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

2023 arXiv
Facebook research

2023.05.17 공학 딥러닝

이소영

Contents

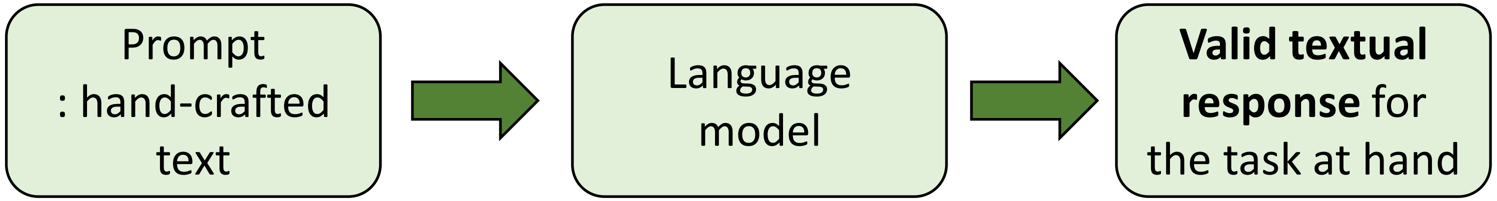
Abstract

- **Segment Anything** (SA) project: a new **task**, **model**, and **dataset** for **image segmentation**

- The **largest segmentation dataset**: over 1 billion masks on 11M images.
- The model is designed and trained to be **promptable**.
- Its capabilities on numerous tasks and find that its **zero-shot performance** is impressive.

Introduction

- Revolutionizing NLP
  - Large language models pre-trained on web-scale datasets = **foundation models**
  - Implemented with **prompt engineering**
  - Abundant text corpora from the web → zero & few-shot performance compare surprisingly well to fine-tuned models



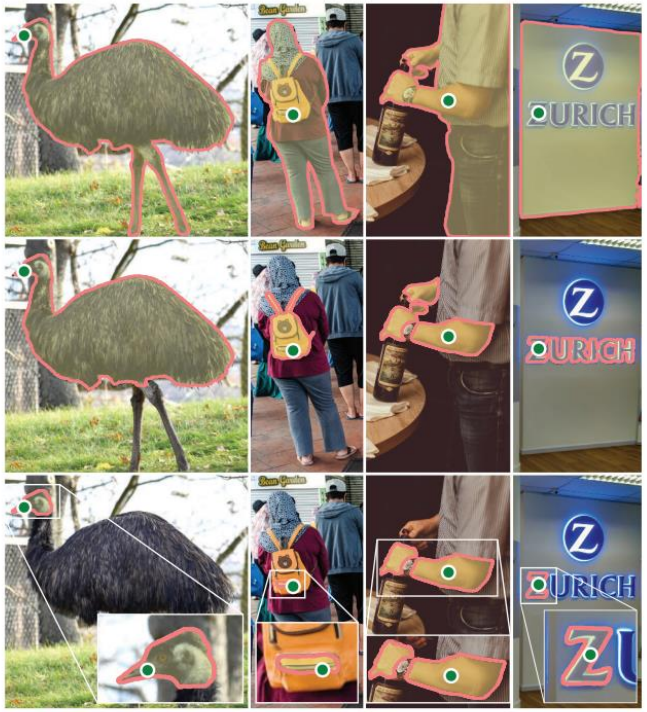
  - Empirical trends: improving with **model scale**, **dataset size**, and **total training compute**
- Foundation models in computer vision
  - Paired text and images align from the web (ex. CLIP)
  - **Abundant training data does not exist**

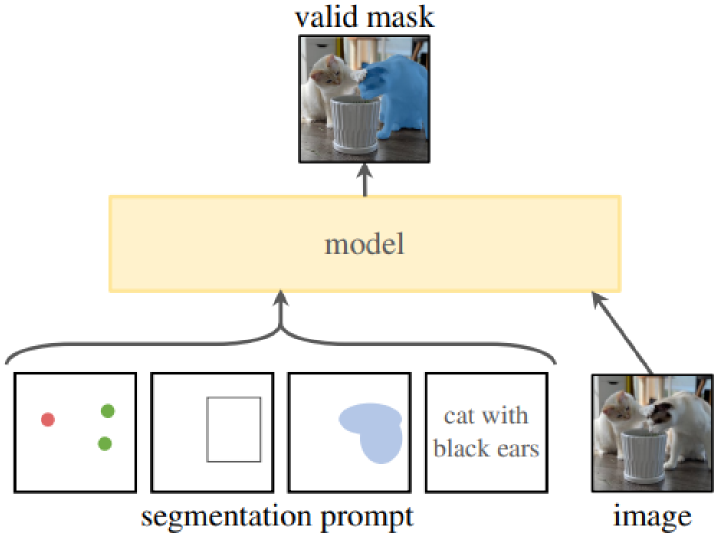
Introduction

- Our goal: **build a foundation model for image segmentation**
- The success of this plan hinges on three components: **task**, **model**, and **data**.
- To develop them, we address the following questions about image segmentation:
  - What **task** will enable zero-shot generalization? → Promptable segmentation task
  - What is the corresponding **model** architecture? → Supports flexible prompting
  - What **data** can power this task and model? → Diverse, large-scale source of data

Segment Anything - Task

- Promptable segmentation task
  - Return a valid segmentation mask given any segmentation prompt





(a) **Task**: promptable segmentation

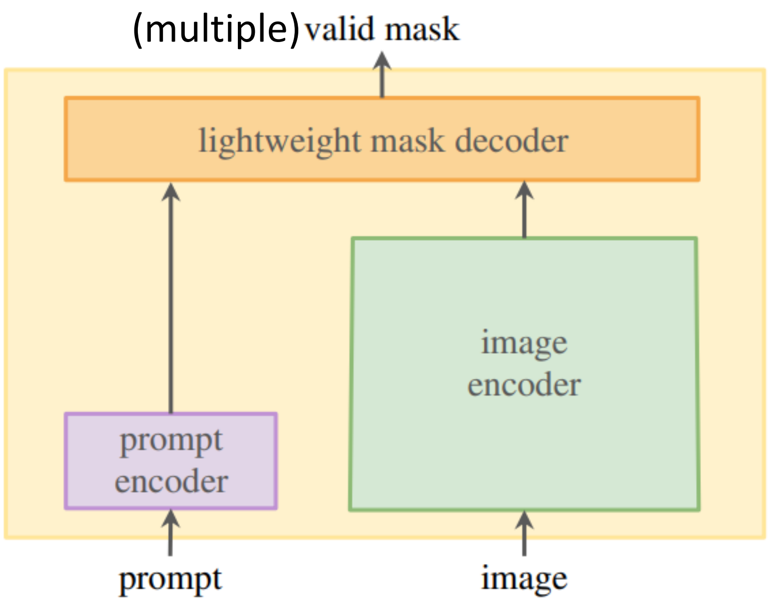
When a prompt is ambiguous and could refer to multiple objects, the output should be a **reasonable mask for at least one of those objects**.

