Segment Anything

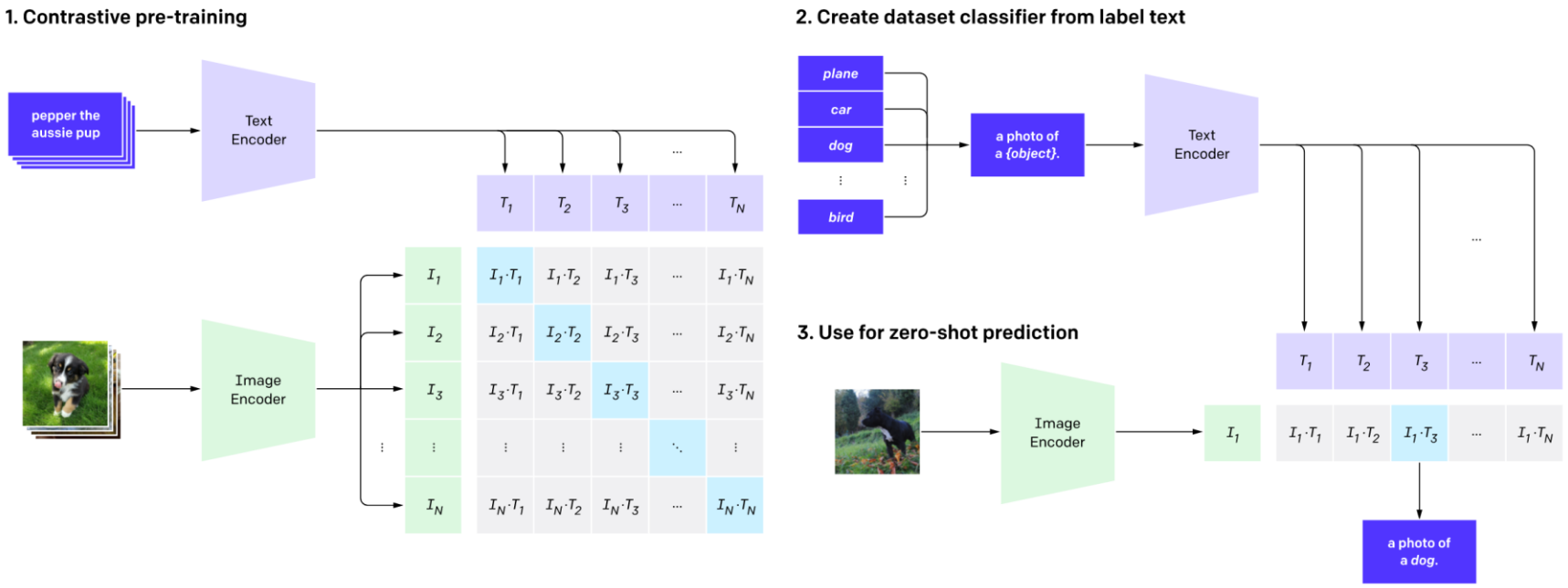
Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR

