Segment Anything Model and its applications

Never Stand Still | Faculty of Engineering | School of Computer Science and Engineering

Lian Xu
Department of Computer Science and Software Engineering, UWA, Perth, Australia
School of Computer Science and Engineering, UNSW, Sydney, Australia
lian.xu@uwa.edu.au

Foundation models

Foundation model refers to any model that is trained on broad data and can be adapted to a wide range of downstream tasks [1].

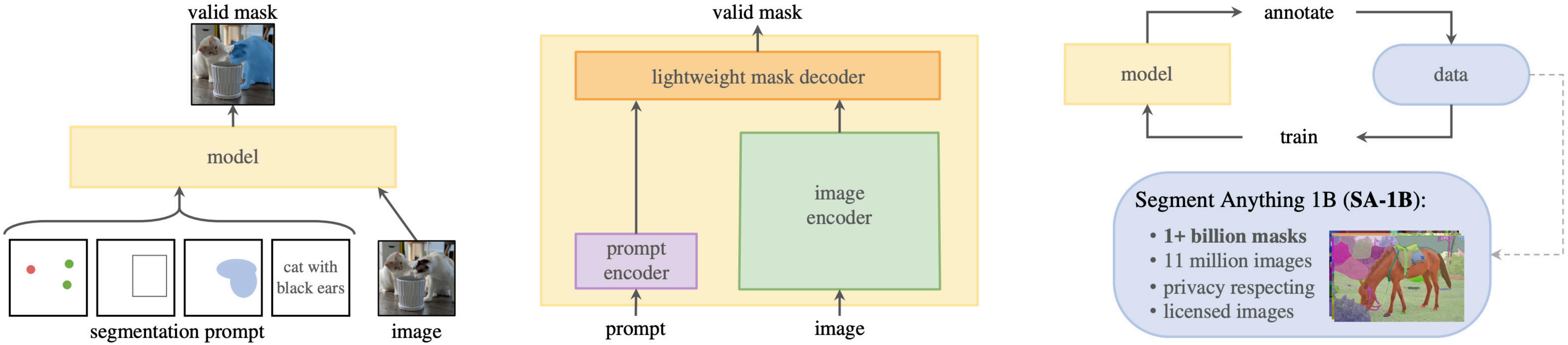
Foundation models in NLP are very popular (e.g., GPT)...with strong zero-shot and few-shot generalization.

- Pre-trained on web-scale datasets
- Solving diverse tasks via prompt engineering

[1] Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).

Foundation model for segmentation

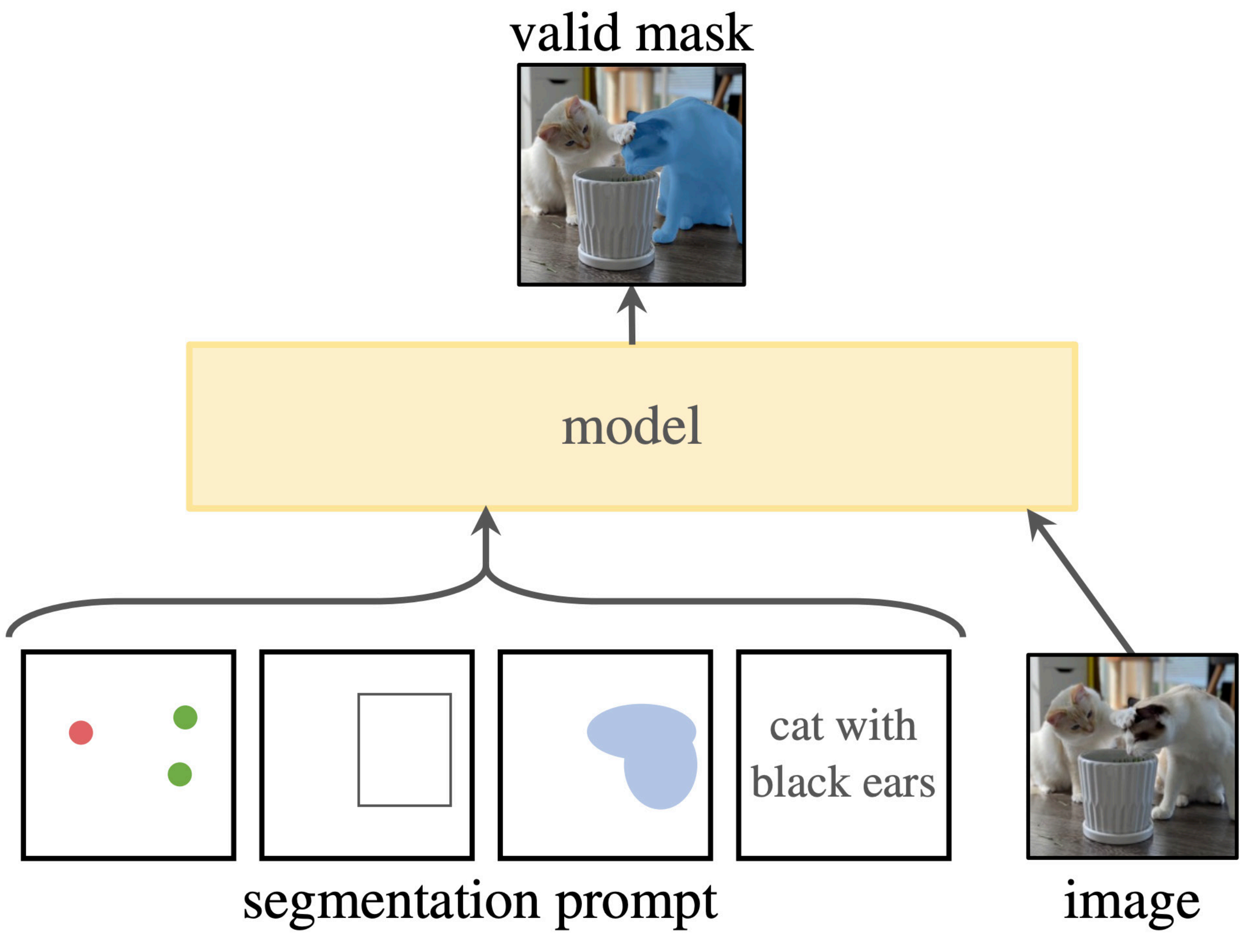
Three keys to the success: Task; Model; Data.



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

Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Promptable segmentation



A prompt can be:

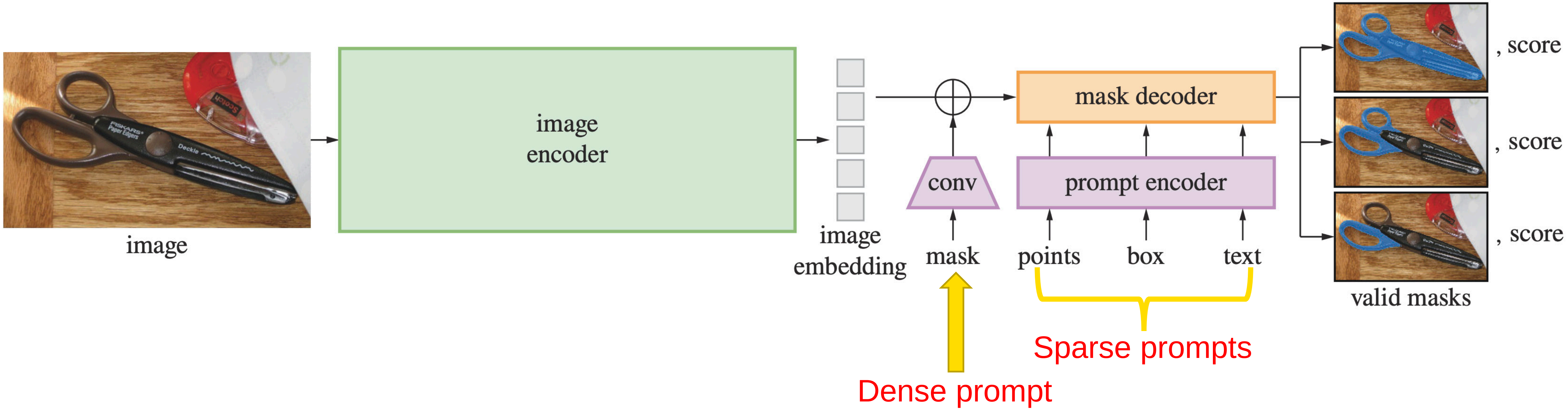
1. a set of foreground/background points
2. A rough box or mask
3. Free-form text

…. any information to indicate what to segment

This task aims to return a *valid* segmentation mask given *any prompt*.