Simply specifies what to segment in an image (e.g., point, bbox, mask, or text information identifying an object)

Segment Anything - Model

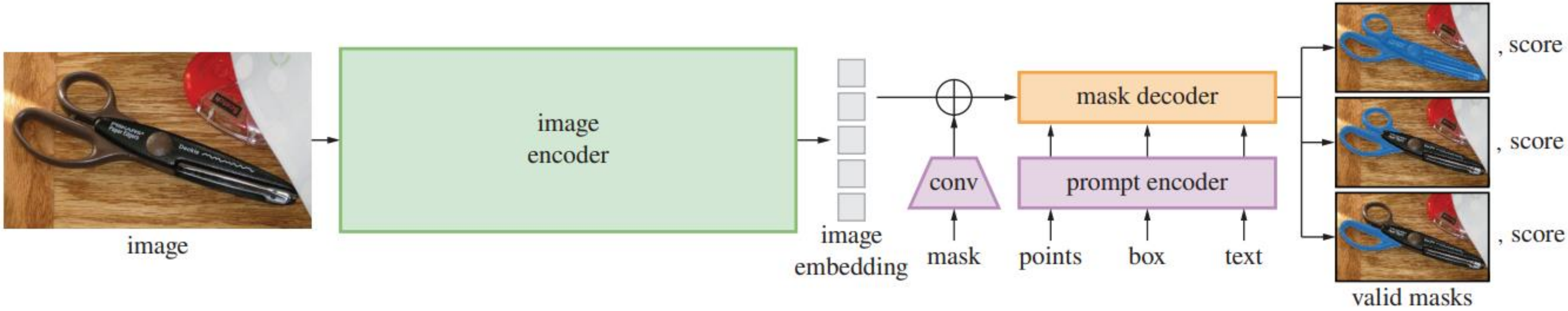
- Constraints on the model architecture
  - Must support flexible prompts
  - Needs to compute masks in amortized real-time to allow interactive use
  - Must be ambiguity-aware

→ Segment Anything Model (SAM)



(b) **Model**: Segment Anything Model (**SAM**)

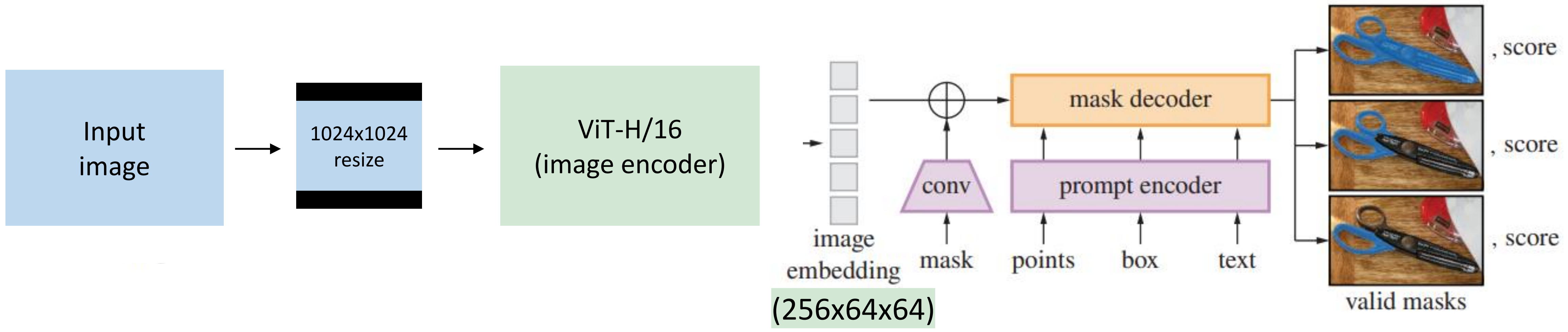
Segment Anything - Model



- **Image encoder**
  - **Masked Autoencoder (MAE)** pre-trained Vision Transformer (ViT): ViT-H/16
  - **Computed only once per image**, not per prompt

→ the prompt encoder and mask decoder predict a mask from a prompt in **~50ms** in a web browser

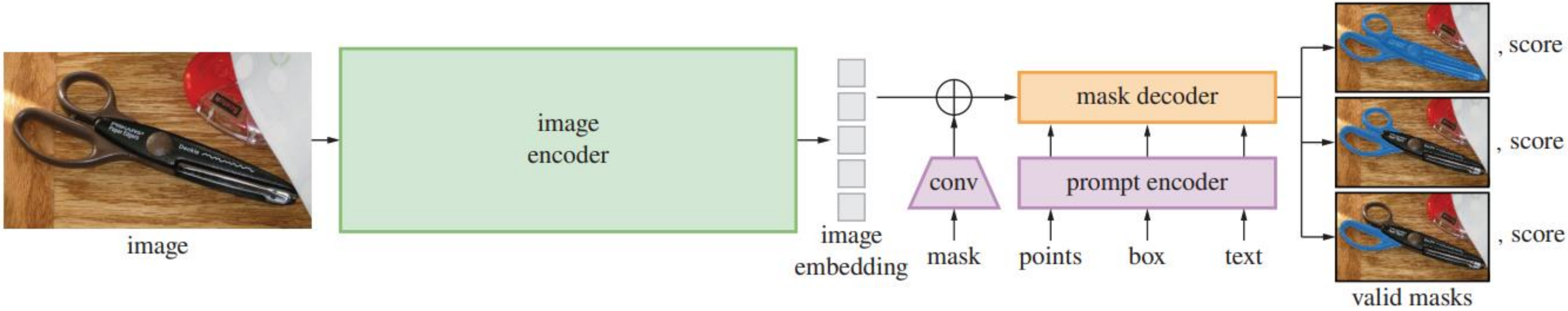
Segment Anything - Model



- **Image encoder**
  - **Masked Autoencoder(MAE)** pre-trained Vision Transformer (ViT): ViT-H/16
  - **Computed only once per image**, not per prompt