arxiv

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

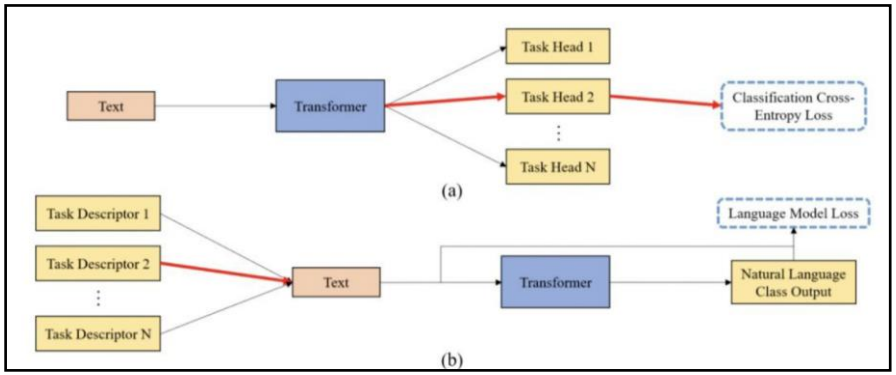
研究背景

- CLIP实现了视觉-语言对齐，使得视觉任务可以扩展到open set

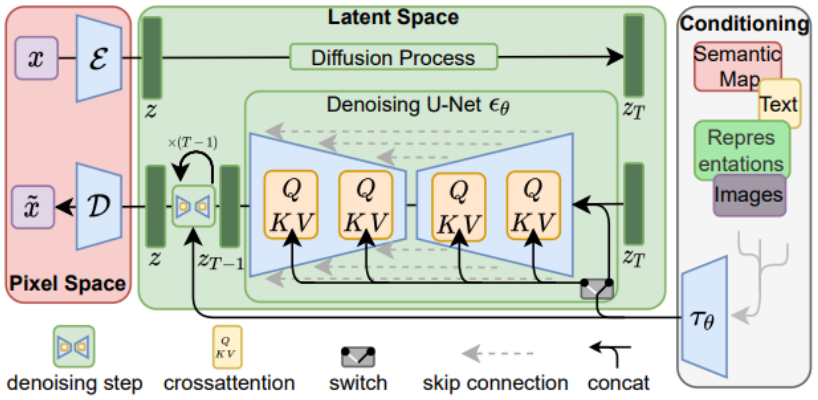


研究背景

- Prompt learning



Prompt learning in NLP



Text to image

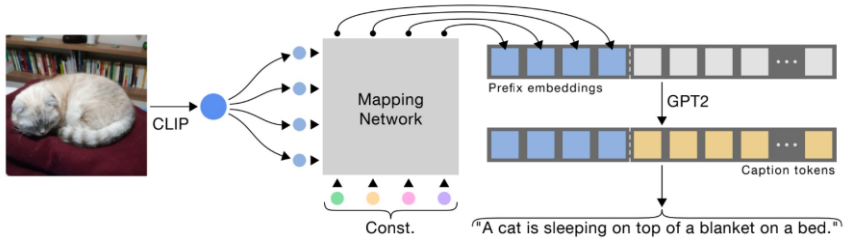
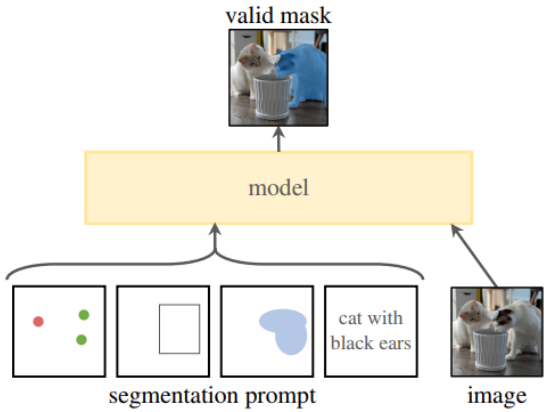