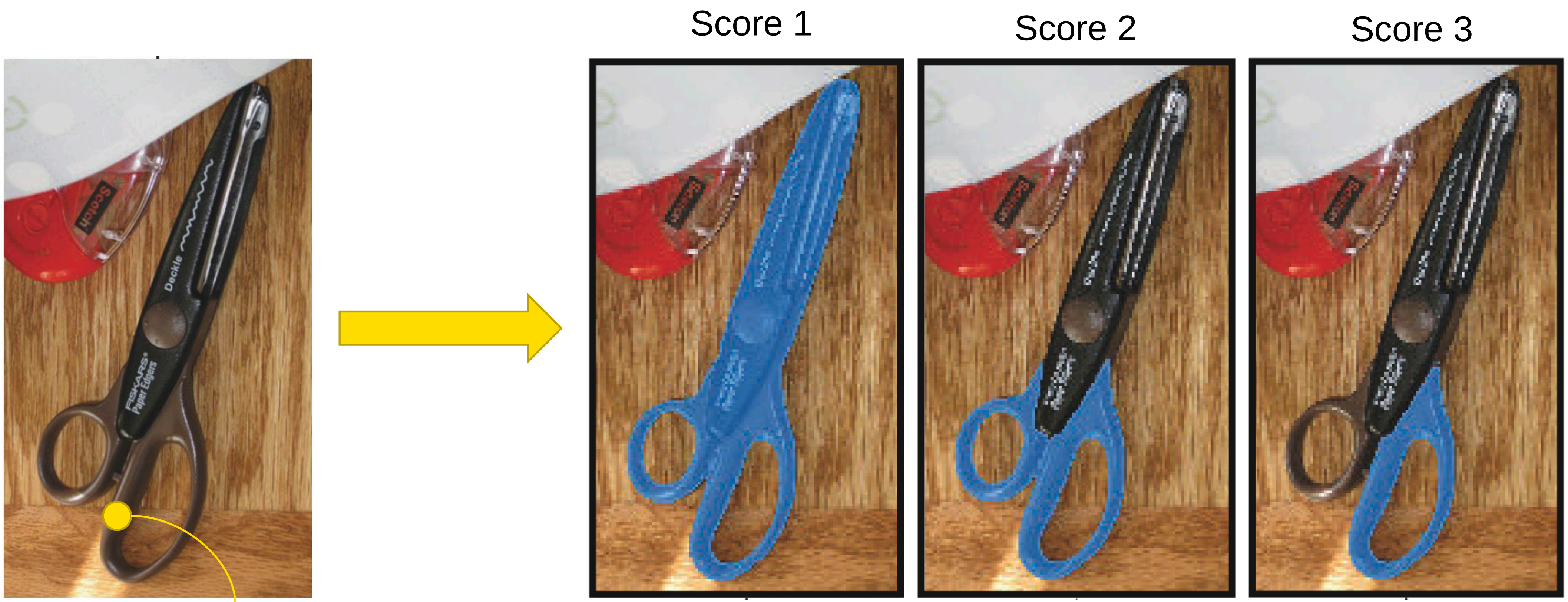
Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

The Overview of SAM



Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Resolving Ambiguity



“Click” (a point prompt)

Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Efficient & Flexible Model Design

Image encoder:

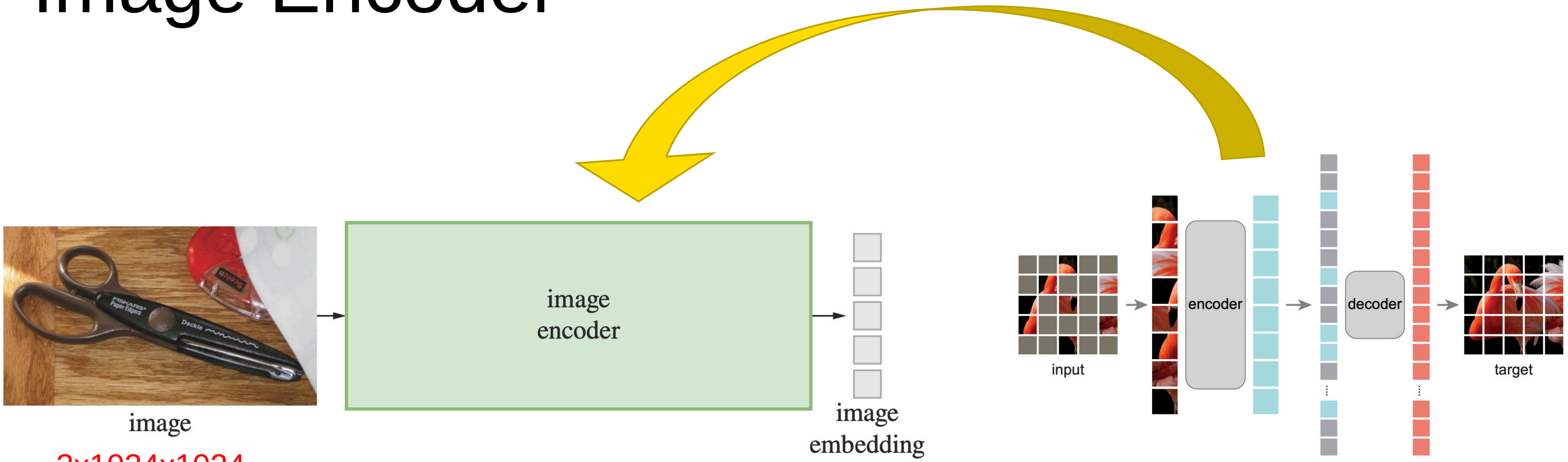
- Runs once per image
- A large ViT model
- Runs on a GPU

Prompt encoder & Mask decoder:

- Runs on each input prompt
- A Lightweight model
- Runs on a web-browser

Video from https://segment-anything.com/

Image Encoder



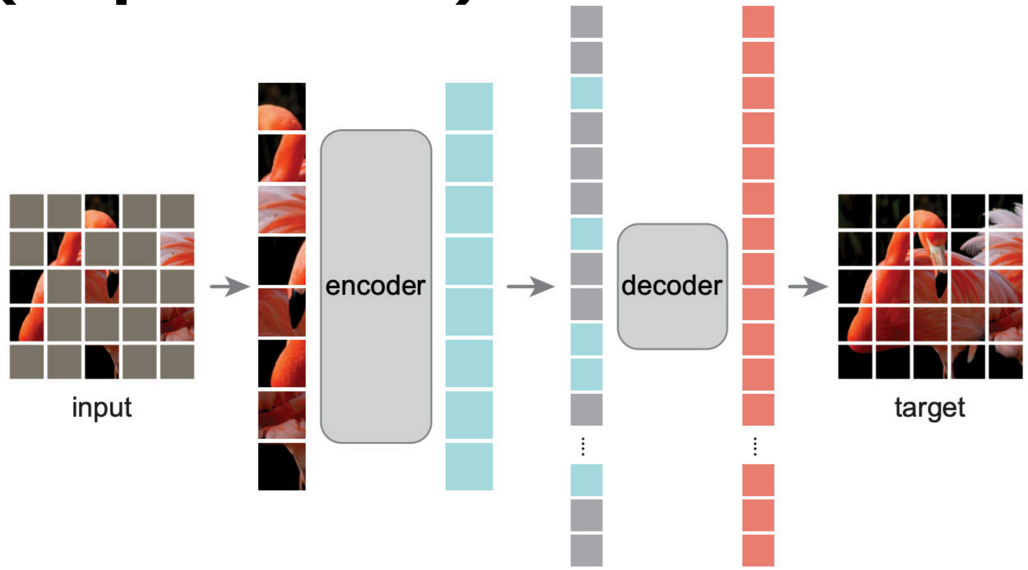
3x1024x1024

256x64x64

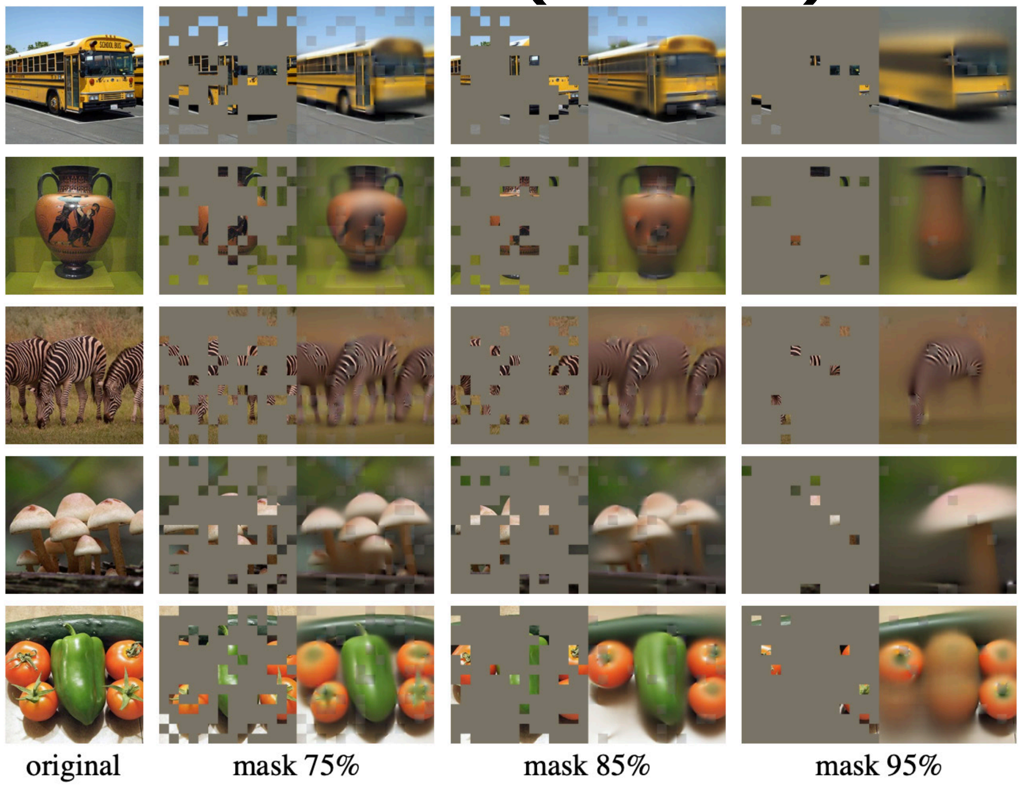
Masked Auto-encoder

Kirillov, Alexander, et al. "Segment anything." ICCV 2023.
He, Kaiming, et al. "Masked autoencoders are scalable vision learners." CVPR 2022.

(Optional) Masked Auto-Encoder (MAE)



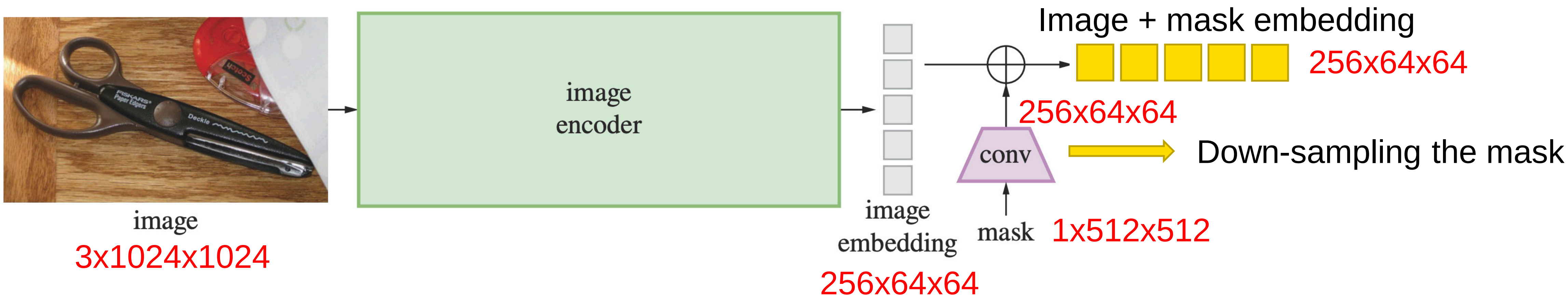
MAE architecture. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.



Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." CVPR 2022.

Dense (Mask) Prompt Encoding



Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Sparse Prompt Encoding

- **Encoding points**
  - A positional encoding (PE) of the point's location (x, y)
  - A learned embedding indicating ”foreground/background” point
- **Encoding bounding boxes**
  - The PE for the “top-left” point + a learned embedding indicating “top-left”
  - The PE for the “bottom-right” point + a learned embedding indicating “bottom-right”
- **Encoding text prompts**
  - Text embeddings form the pre-trained CLIP text encoder