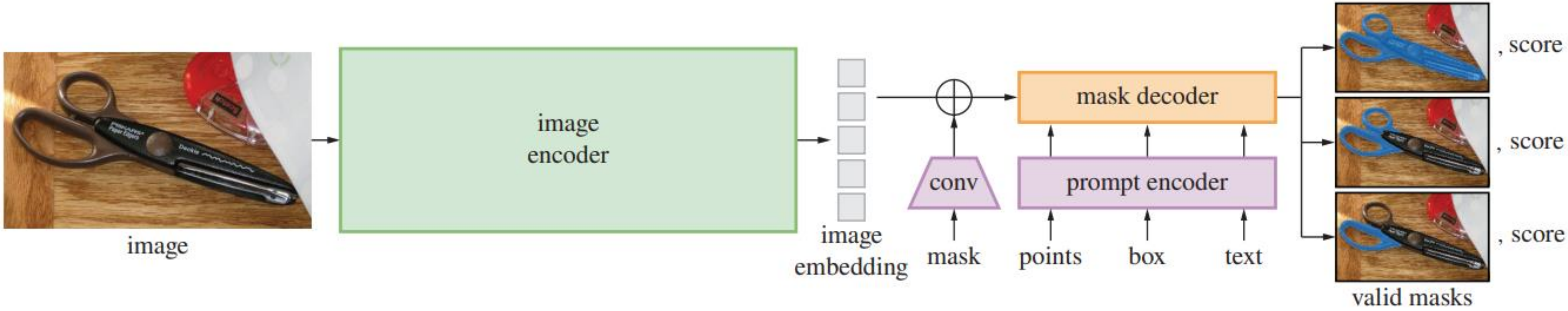
  → the prompt encoder and mask decoder predict a mask from a prompt in **~50ms** in a web browser

Segment Anything - Model



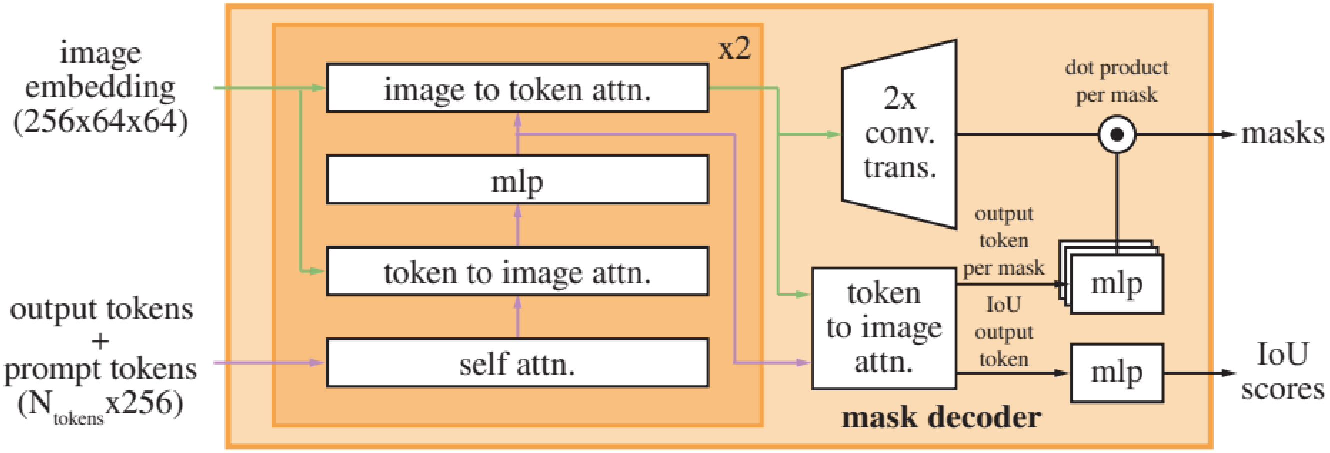
- **Prompt encoder**
  - **Sparse** prompts (points, boxes, text): mapped to 256- dimensional vectorial embeddings
    - Points: positional encoding of the point’s location + foreground/background embedding (learned)
    - Boxes: pair of two components
      - (1) the positional encoding of its top-left corner + learned embedding representing “top-left corner”
      - (2) the same structure but using a learned embedding indicating “bottom-right corner”
    - Free-from text: text encoder from CLIP
  - **Dense** prompts (masks): convolutions and summed element-wise with the image embedding