Image Caption

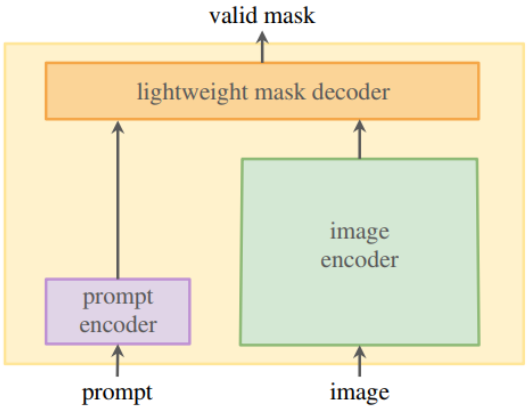
Visual Grounding

研究动机

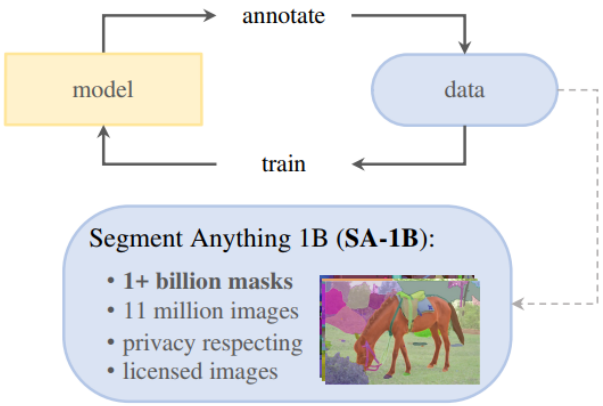
- Promptable segmentation task
- Large scale dataset



(a) **Task**: promptable segmentation



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



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

Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a promptable segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

Architecture

- Image encoder: ViT(pre-trained MAE)
- prompt:
  - Dense prompt: mask → conv encoder
  - Sparse prompt: points, box, text → prompt encoder
- Mask decoder: cross-attention

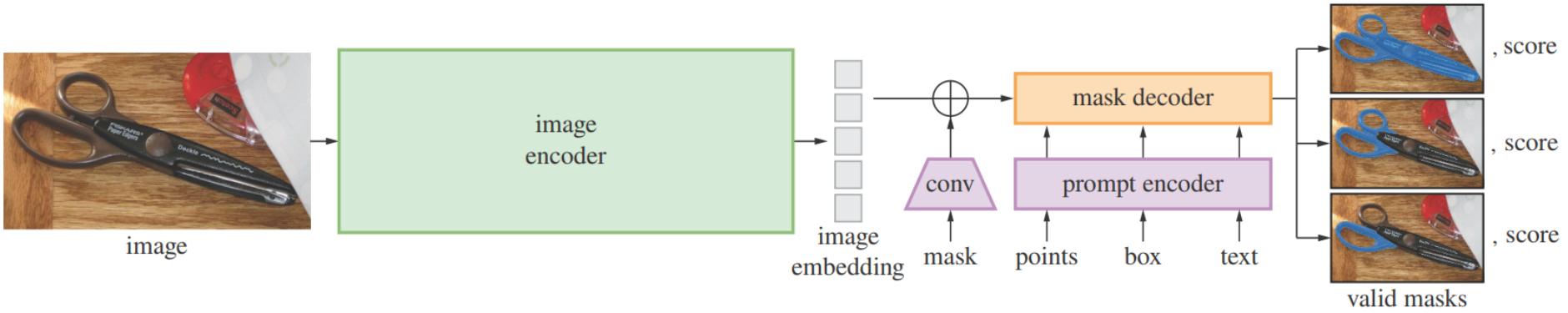
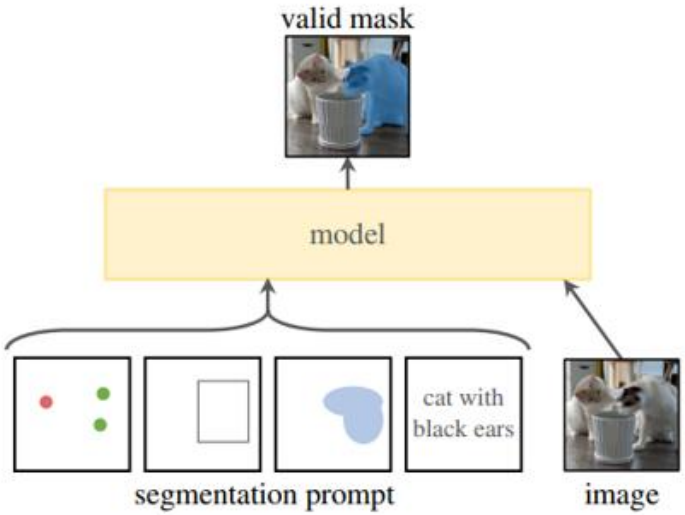