(Optional) CLIP

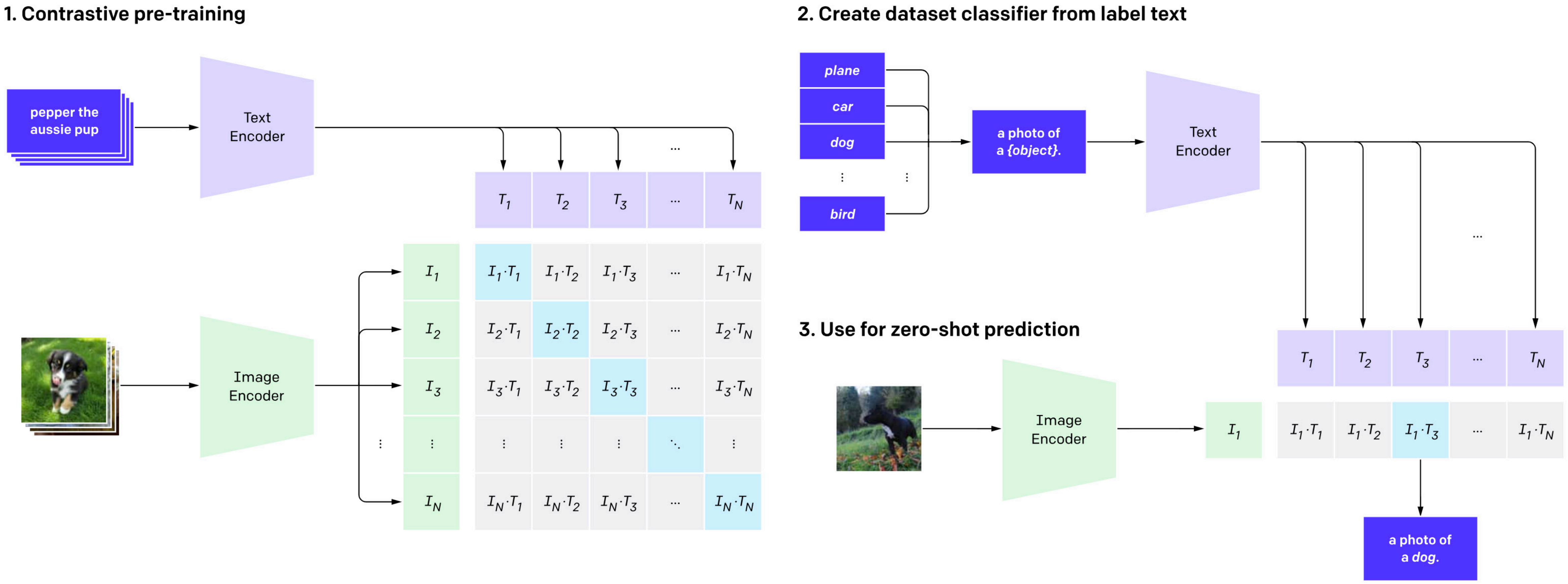
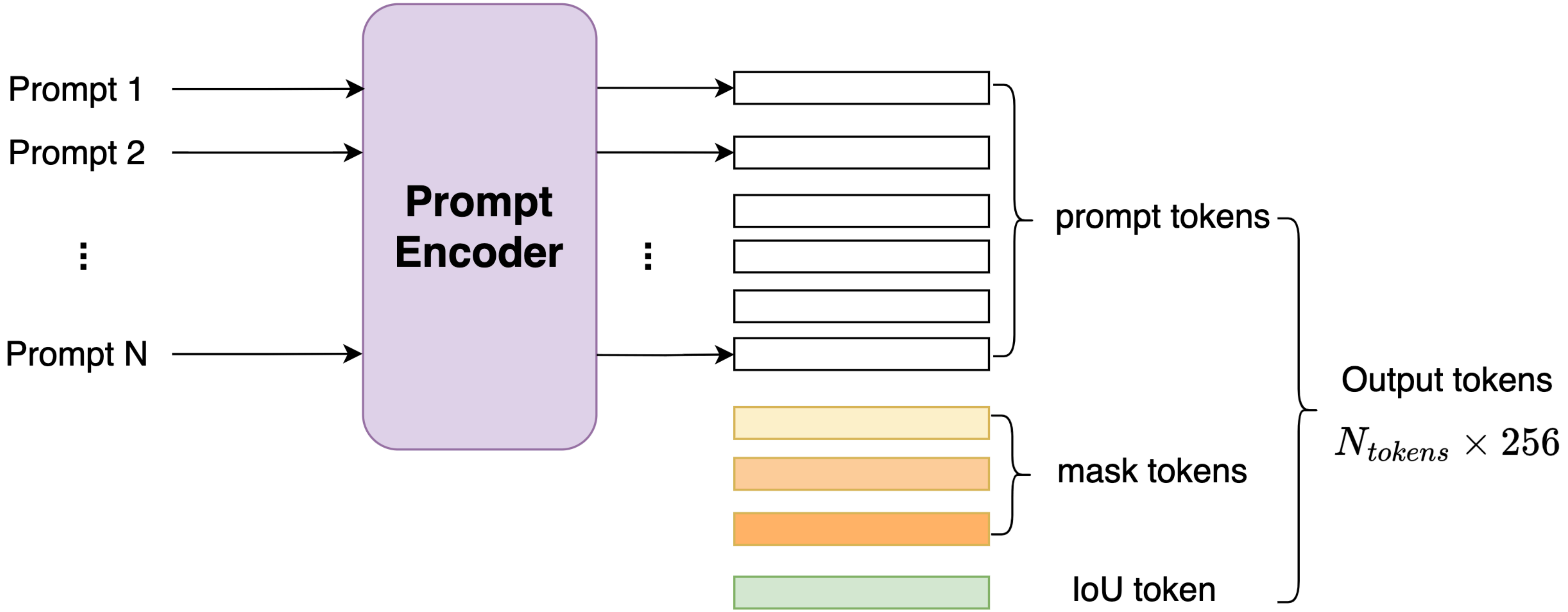
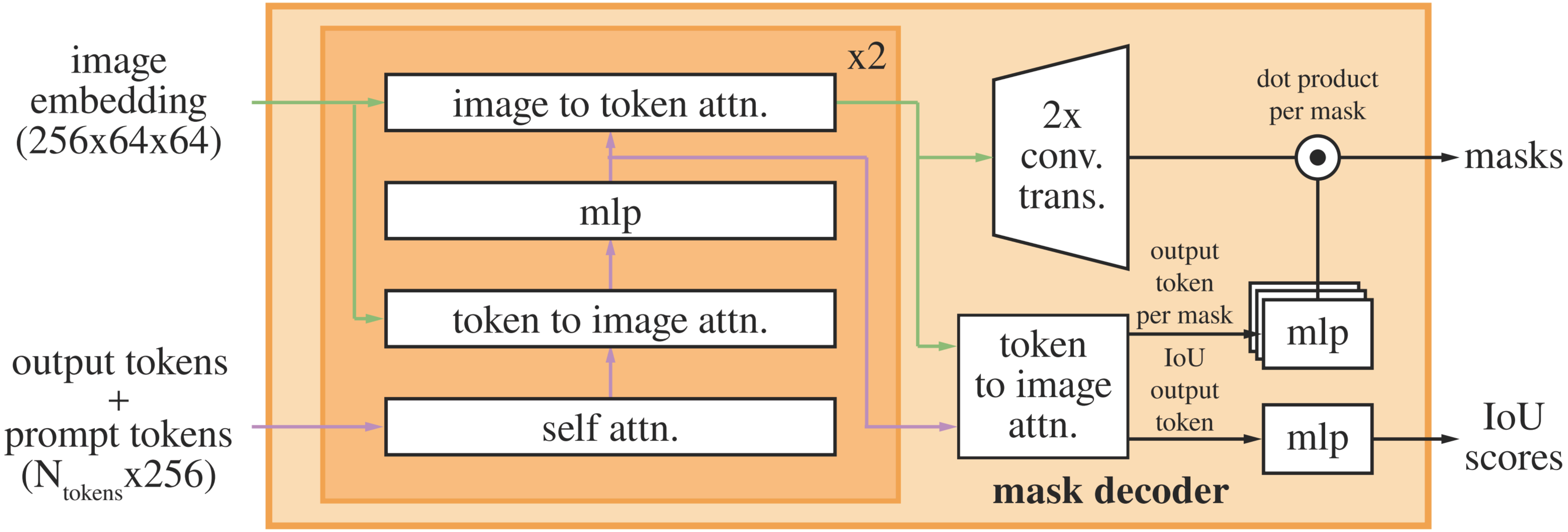


Image from https://openai.com/research/clip

Total prompt encoding



Mask decoder



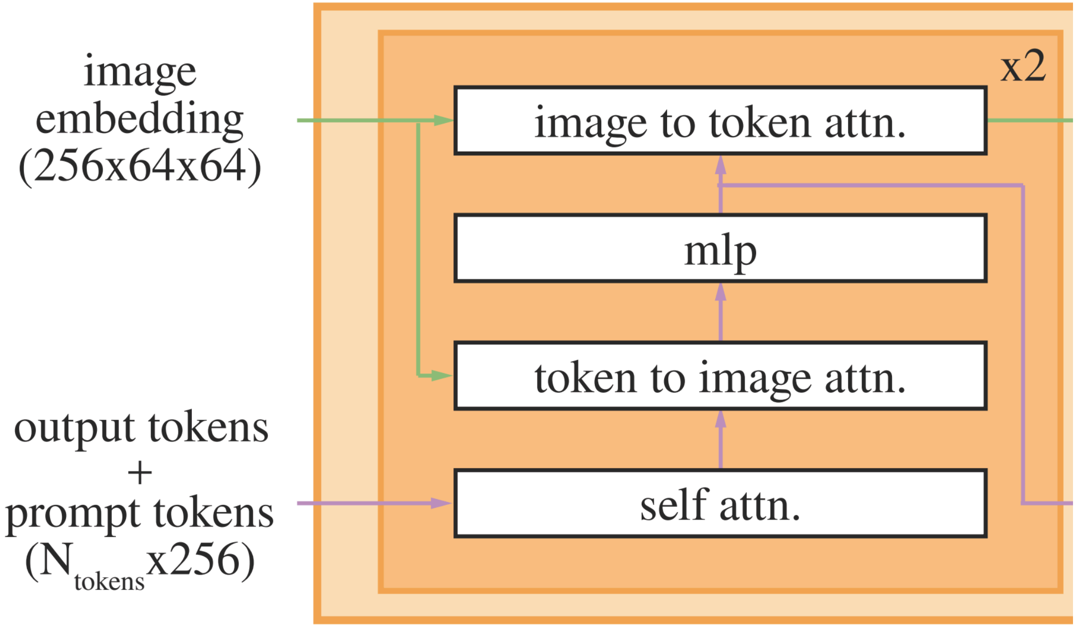
Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Mask decoder – 3 types of attention

- Self-attention of the tokens
  Update each prompt/out embedding with contextual knowledge of other tokens

- Cross-attention: tokens → image embedding
  Update the tokens with image context

- Cross-attention: image embedding → tokens
  Update the image embedding with prompt information