Segment Anything - Model

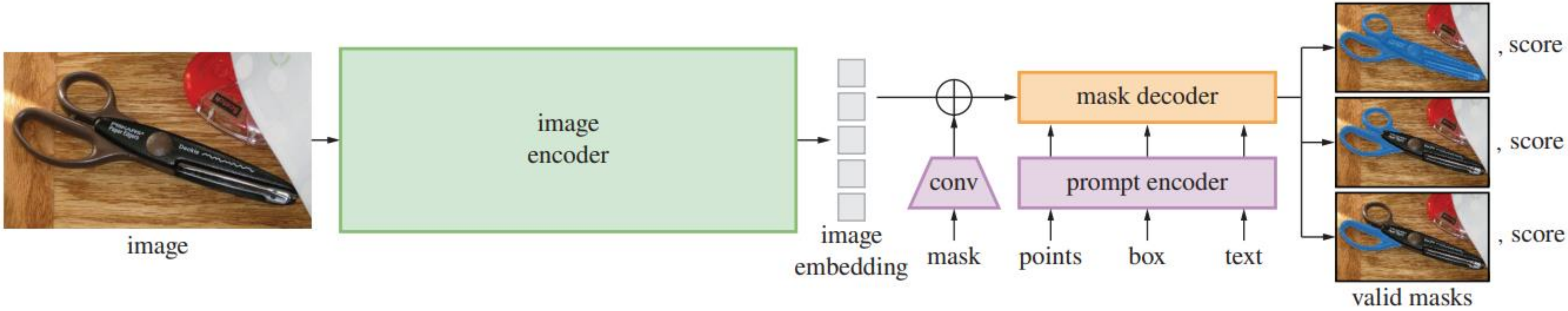


- **(Lightweight) Mask decoder**

similar to
ViT [class] token

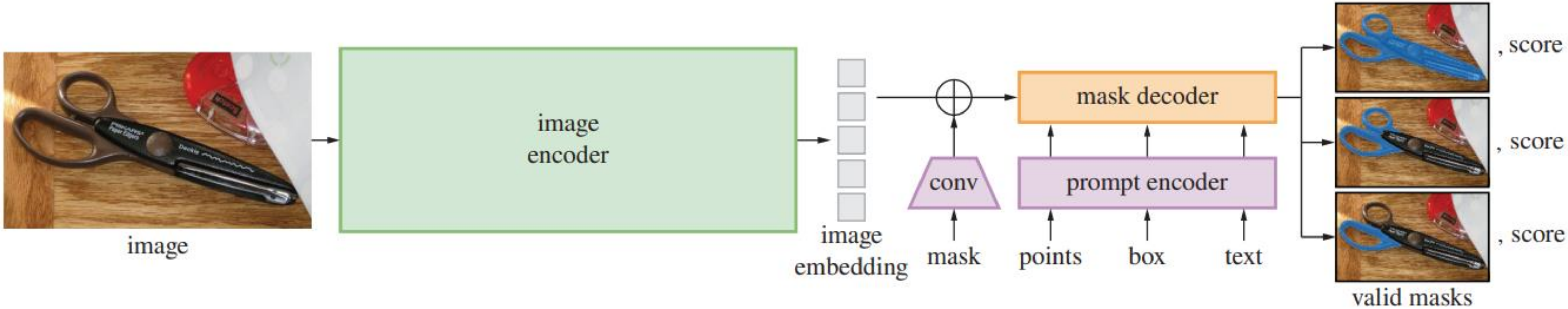


Segment Anything - Model



- **Resolving ambiguity**
  - Predict **3 masks**: whole, part, and subpart
  - Compute the loss between the ground truth and **each of the predicted masks**
  - but only **backpropagate from the lowest loss**
  - Add a small head: that **estimates the IoU (score)** between each predicted mask and the object it covers

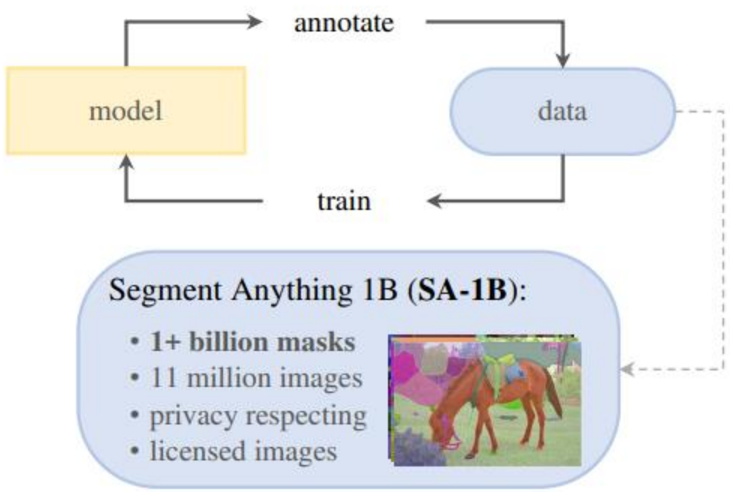
Segment Anything - Model



- Loss and training
  - **Mask** prediction: Focal loss + dice loss (20:1)
  - **IoU** prediction: mean-square-error (MSE) loss

Segment Anything - Data Engine

- **1.1B mask dataset, SA-1B**
- The data engine has three stages:
  1. **Model-assisted manual annotation stage**
  2. **Semi-automatic stage** with a mix of automatically predicted masks and model-assisted annotation
  3. **Fully automatic stage** in which our model generates masks without annotator input.



(c) **Data**: data engine (top) & dataset (bottom)

Segment Anything - Data Engine

1. Assisted-manual stage

- Human
  - Clicking foreground / background object points → refine
  - Proceed to the next image once a mask took over 30 seconds to annotate
- SAM
  - Start: trained using common **public segmentation datasets**
  - **retrained** using only newly annotated masks (**retraining 6 times**)
  - SAM scaled from ViT-B to ViT-H

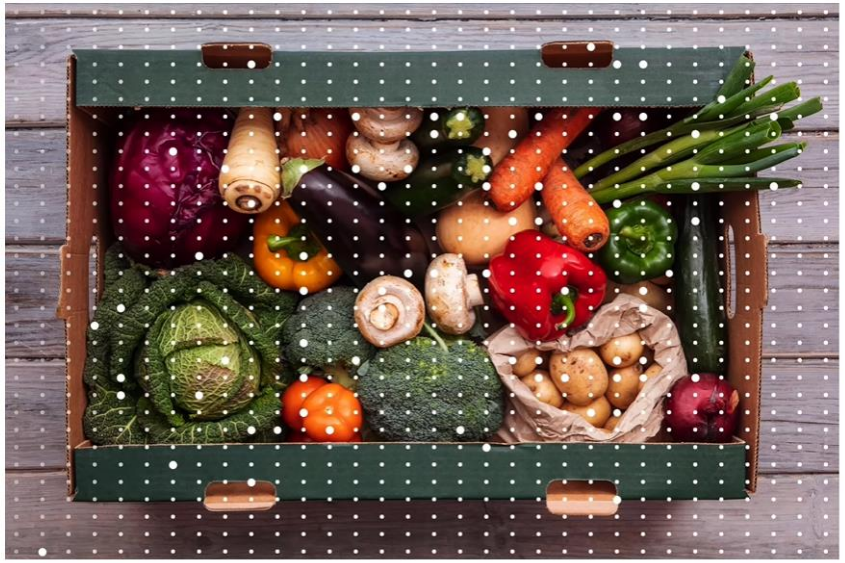
- Overall, we collected **4.3M masks from 120k images** in this stage.

Segment Anything - Data Engine

2. Semi-automatic stage

- Aim: **increase the diversity** of masks
- First **automatically (SAM) detected confident masks → annotate (human) any additional** unannotated objects
- Detect **confident** masks
    - Trained a bounding box **detector(Faster R-CNN)** on all first stage masks using a generic "object" category.
- Collected an **additional 5.9M masks in 180k images** (for a total of 10.2M masks).

Segment Anything - Data Engine



3. Fully automatic stage

- Developed the **ambiguity-aware model**
- Prompt: the model with a **32×32 regular grid of points**
- Select confident masks
  - A mask stable if thresholding the probability map at $0.5 - \delta$ and $0.5 + \delta$ results in similar masks
- Filter duplicates: using non-maximal suppression (NMS)

- **Fully automatic** mask generation to all **11M images**, producing a total of **1.1B high-quality masks**

Segment Anything - Dataset

- **Dataset: SA-1B**
- Images
  - 11M diverse, licensed
  - High-resolution: 3300×4950 pixels on average
- Masks
  - Data engine produced 1.1B masks, 99.1% of which were generated fully automatically.
  - Quality analysis → **SA-1B only includes automatically generated masks**

Segment Anything - Dataset

- **Mask quality**
  - Randomly sampled 500 images (~50k masks)

    → professional annotators to improve the quality of all masks in these images.
  - Pairs of automatically predicted and professionally corrected masks
    - **94% of pairs have greater than 90% IoU** (97% of pairs have greater than 75% IoU)
    - Prior work estimates **inter-annotator consistency at 85-91% IoU**

