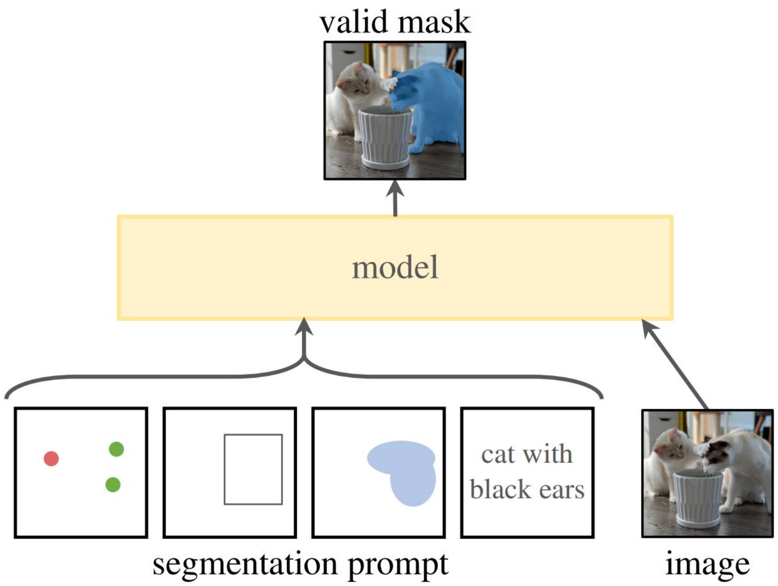
Segment Anything

Segment Anything

- foundation model for semantic segmentation
- promptable segmentation task
- data collection pipeline



Result example

Architecture description

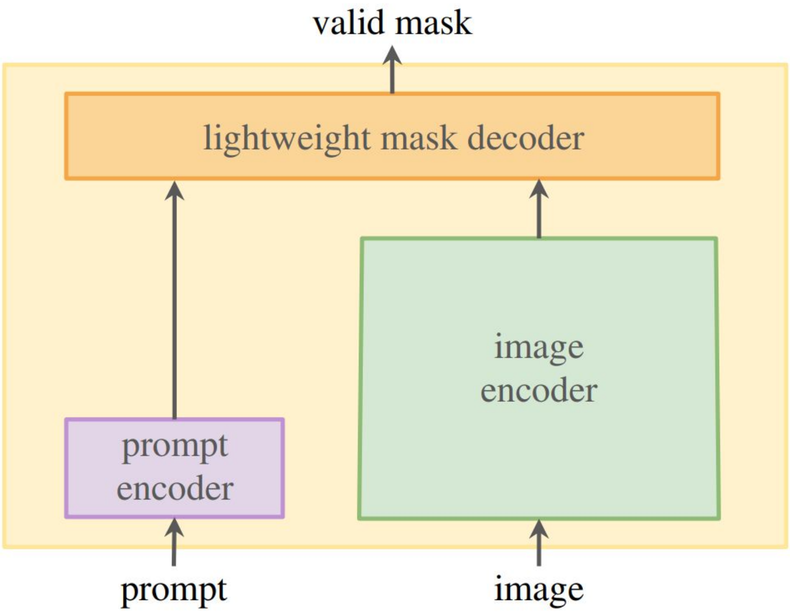
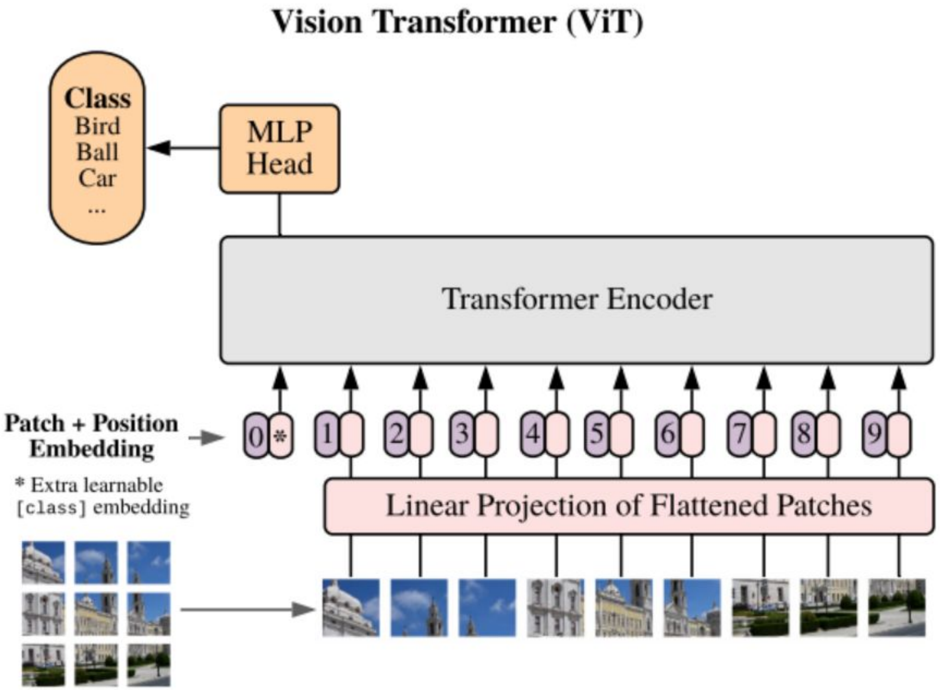