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

Prompt encoder

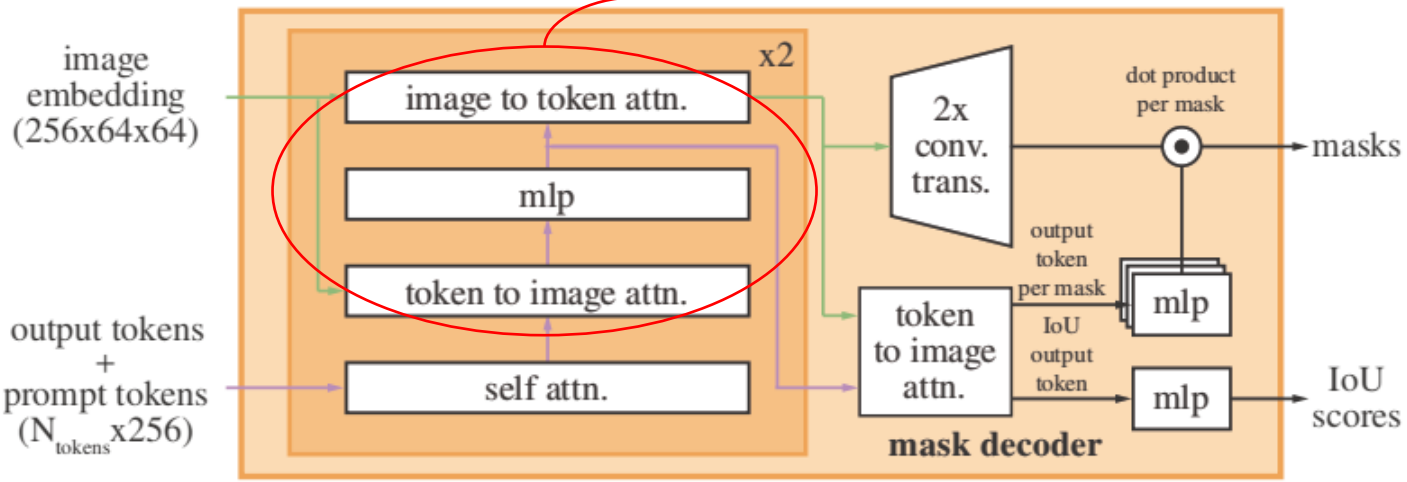
- Sparse prompt
  - Point：sine position encoding
  - Bbox： sine position encoding
  - Text：CLIP text encoder
  - **[mask token, prompt]，token拼接**
- Dense prompt：
  - Mask：conv layer
  - **image embed + prompt，逐像素相加**



Mask

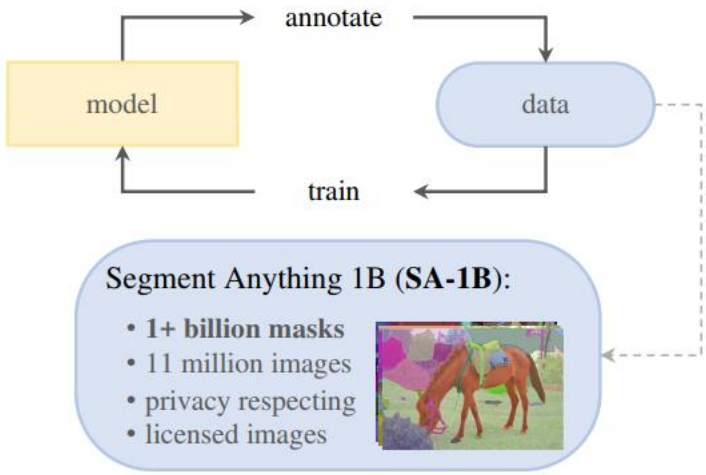
- Two-w
  - tok
  - Ima
  - Mas

```
# Cross attention block, tokens attending to image embedding
q = queries + query_pe
k = keys + key_pe
attn_out = self.cross_attn_token_to_image(q=q, k=k, v=keys)
queries = queries + attn_out
queries = self.norm2(queries)

# MLP block
mlp_out = self.mlp(queries)
queries = queries + mlp_out
queries = self.norm3(queries)

# Cross attention block, image embedding attending to tokens
q = queries + query_pe
k = keys + key_pe
attn_out = self.cross_attn_image_to_token(q=k, k=q, v=queries)
keys = keys + attn_out
keys = self.norm4(keys)
```