Segment Anything - Dataset

- **Mask properties**
  - Greater coverage of image corners

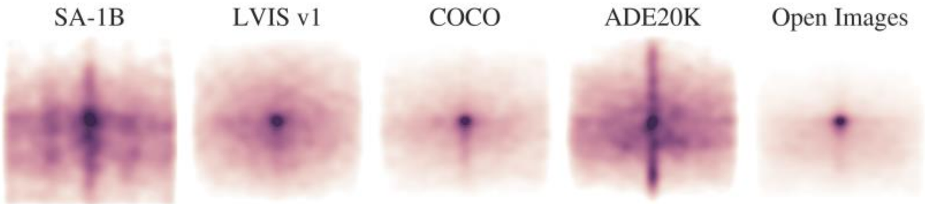
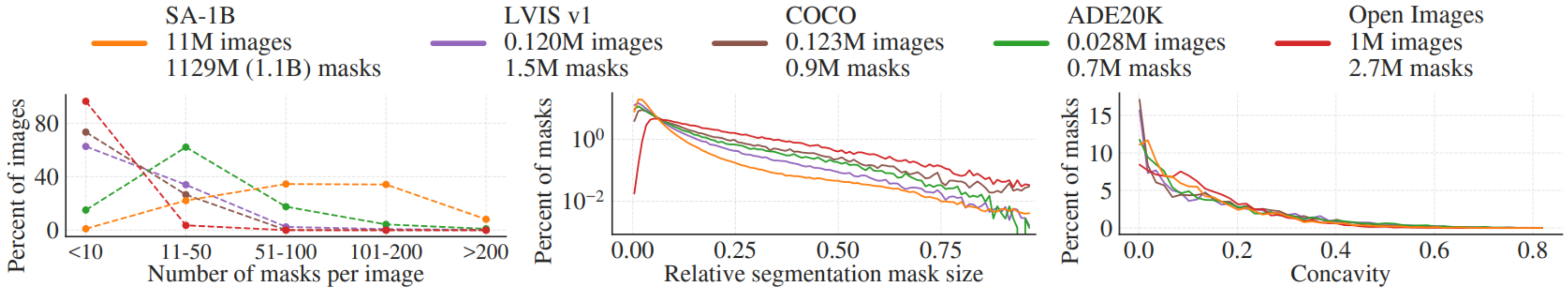


Figure 5: Image-size normalized mask center distributions.

  - Compare by size



more masks per image

greater percentage of small and medium relative-size masks

similar shape complexity

Zero-Shot Transfer Experiments

- The datasets may include **novel image distributions**, such as underwater or ego-centric images that, to our knowledge, do not appear in SA-1B.

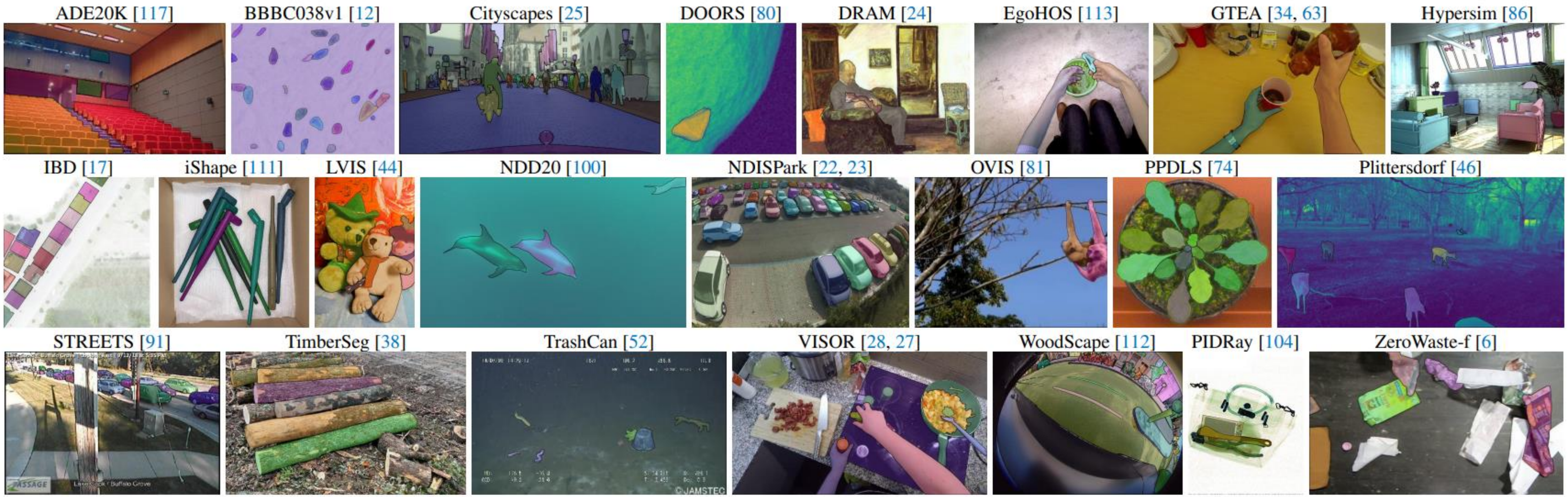
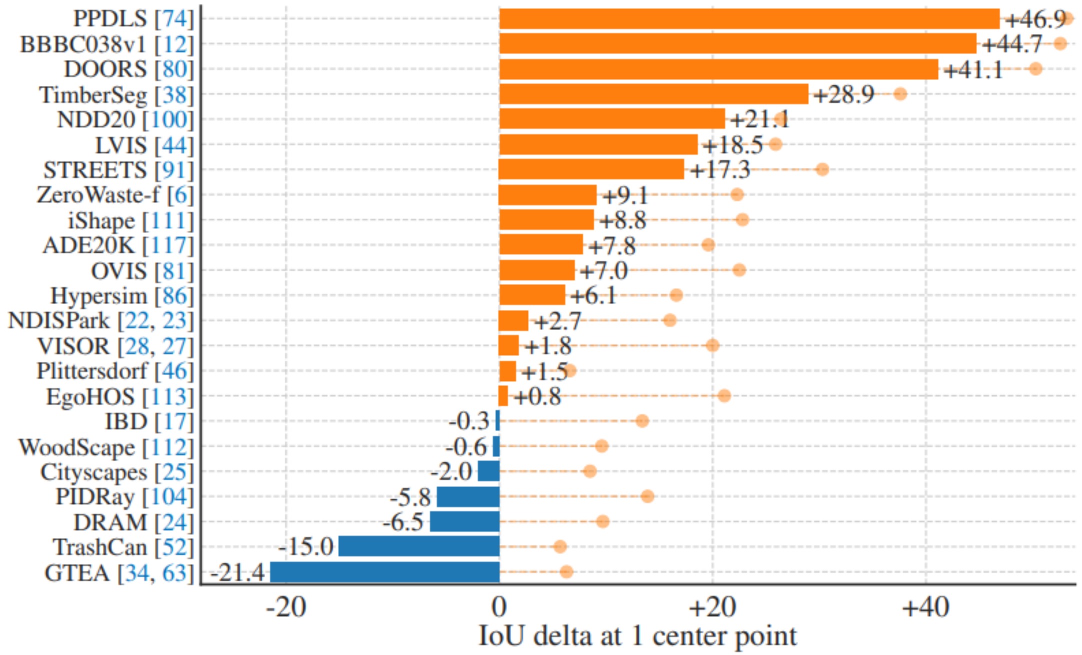


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.

