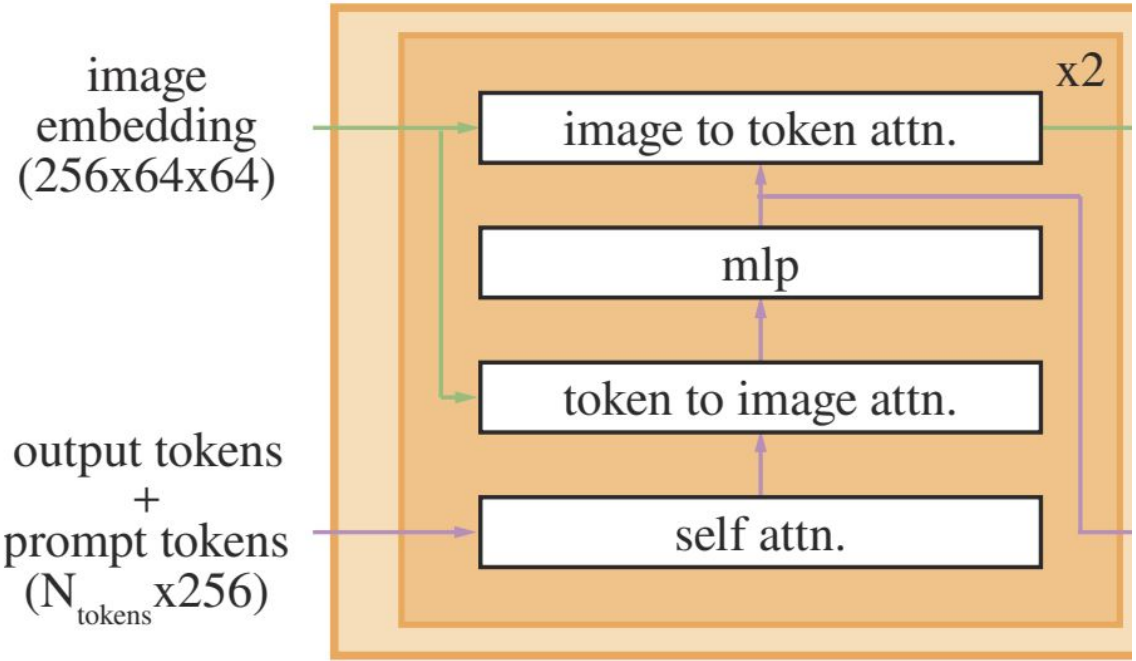
Encoding text prompts

- Pretrained CLIP model text encoder

Prompt Encoder



Mask Decoder



Mask Decoder

Self Attention with the prompt + out tokens

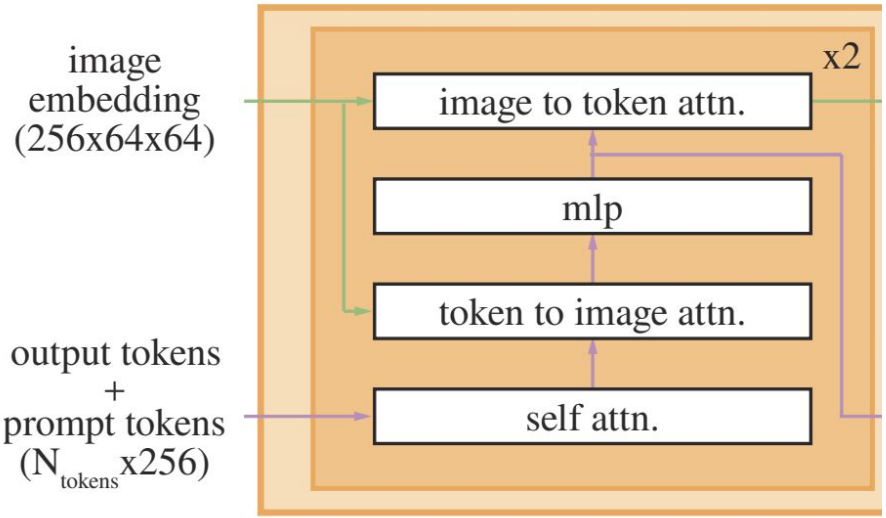
- Updates each prompt/out embedding with contextual knowledge about other prompt/out embeddings