Data engine

- 1 billion masks（400x previous datasets）



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

- *Stage1：assisted-manual*
  - 使用公共数据集训练SAM
  - 人工修改mask，用新数据重新训练
  - 循环6次，**4.3M masks**
- *Stage2：semi-automatic*
  - 找出置信度低的mask（在mask上做目标检测）
  - 人工修改mask，用新数据重新训练
  - 循环5次，**5.9M masks**
- *Stage3：fully automatic*
  - 过滤置信度低的mask
  - 过滤不稳定的mask（略微改变threshold，mask变化应该不大）
  - **1.1B masks，11M images**

dataset

- 三阶段标注数据vs自动标注数据。SA-1B只包含自动标注的数据
- 10% SA-1B 的训练效果与全量数据接近
- 扩大Image encoder

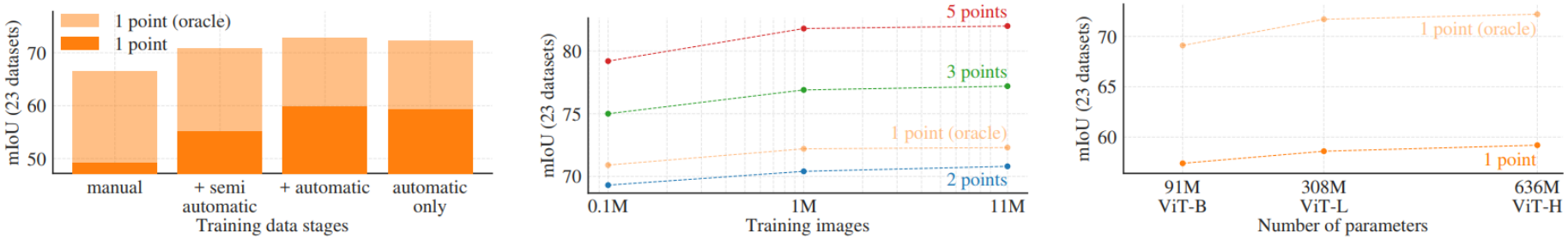


Figure 13: Ablation studies of our data engine stages, image encoder scaling, and training data scaling. (Left) Each data engine stage leads to improvements on our 23 dataset suite, and training with only the automatic data (our default) yields similar results to using data from all three stages. (Middle) SAM trained with ~10% of SA-1B and full SA-1B is comparable. We train with all 11M images by default, but using 1M images is a reasonable practical setting. (Right) Scaling SAM's image encoder shows meaningful, yet saturating gains. Nevertheless, smaller image encoders may be preferred in certain settings.

dataset

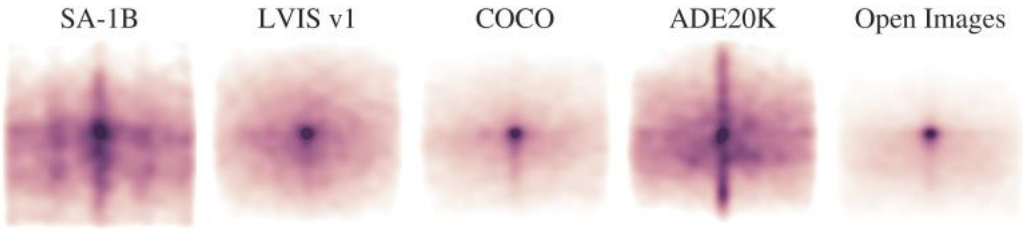


Figure 5: Image-size normalized mask center distributions.