Zero-Shot Transfer Experiments

1. Zero-Shot **Single Point Valid Mask Evaluation Task**

- Task: segmenting an object from a single foreground point
  - evaluate only the model's most confident mask by default
- Result



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

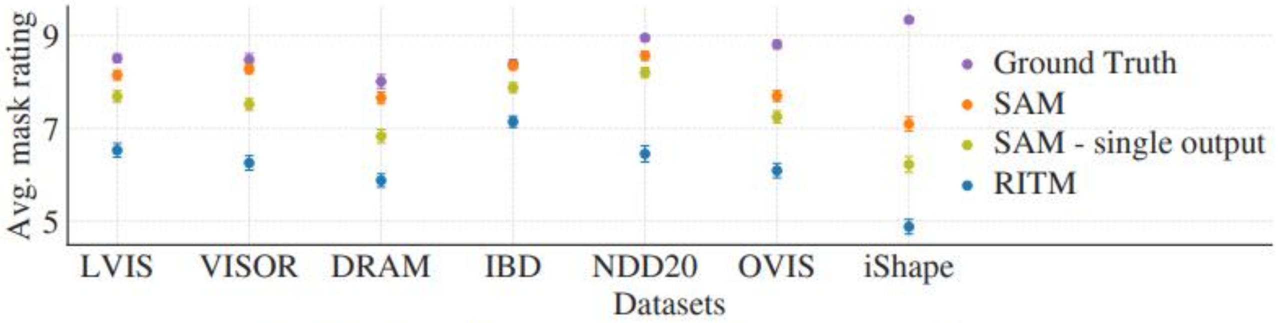
SAM yields higher results on 16 of the 23 datasets
"oracle" result: SAM outperforms RITM on all datasets

"oracle" result
- Most relevant of SAM's 3 masks is selected by comparing them to the ground truth
- Reveals the impact of ambiguity on automatic evaluation

Zero-Shot Transfer Experiments

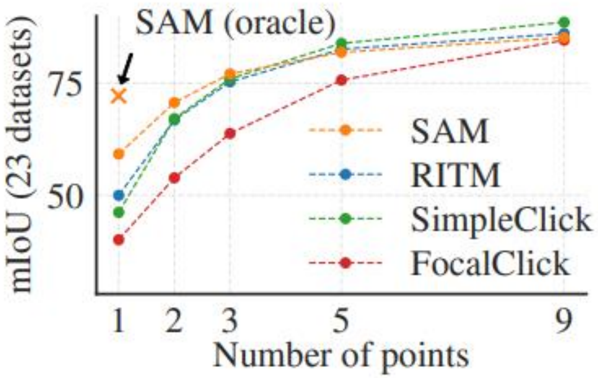
1. Zero-Shot **Single Point Valid Mask Evaluation Task**

- Result

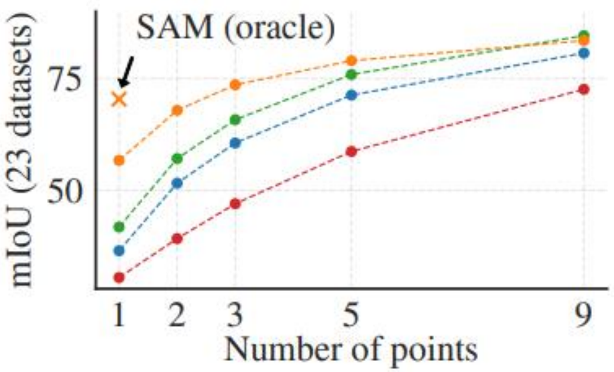


(b) Mask quality ratings by human annotators

Annotators consistently rate the quality of SAM's masks substantially higher than RITM



(c) Center points (default)



(d) Random points

the number of points increases from 1 to 9
→ the gap between methods decreases
: task becomes easier

random point sampling
→ the gap between SAM and the baselines grows

Zero-Shot Transfer Experiments

2. Zero-Shot **Edge Detection**

- Approach: evaluate SAM on the classic low-level task of edge detection using BSDS500
    - Using a simplified version of our automatic mask generation pipeline
    - edge maps are computed
        - Sobel filtering of unthresholded mask probability maps
        - Standard lightweight postprocessing, including edge NMS

Zero-Shot Transfer Experiments

2. Zero-Shot **Edge Detection**

4 standard metrics for edge detection
ODS: Optimal Dataset Scale
OIS: Optimal Image Scale
AP: Average Precision
R50: Recall at 50% precision

- Results

method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

Table 3: Zero-shot transfer to edge detection on BSDS500.

Qualitatively: produces reasonable edge maps

SAM predicts more edges (bias),
including sensible ones that are not annotated in BSDS500

SAM's bias: high R50, but low AP

Better than zero-shot transfer methods

Zero-Shot Transfer Experiments

3. Zero-Shot **Object Proposals**

- Approach: evaluate SAM on the mid-level task of object proposal generation
  - Compute the standard average recall (**AR**) metric for masks at 1000 proposals on LVIS v1
- Results

	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

Table 4: Object proposal generation on LVIS v1. SAM is applied zero-shot, *i.e.* it was not trained for object proposal generation nor did it access LVIS images or annotations.