Prompt -> Image attention

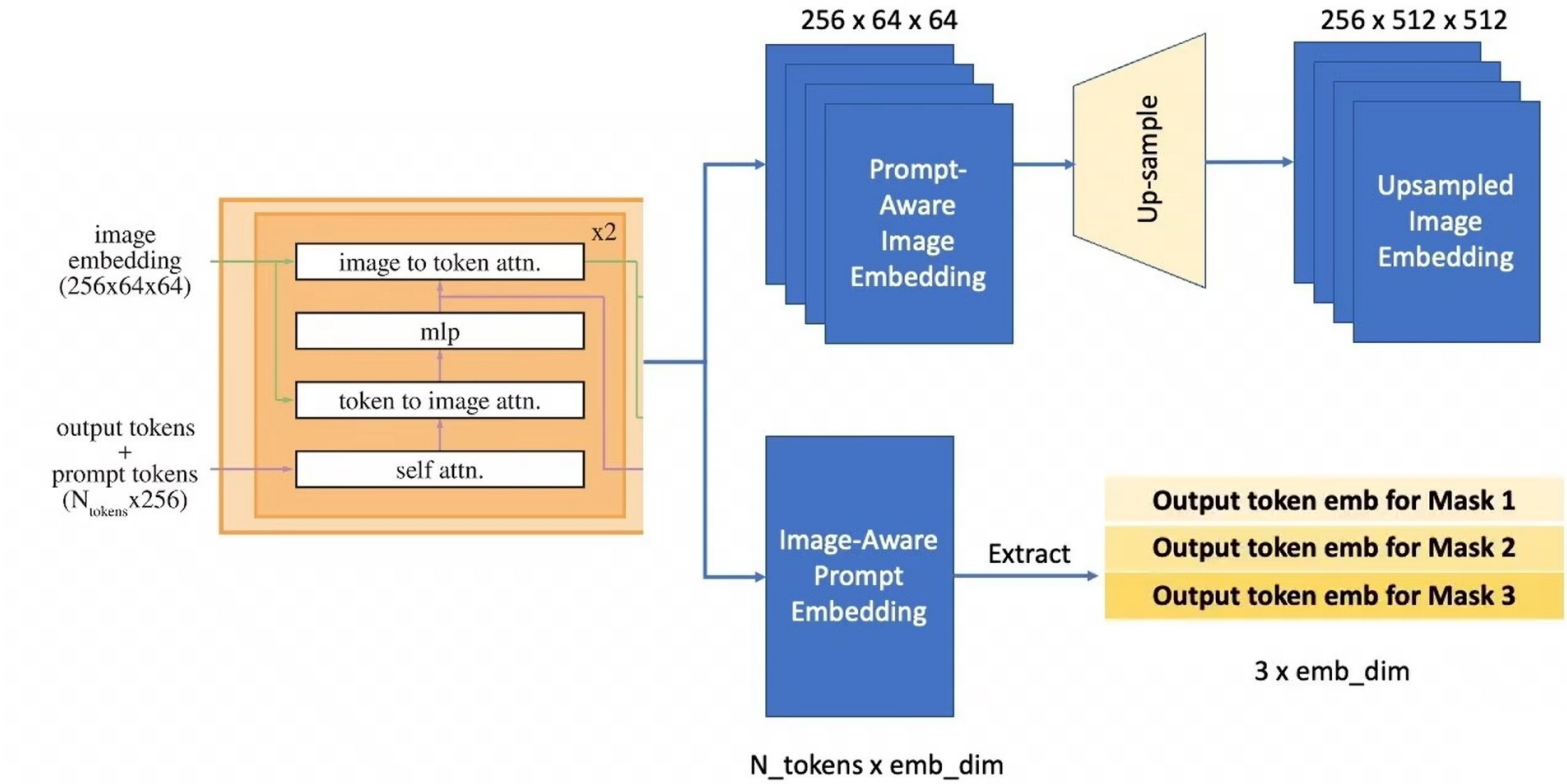
- Updates prompt/out token embeddings with contextual information from the image