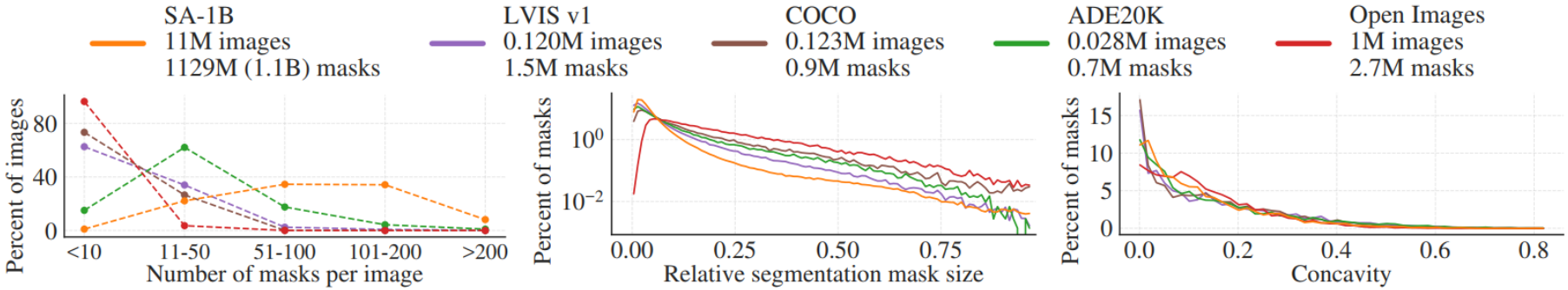


Figure 6: Dataset mask properties. The legend references the number of images and masks in each dataset. Note, that SA-1B has 11× more images and 400× more masks than the largest existing segmentation dataset Open Images [60].

	# countries	SA-1B #imgs	SA-1B #masks	% images SA-1B	% images COCO	% images 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%

Table 1: Comparison of geographic and income representation. SA-1B has higher representation in Europe and Asia &