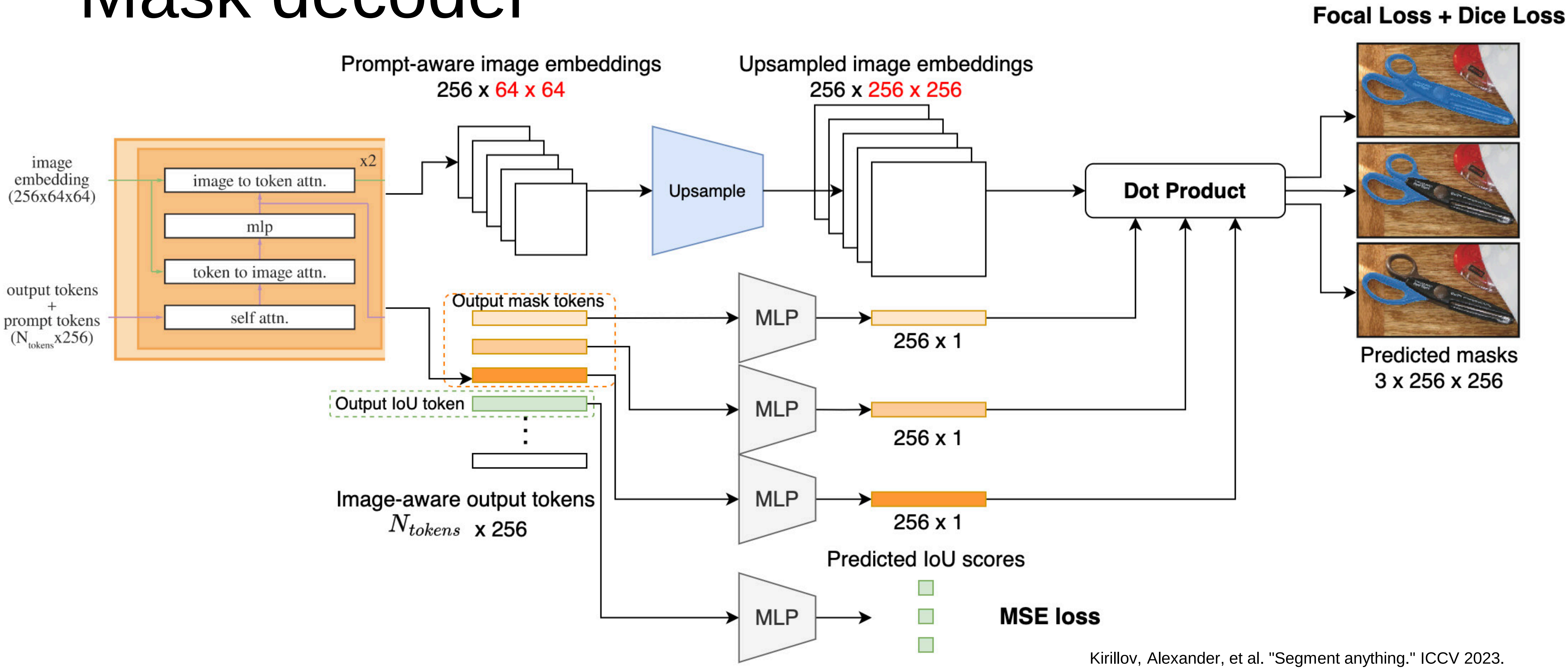
Output tokens: 1. mask tokens; 2. IoU token

$N_{tokens} = N_{output\_mask} + 1 + N_{prompts}$

Here, $N_{output\_mask} = 3$ (whole, part, sub-part)

Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Mask decoder



Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Training losses – Focal Loss

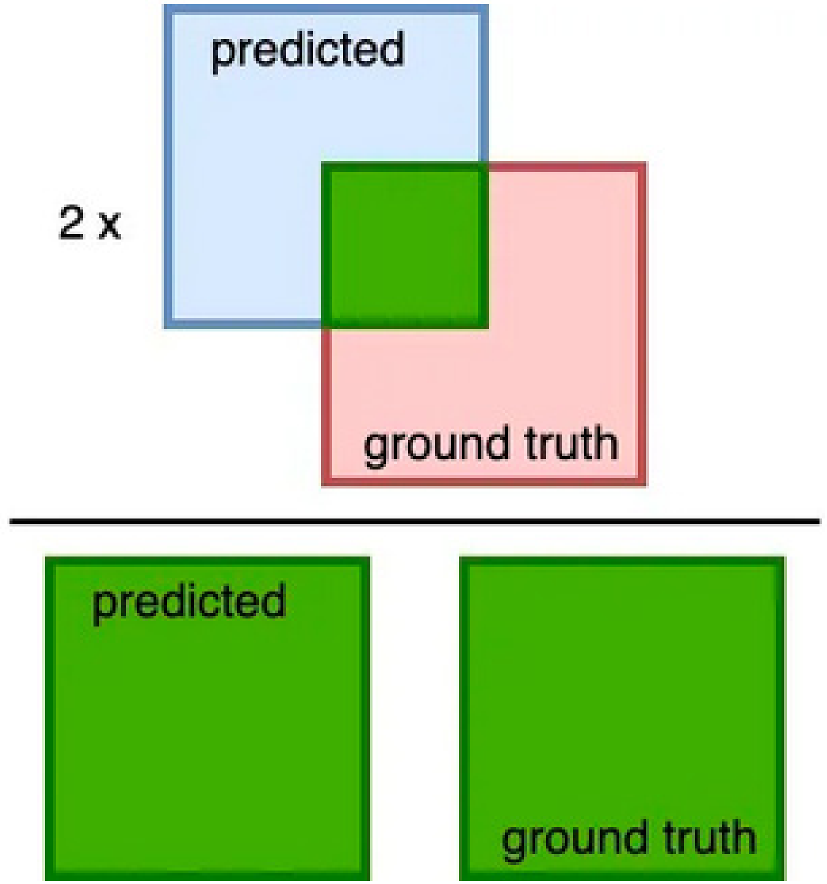
Focal Loss modifies the Cross-Entropy Loss by focusing learning on hard mis-classified examples and down-weighting easy samples.

$$\text{Focal Loss} = -(1 - p_t)^{\gamma} \log(p_t)$$

Lin, Tsung-Yi, et al. "Focal loss for dense object detection." ICCV 2017.

Training losses – Dice Loss

$$\text{Dice Loss} = \frac{2 \times area\ of\ overlapped\ (green)}{total\ area\ (green)} =$$



Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." 3DV 2016.

Training Algorithm

An iterative and interactive segmentation setup:

3 stages with different prompts in a total of 11 iterations:

- First iteration: randomly select a point or a box as prompt
- 2-9 iterations: the predicted mask with the highest predicted IoU and a point sampled from the error prediction of that mask
- 10-11 iterations: the predicted mask with the highest predicted IoU

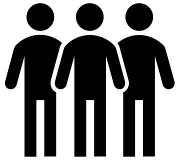
SAM Training with Data Engine

Stage 1 – Assisted-Manual

SAM

Retrain with newly annotated labels

Manually revise



Repeating for 6 times

- Train an initial segmentation model on publicly available datasets
- Annotators revise the predicted masks
- Use the newly annotated labels to train the model

resulting in 120K annotated images with 4.3M masks, ~44 masks per image

Images from https://segment-anything.com/

SAM Training with Data Engine

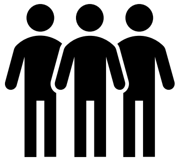
Stage 2 – Semi-Automatic: To improve the diversity of masks

SAM

Repeating for 5 times

- Train SAM on the collected data
- Annotators label additional segments SAM missed