Image Encoder

- MAE (masked autoencoder) Vision Transformer
- Pre Training objective: reconstruct masked random patches
- One run per image
- Computationally heavy
- ViT-H/16 with 14×14 windowed attention
- Output 64×64 is a 16× downscaled embedding of the input image 1024x1024
- Post-processing CNN (1x1 conv + 3x3 conv) to reduce channel dim

Image Encoder: ViT



Prompt encoder

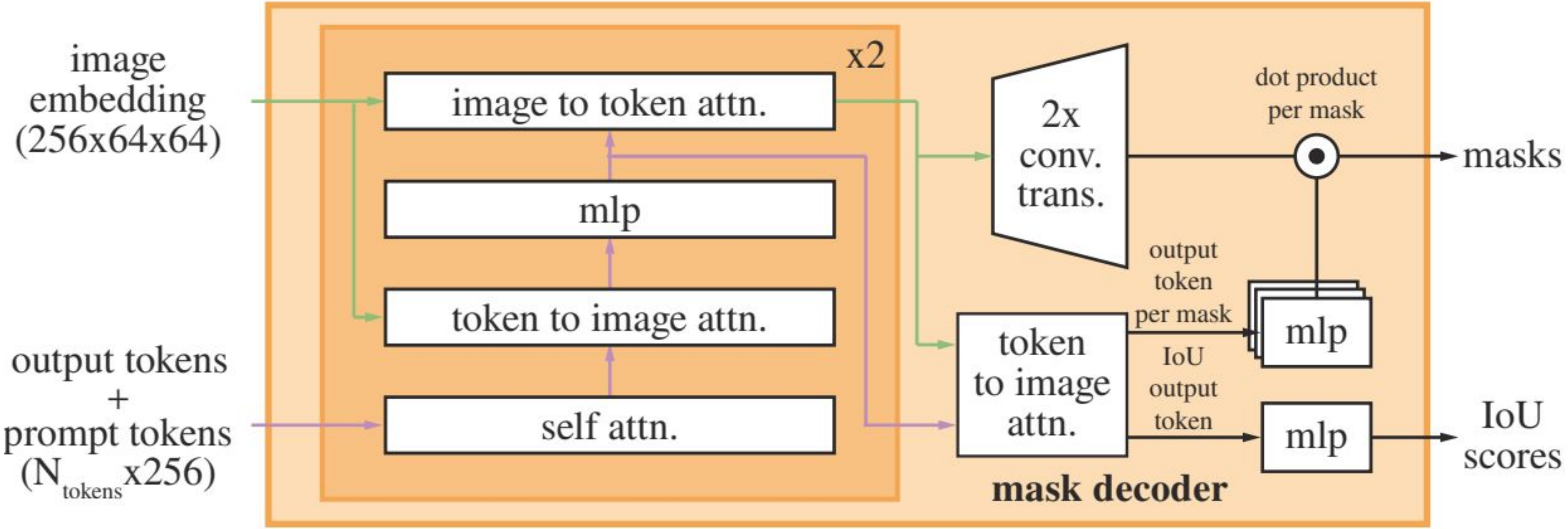
- Sparse prompts: points, boxes, text
- Dense prompts: masks
- Point/boxes prompts = positional embeddings + type embeddings
- Text prompts = text embeddings (CLIP)
- Dense prompts = convolutions + element-wise summation

```
self.mask_input_size = (4 * image_embedding_size[0], 4 * image_embedding_size[1])
self.mask_downscaling = nn.Sequential(
    nn.Conv2d(1, mask_in_chans // 4, kernel_size=2, stride=2),
    LayerNorm2d(mask_in_chans // 4),
    activation(),
    nn.Conv2d(mask_in_chans // 4, mask_in_chans, kernel_size=2, stride=2),
    LayerNorm2d(mask_in_chans),
    activation(),
    nn.Conv2d(mask_in_chans, embed_dim, kernel_size=1),
)
```