ViTDet-H as object proposals performs the best overall

SAM does remarkably well on several metrics

Zero-Shot Transfer Experiments

4. Zero-Shot **Instance Segmentation**

- Approach: higher-level vision
  - Object detector (same as ViTDet) output boxes = prompt SAM
- Results

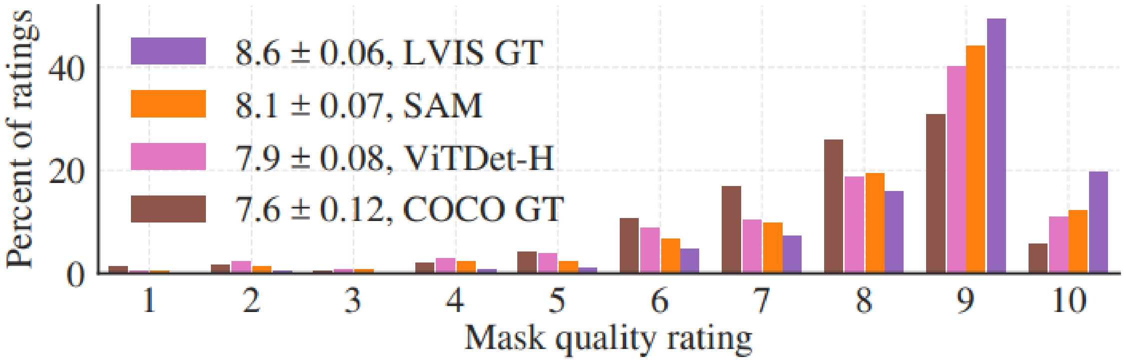
	COCO [66]				LVIS v1 [44]			
method	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L
ViTDet-H [62]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5

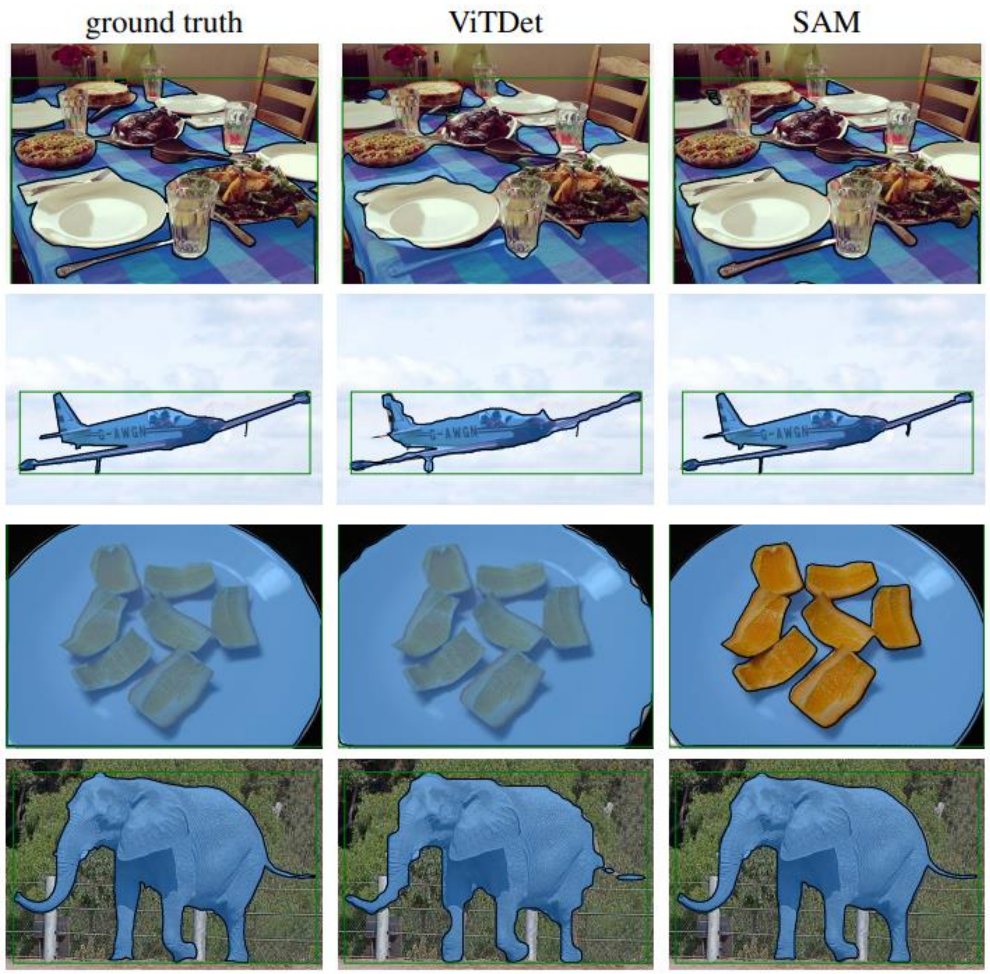
AP metric: ViTDet-H > SAM on both datasets