Image -> Prompt attention

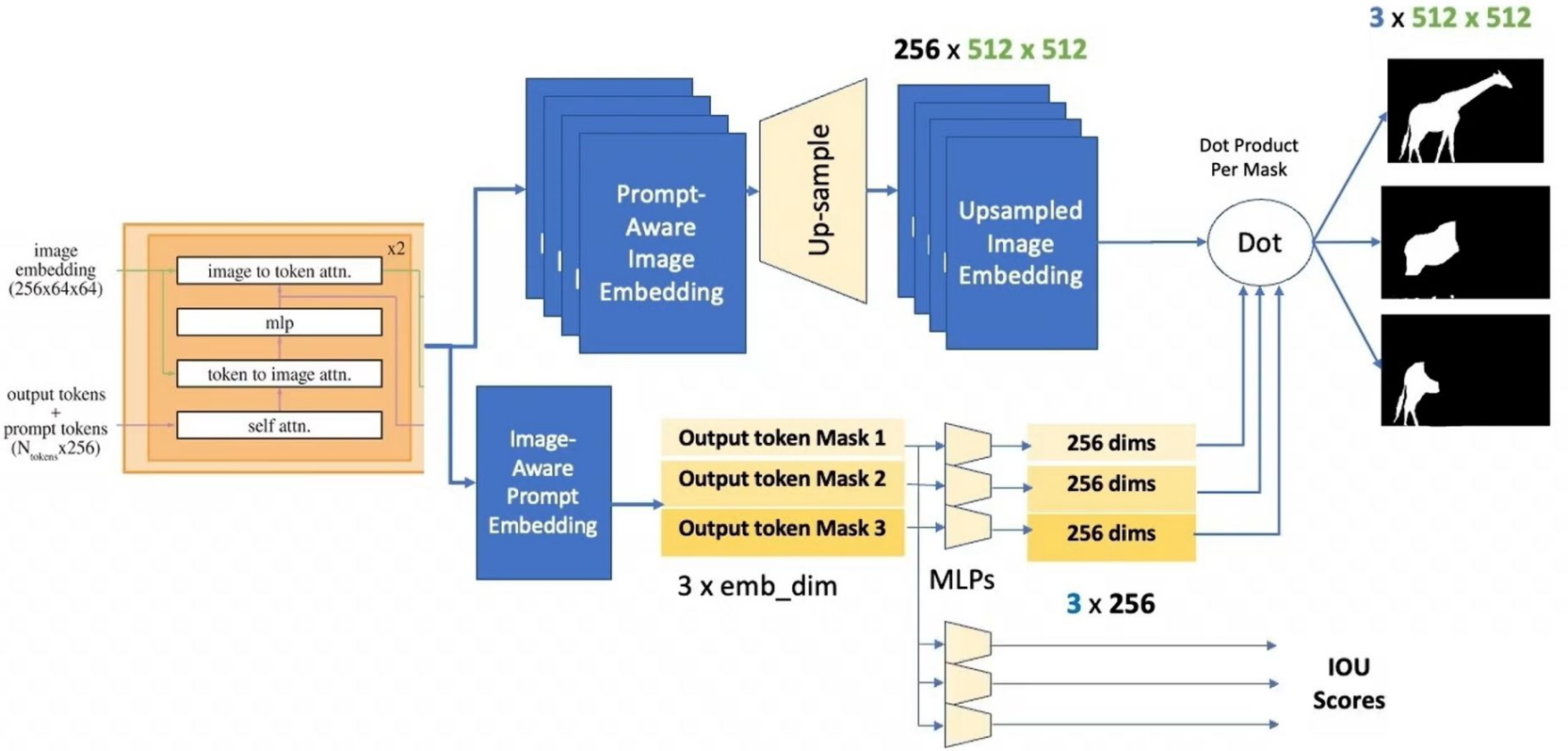
- Updates the image embeddings with contextual information from the prompt/out tokens.