Dense prompt cnn

Mask decoder

- Positional embedding are added to image embeddings every time
- Flattening at cross-attention
- Original prompt tokens are re-added to the updated ones at attention layers
- Reduced dim in cross-attentions
- Image embeddings are copied for each mask



Training: losses

- Ambiguity dealing: predict 3 masks and choose the best for backprop.
- No ambiguity dealing if given 2+ prompts
- Focal loss + Dice loss for mask prediction
- MSE loss for IoU head

$$DiceLoss(y, \overline{p}) = 1 - \frac{(2y\overline{p} + 1)}{(y + \overline{p} + 1)}$$

$$\mathrm{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$$

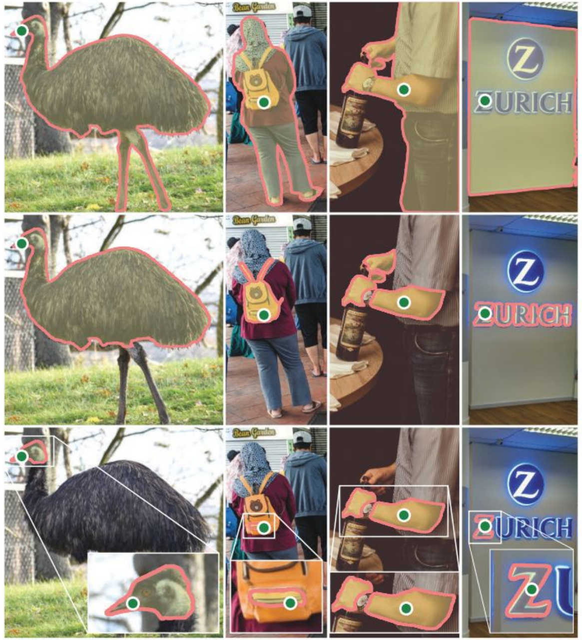


Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

Training: algorithm

Simulation of interactive segmentation:

1. choose foreground point (50%) or bounding box (50%)
   - if box, then use target bbox + noise (10% of side length, max 20px)
   - if point, then choose uniformly one from the ground truth mask
2. choose new point from the mispredicted area
   - add previous mask logits to prompt

1 initial iteration

8 iterations with new points

2 iteration with previous predictions, but without new points.

Training: misc

Drop path with rate 0.4

No data augmentation is applied

Large masks (covering 90% of image) are removed

Data Engine

1. Assisted-manual stage
2. Semi-automatic stage
3. Fully automatic stage

Assisted-manual stage

SAM is trained on previous datasets at the start

6 retraining, ViT-B to ViT-H scaling

Image embeddings are precomputed for in-browser model usage

“Brush” and “Eraser” tools

4.3M masks from 120k images

34 (initial) to 14 (final) seconds per mask (average time)

Semi-automatic stage

Goal: to increase the diversity of masks

Confident masks are automatically detected and presented to humans

Annotators label **only** unannotated objects

+5.9M masks in 180k images (for a total of 10.2M masks)

44 to 72 masks per image (including automatic)

34 seconds per mask (average time)

Fully automatic stage

- Use image and 20 zoomed-in overlapping crops
- Make 32×32 regular grid of points and prompt every point.
- Select only confident and stable masks
- Remove too large masks
- Apply NMS (in every crops firstly, then between crops)
- Remove mask of components with <100px area
- Fill holes of <100px area
- Different model

Fully automatic stage: model

- Larger training time 90k -> 177k
- Larger model
- Trained only on data from manual-assisted and semi-automatic stages
- Color jitter augmentation
- No box prompts
- Only 4 points per mask during training

Segment Anything Dataset

- 11M images
- downsampled (3300×4950 average -> 1500xN) and blurred where needed
- 1.1B masks (autogenerated)
- 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%
- Less photographer biases