Label additional segments

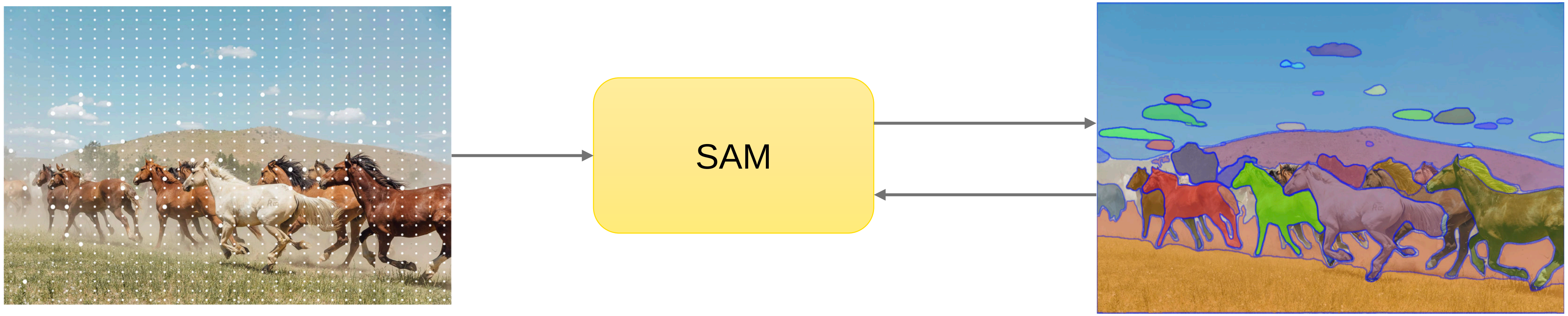


resulting in 180K annotated images with 5.9M masks, ~72 masks per image

Images from https://segment-anything.com/

SAM Training with Data Engine

Stage 3 – Fully automatic



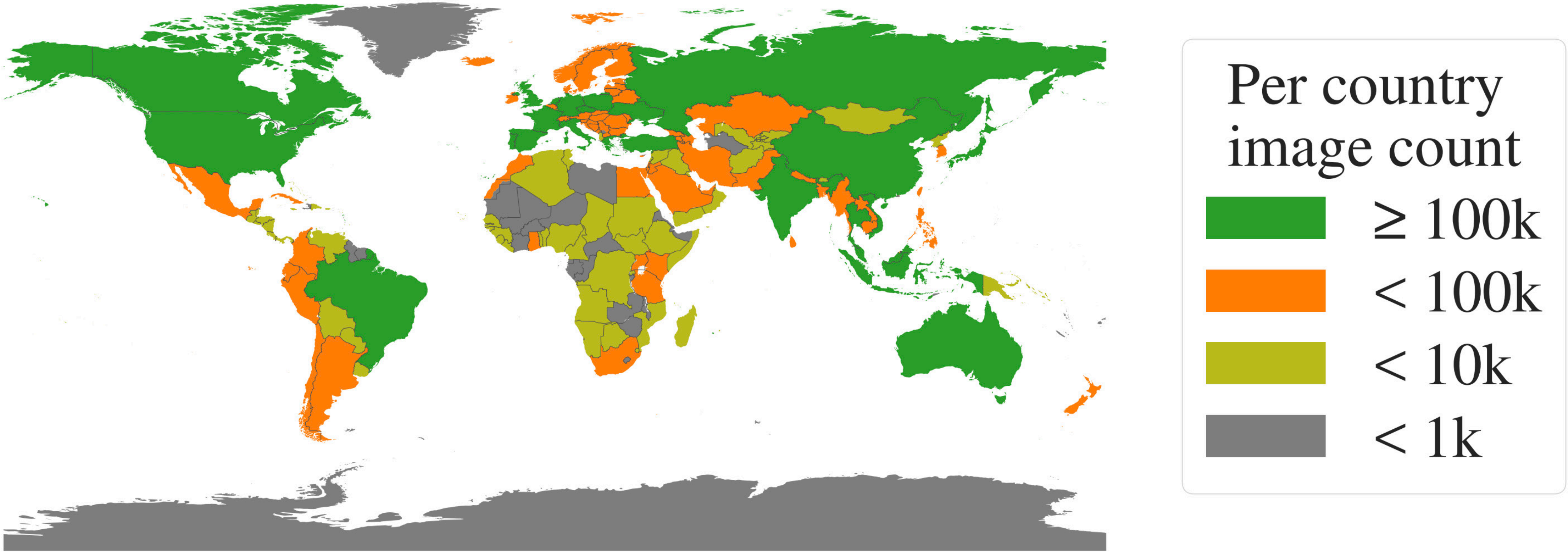
Prompt with 32 x 32 grid points

Train SAM on the collected data so far (300 K images with 10.2 M masks)
Predict 3 outputs, i.e., whole, part, and subpart.

Resulting in the SA-1B dataset consisting of 11M *high-resolution* images (3300x4950) with *automatically generated* 1.1B masks

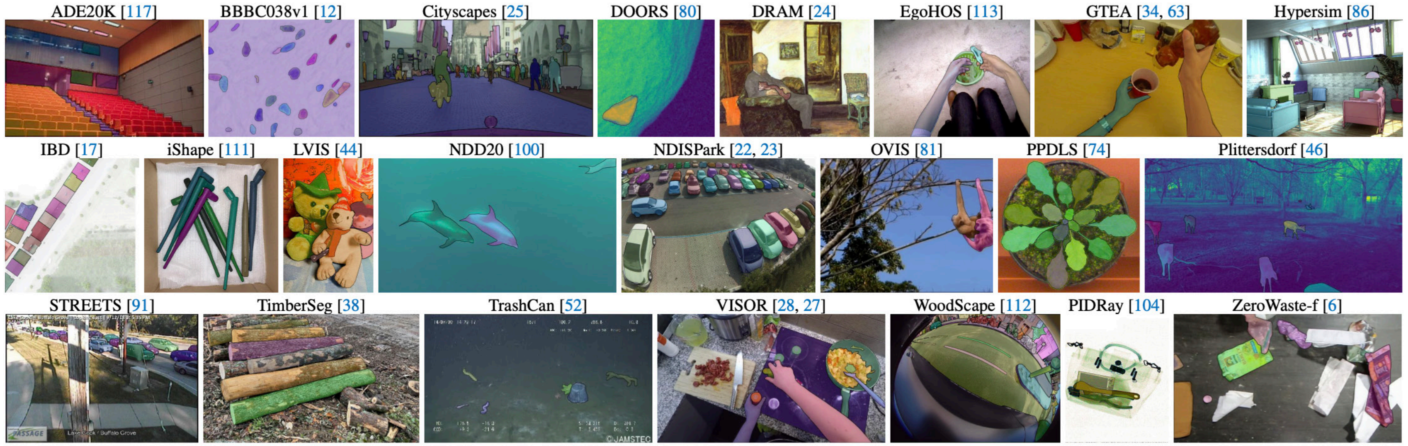
Images from https://segment-anything.com/

SA-1B Dataset -- Geographic Distribution

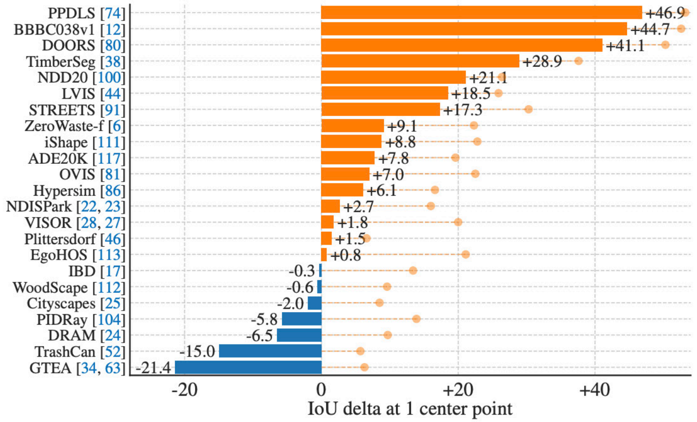


Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

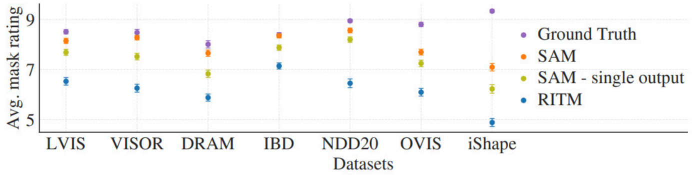
Zero-shot Single Point Valid Mask Evaluation



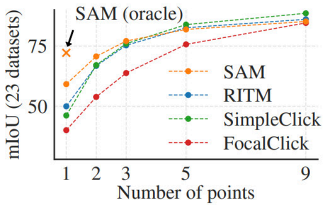
Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.