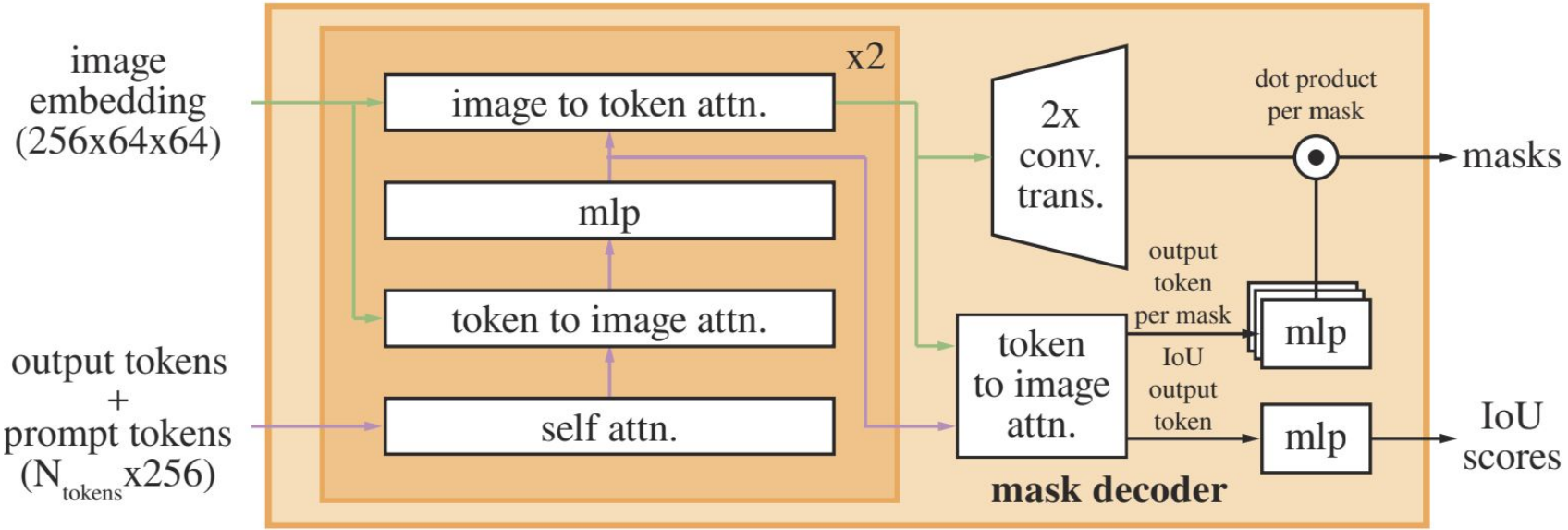
Mask Decoder



Mask Decoder



Mask Decoder



Data Collection

Assisted-manual stage:

- SAM is pretrained on publicly available datasets
- Annotators label prominent segments in the images
- Annotated 120K images with 4.3M masks

Semi-automatic stage:

- SAM is trained on collected data so far
- Annotators label additional segments SAM missed
- Annotated 180K images with 5.9M masks

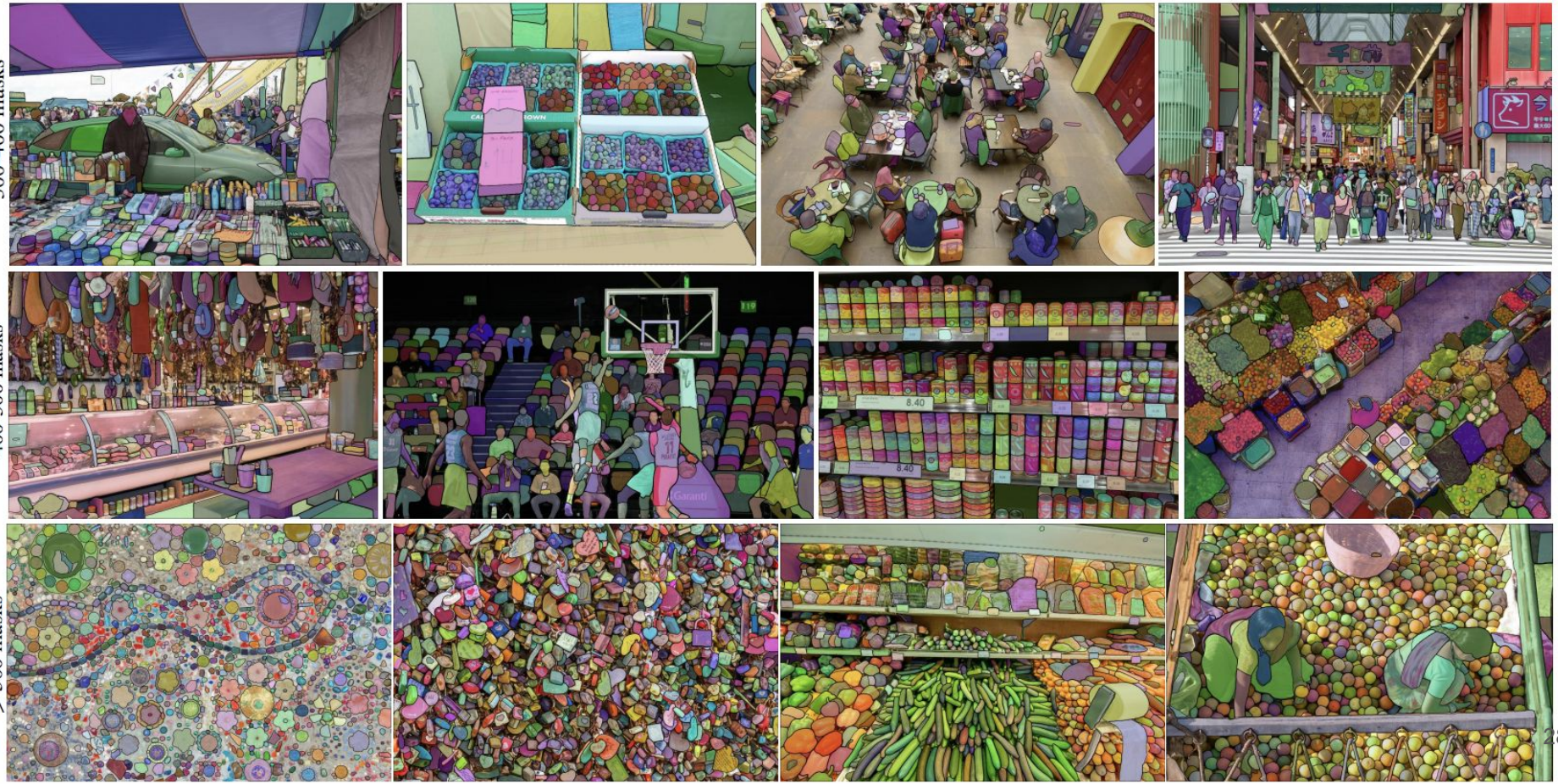
Fully-automatic stage:

- SAM is trained on collected data so far (300K images with 10.2M masks)
- Annotates 11M images with 1B masks autonomously

SA-1B Dataset

SA-1B Dataset

SA-1B Dataset

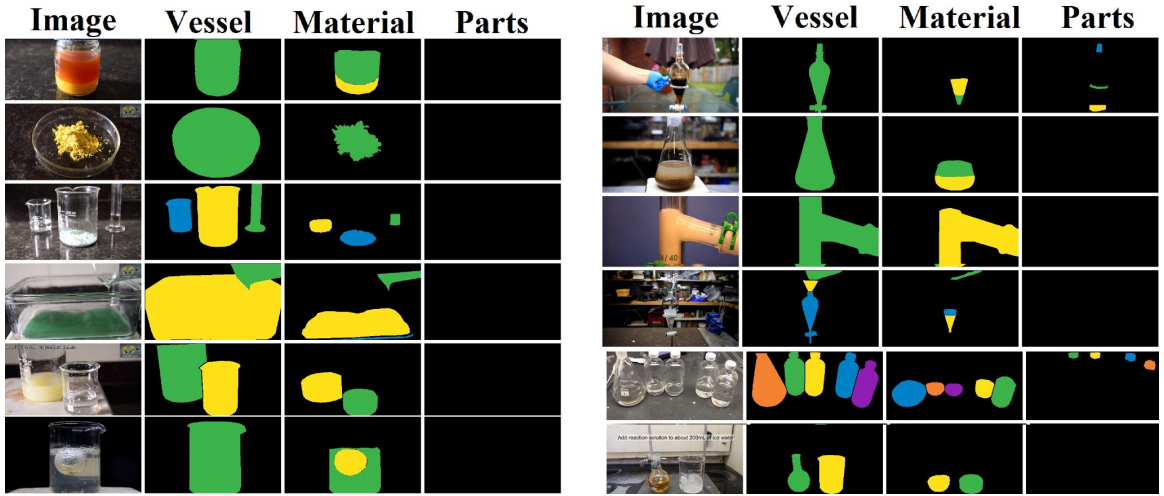


LabPicsV1 Dataset

A specialized dataset with annotated images of lab equipment, reagents, containers, and other lab objects.

Applications:

- Object detection and segmentation in laboratory settings
- Automated inventory management and safety monitoring



Reference

https://www.v7labs.com/blog/segment-anything-model-sam

https://viso.ai/deep-learning/segment-anything-model-sam-explained/

https://www.youtube.com/watch?v=OhxJkqD1vuE&ab_channel=NeuralBreakdownwithAVB

Segment Anything, Kirillov et al - 2023