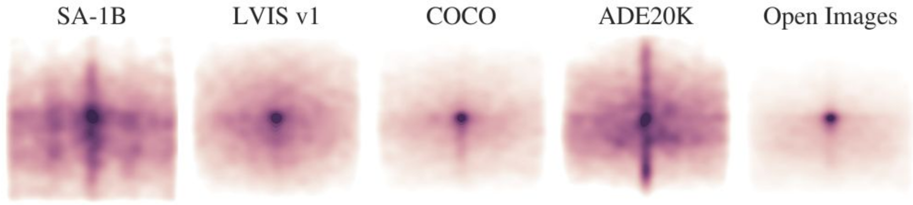


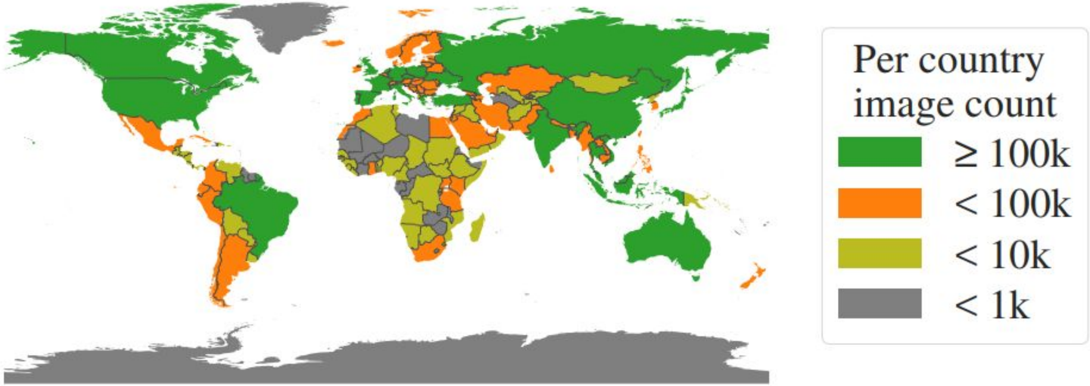
Figure 5: Image-size normalized mask center distributions.

Examples

Responsible AI (RAI)

- What countries are represented?
- Are they rich or poor? Where are they?
- Locations are inferred from captions for SA-1B, using Flickr API for others
- Average number of masks per image is fairly consistent across region and income

		SA-1B		% images		
	# countries	#imgs	#masks	SA-1B	COCO	O.I.
Africa	54	300k	28M	2.8%	3.0%	1.7%
Asia & Oceania	70	3.9M	423M	36.2%	11.4%	14.3%
Europe	47	5.4M	540M	49.8%	34.2%	36.2%
Latin America & Carib.	42	380k	36M	3.5%	3.1%	5.0%
North America	4	830k	80M	7.7%	48.3%	42.8%
high income countries	81	5.8M	598M	54.0%	89.1%	87.5%
middle income countries	108	4.9M	499M	45.0%	10.5%	12.0%
low income countries	28	100k	9.4M	0.9%	0.4%	0.5%



RAI

- Simulated interactive segmentation with random sampling of 1&3 point(s)
- Which groups are underrepresented in the training data?
- Is the model quality consistent on these groups?

	mIoU at 1 point	mIoU at 3 points
perceived gender presentation		
feminine	54.4 ±1.7	90.4 ±0.6
masculine	55.7 ±1.7	90.1 ±0.6
perceived age group		
older	62.9 ±6.7	92.6 ±1.3
middle	54.5 ±1.3	90.2 ±0.5
young	54.2 ±2.2	91.2 ±0.7

	mIoU at 1 point	mIoU at 3 points
perceived skin tone		
1	52.9 ±2.2	91.0 ±0.9
2	51.5 ±1.4	91.1 ±0.5
3	52.2 ±1.9	91.4 ±0.7
4	51.5 ±2.7	91.7 ±1.0
5	52.4 ±4.2	92.5 ±1.4
6	56.7 ±6.3	91.2 ±2.4

People segmentation

	mIoU at 1 point	mIoU at 3 points
perceived gender presentation		
feminine	76.3 ±1.1	90.7 ±0.5
masculine	81.0 ±1.2	92.3 ±0.4

	mIoU at 1 point	mIoU at 3 points
perceived age group		
older	81.9 ±3.8	92.8 ±1.6
middle	78.2 ±0.8	91.3 ±0.3
young	77.3 ±2.7	91.5 ±0.9