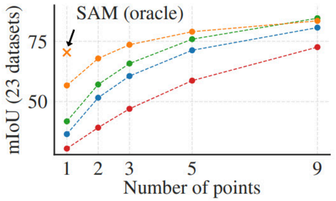
(a) SAM *vs*. RITM [92] on 23 datasets



(b) Mask quality ratings by human annotators



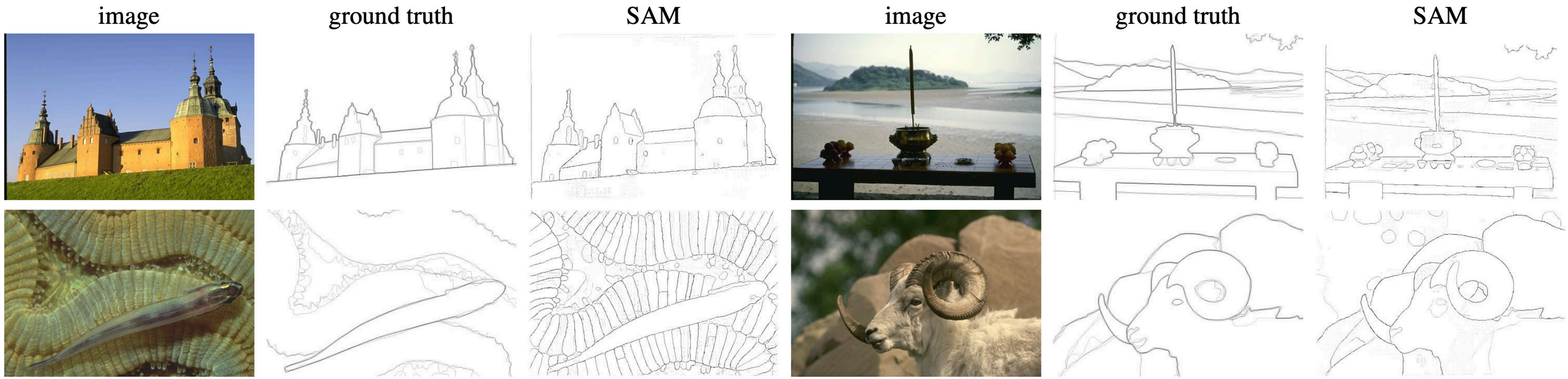
(c) Center points (default)



(d) Random points

Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

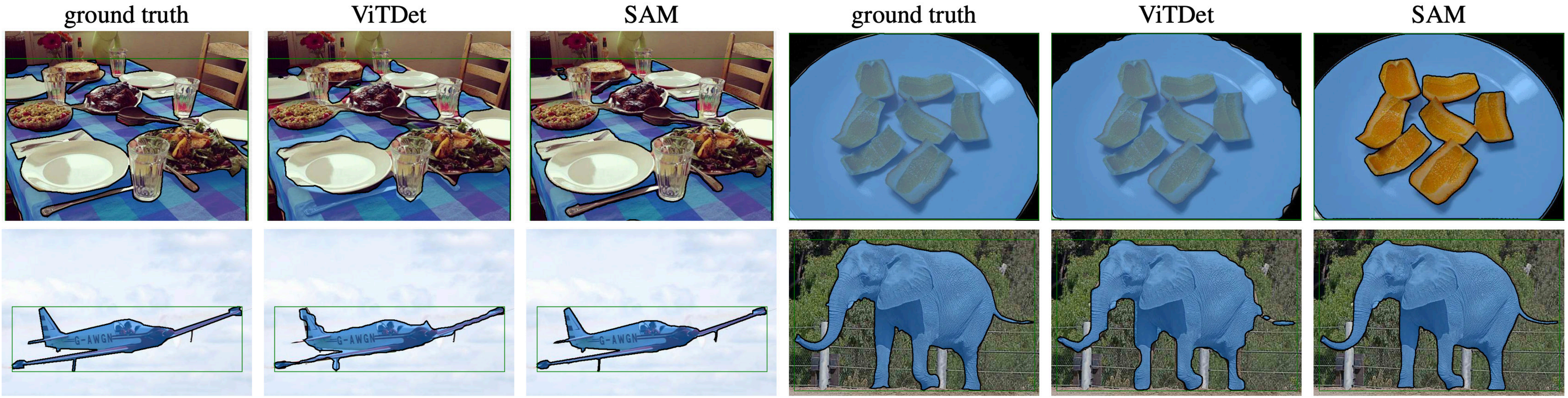
Zero-shot edge detection



Additional visualizations of zero-shot edge predictions on BSDS500. Recall that SAM was not trained to predict edge maps and did not have access to BSDS images and annotations during training.

Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

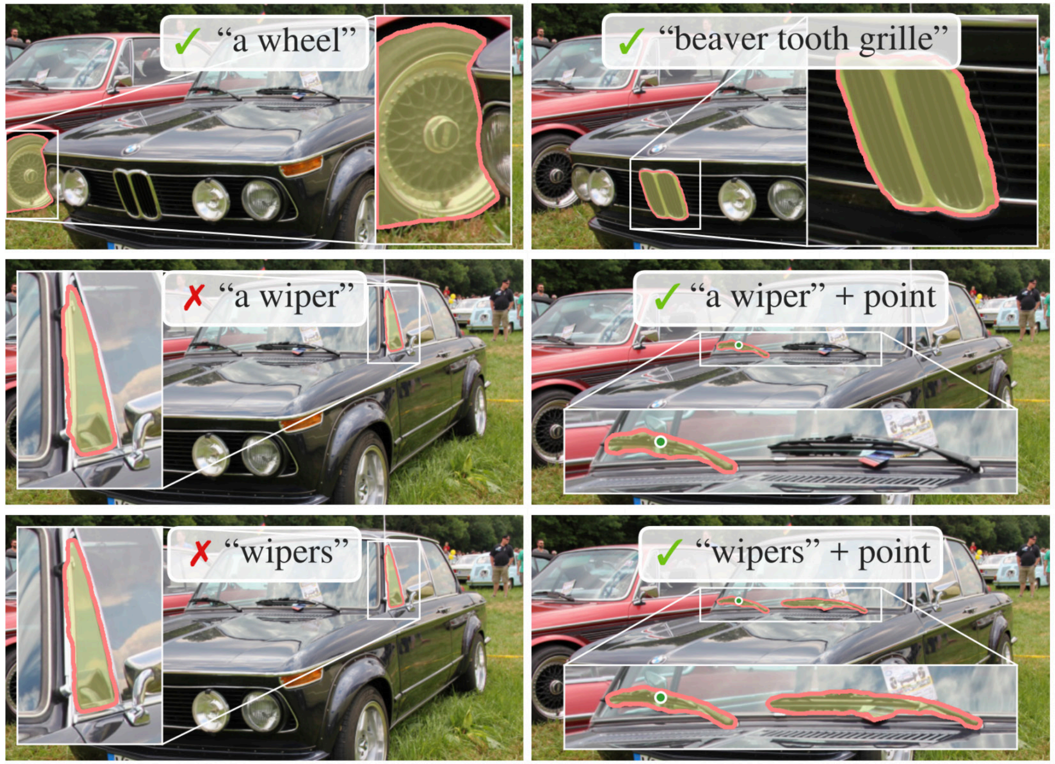
Zero-shot Instance Segmentation



Zero-shot instance segmentation on LVIS v1. SAM produces higher quality masks than ViTDet. As a zero-shot model, SAM does not have the opportunity to learn specific training data biases; see top-right as an example where SAM makes a modal prediction, whereas the ground truth in LVIS is amodal given that mask annotations in LVIS have no holes.

Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Zero-shot Text-to-Mask



Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Kirillov, Alexander, et al. "Segment anything." ICCV 2023.