Zero-Shot Transfer Experiments

4. Zero-Shot **Instance Segmentation**

- Results
  - SAM masks are often **qualitatively better** than those of ViTDet





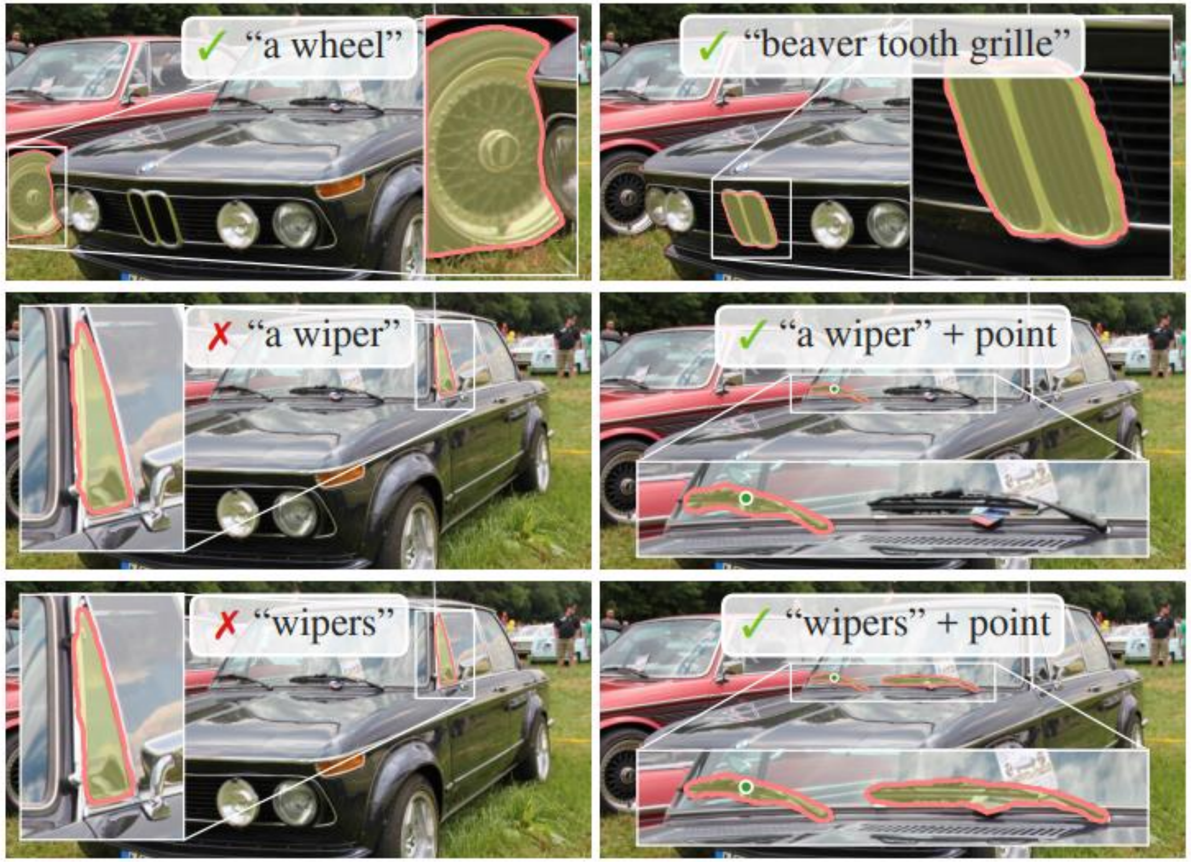
Zero-Shot Transfer Experiments

5. Zero-Shot **Text-to-Mask**

- Approach: a proof-of-concept of SAM's **ability to process text prompts**

- Results
  - SAM can segment objects based on simple text prompts
  - SAM fails to make a correct prediction

  → an additional point prompt can help

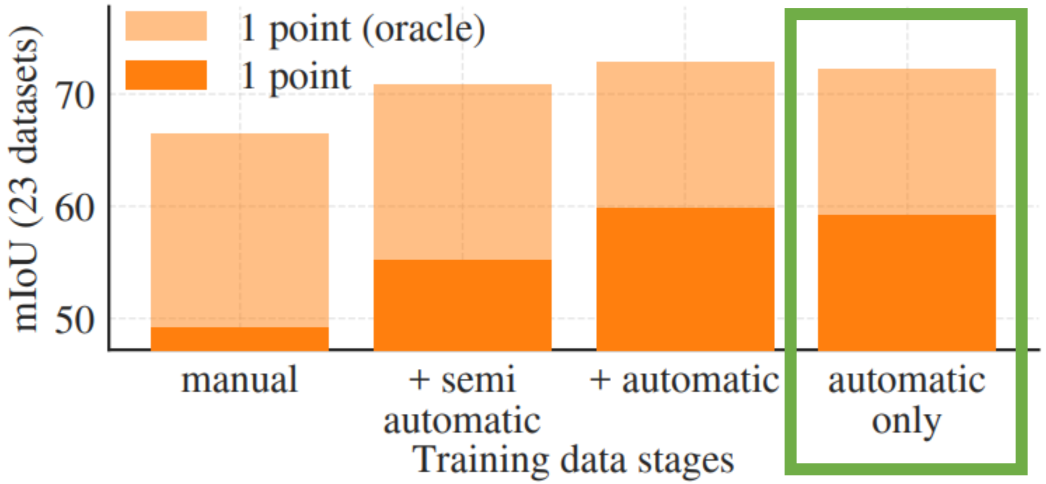


Zero-Shot Transfer Experiments

6. **Ablations**

- Data engine stages
  - Each stage increases mIoU
  - Tested a fourth setup that uses only the automatically generated masks
    - to simplify the training setup

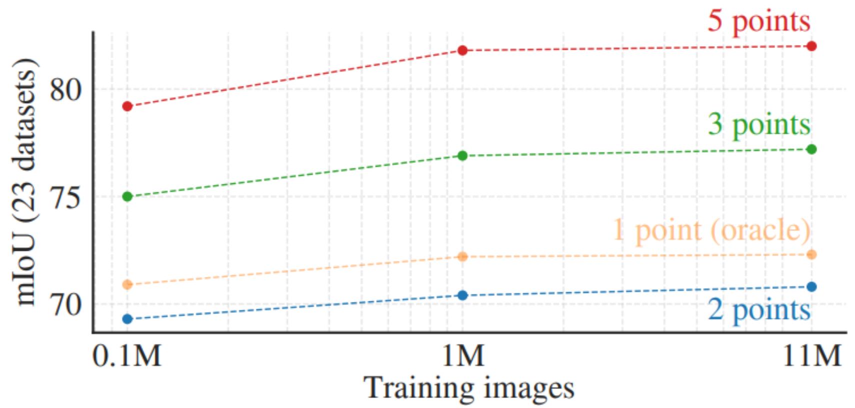
using



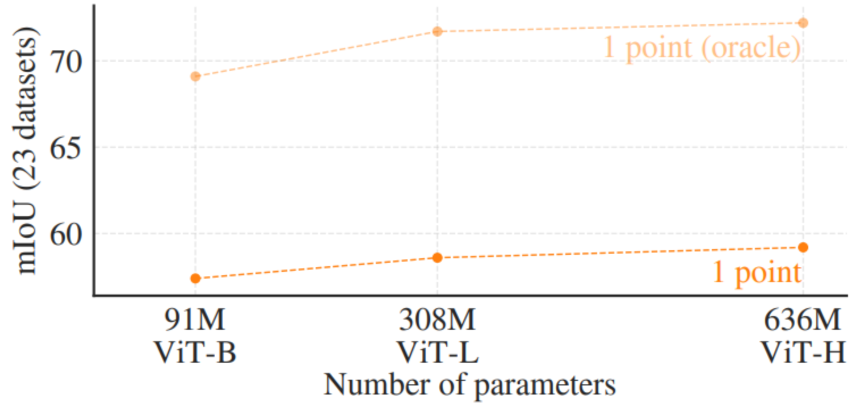
Zero-Shot Transfer Experiments

6. **Ablations**

- Training data scaling
  - 1M images: results comparable to using the full dataset



- Image encoder scaling
  - Further image encoder scaling does not appear fruitful at this time



Discussion

- **Foundation models**
  - Pre-trained model → foundation model (rebranded)
  - Trained on broad data at scale and are adaptable to a wide range of downstream tasks
- **Compositionality**
  - New capabilities: used **as a component** in larger systems
  - Our goal is to make this kind of **composition straightforward** with SAM.
    - SAM to predict a valid mask for a wide range of segmentation prompts

    → create a reliable interface between SAM and other components

Discussion

- **Limitations**
  - While SAM performs well in general, it is **not perfect**
  - Dedicated interactive segmentation methods to **outperform SAM when many points are provided**
  - Not real-time when using a heavy image encoder
  - **Text-to-mask task is exploratory and not entirely robust**
  - **Unclear how to design simple prompts** that implement semantic and panoptic segmentation
- **Conclusion**
  - The Segment Anything project is an attempt to lift image segmentation into the era of foundation models.
  - Our principal contributions: **a new task (promptable segmentation), model (SAM), and dataset (SA-1B)**

END