Clothes segmentation

* 1 is lightest skin tone, 6 is darkest one

Zero-shot experiments

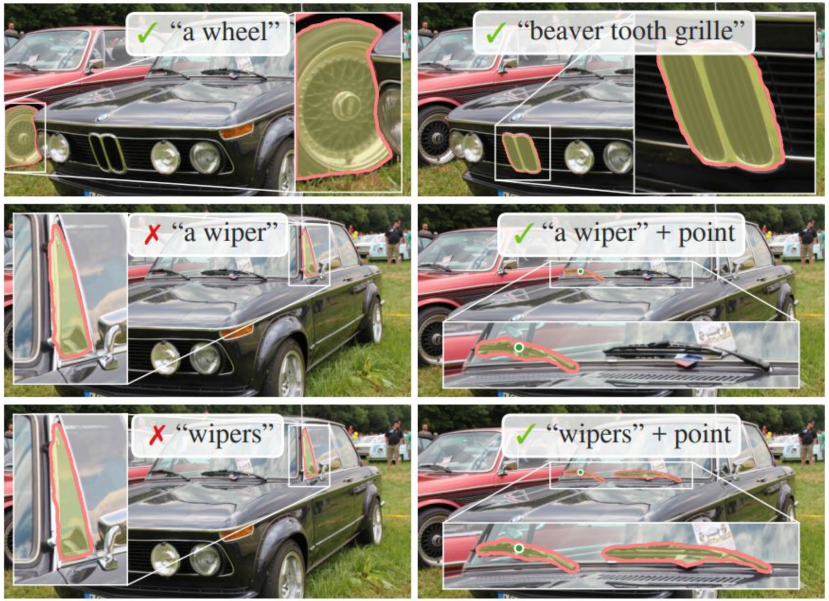
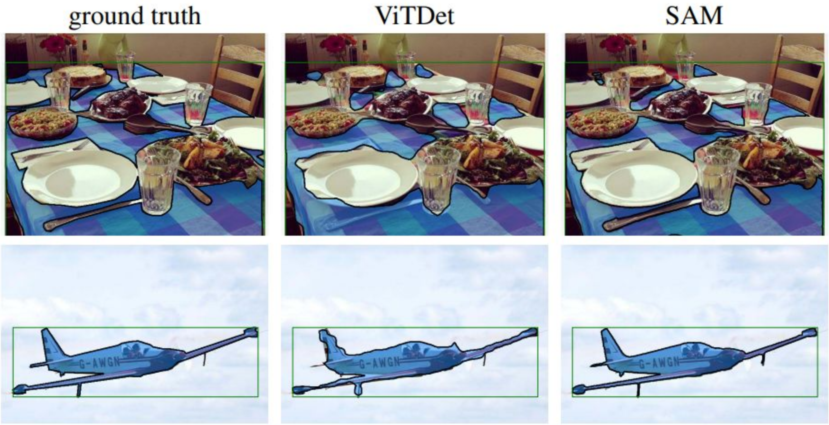
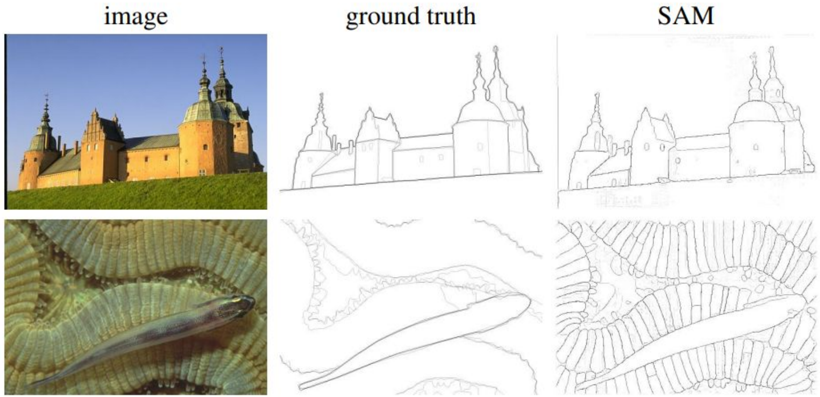


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.





Key points

- Promptable Foundation Model for Semantic Segmentation
- 3-stage Data Engine pipeline
- Public high-quality dataset
- Responsible AI research

Disadvantages of Segment Anything Model (SAM)

- SAM can miss fine structures, hallucinates small disconnected components at times
- SAM does not produce crisply boundaries as “zoom-in” methods
- SAM could be outperformed by dedicated interactive segmentation methods when many points are provided
- Real-time performance is per prompt, not per image
- Not domain specific
- Unclear how to design simple prompts that implement semantic and panoptic segmentation