Grounded-SAM: Grounded DINO + SAM

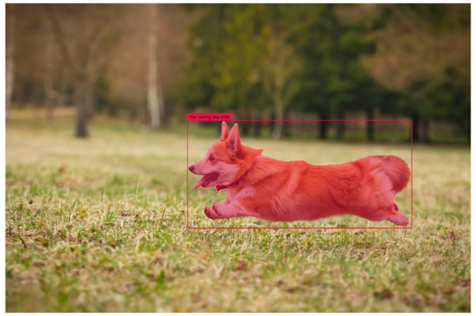
Grounded DINO: Detect anything with text prompt

Grounded SAM: Detect and segment anything with text prompt

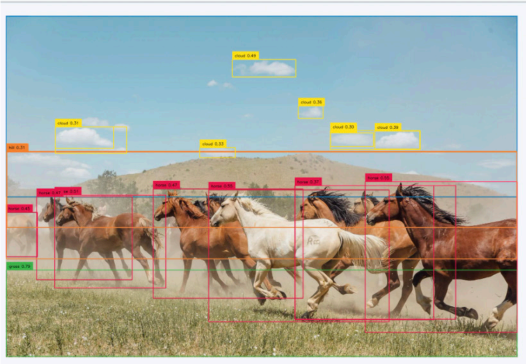
The running dog

Grounded DINO

SAM



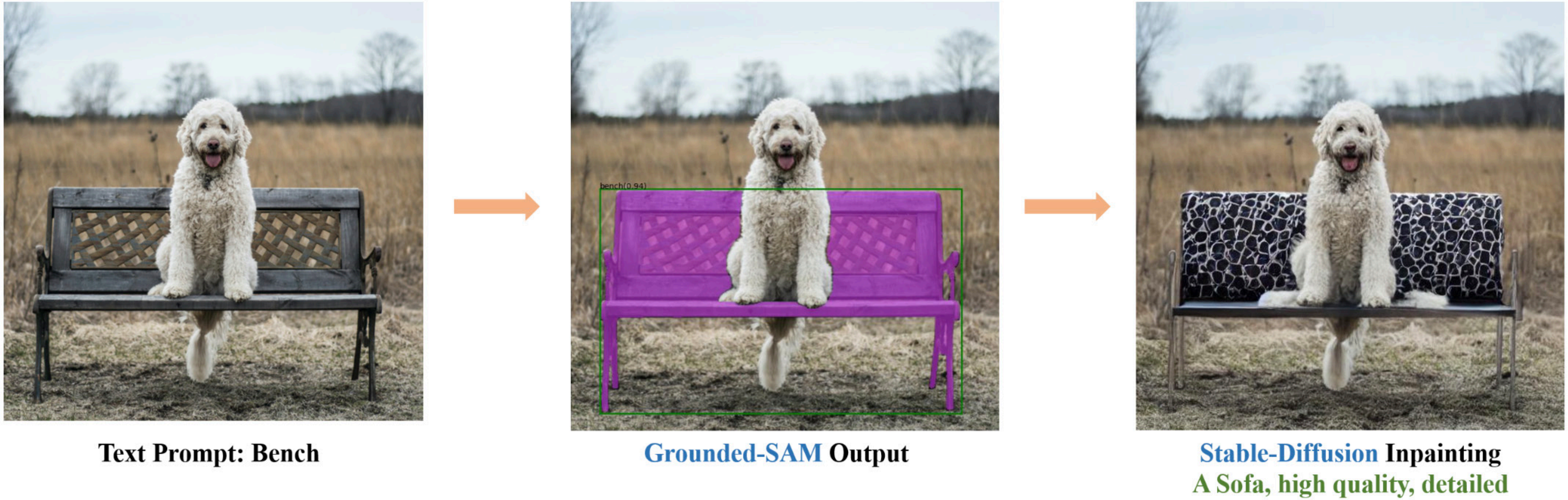
Horse.
Clouds.
Grasses.
Sky. Hill



Images from https://github.com/IDEA-Research/Grounded-Segment-Anything

Grounded-SAM + Stable-Diffusion Inpainting

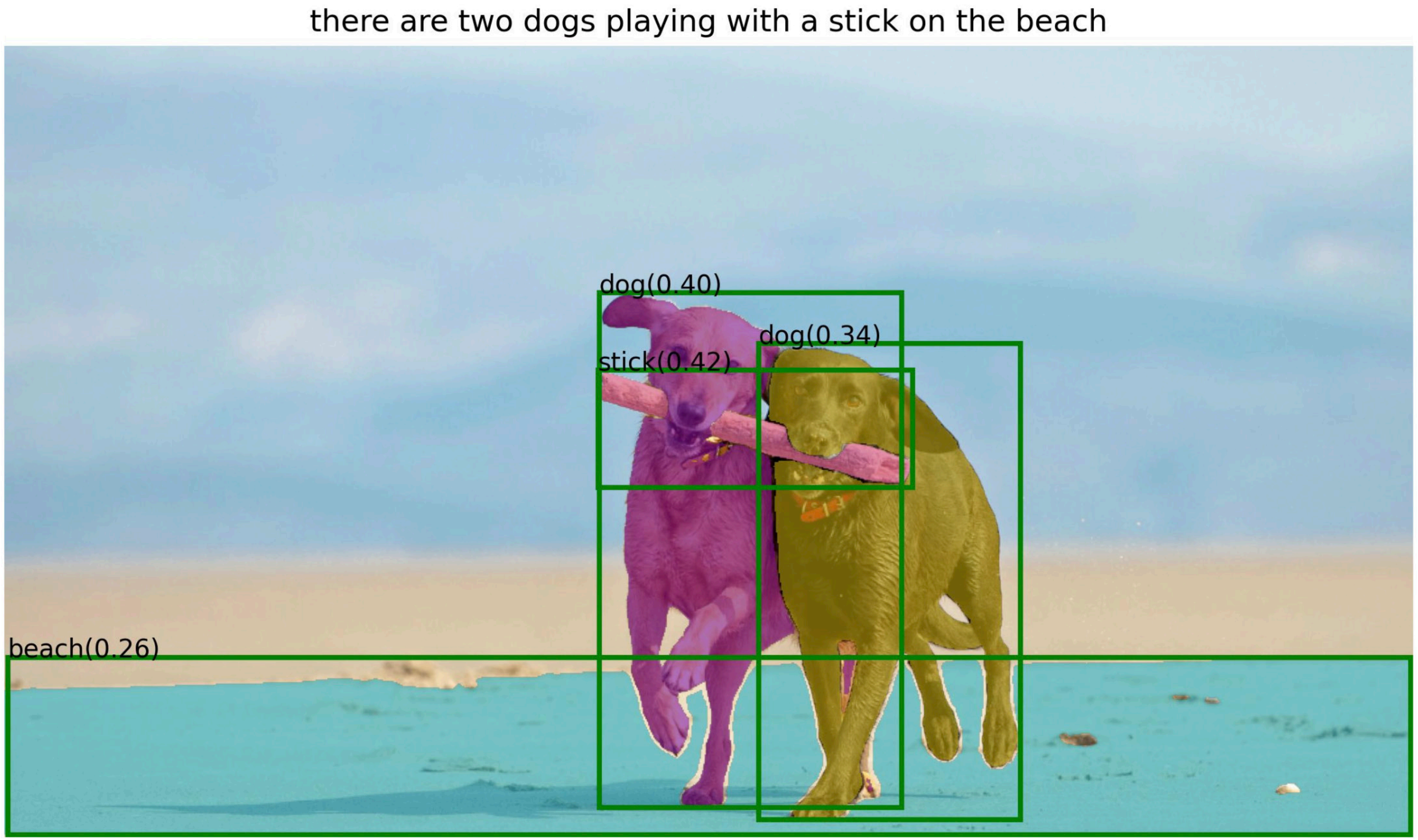
Generating new data!



Images from https://github.com/IDEA-Research/Grounded-Segment-Anything

BLIP + Grounded-SAM

Automatic Labeling



Images from https://github.com/IDEA-Research/Grounded-Segment-Anything

Grounded-SAM + Whisper

Detect anything with text prompt with speech

Images from https://github.com/IDEA-Research/Grounded-Segment-Anything

Conclusion

SAM

- defines a generalized segmentation approach: *promptable segmentation*
- builds a model that supports flexible prompting and real-time inference
- build a data engine that acquired the largest ever segmentation dataset SA-1B

Questions?