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

- 稀疏

- 密集

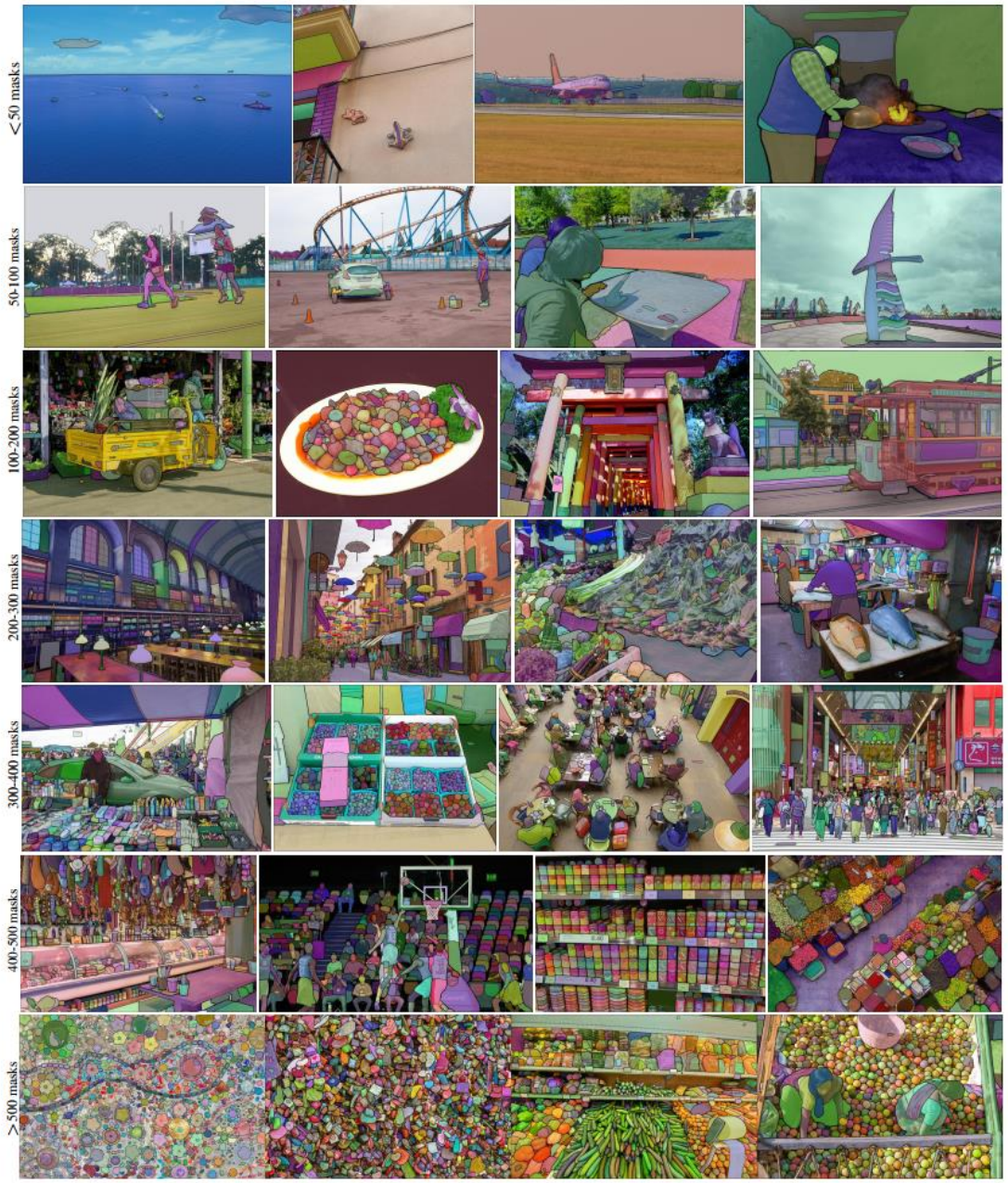


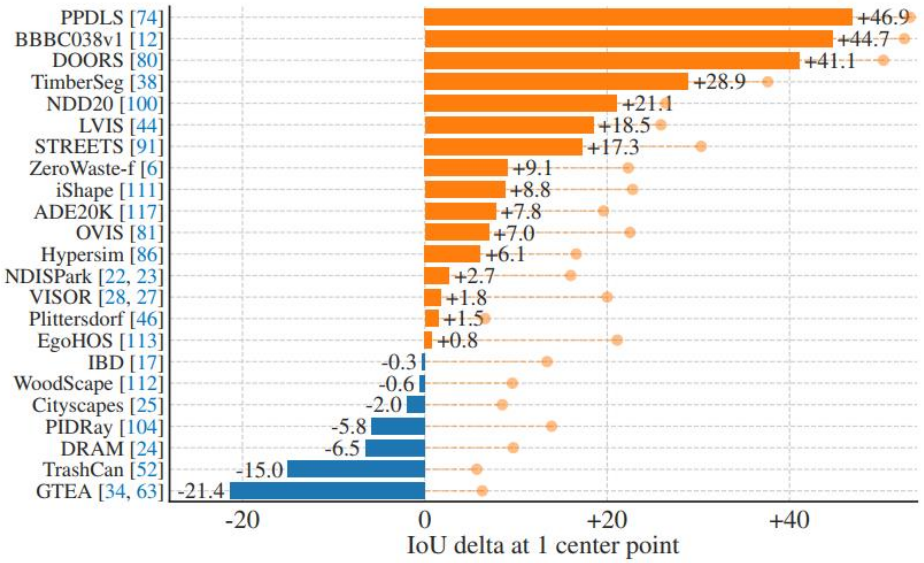
Figure 2: Example images with overlaid masks from our newly introduced dataset, **SA-1B**. SA-1B contains 11M diverse, high-resolution, licensed, and privacy protecting images and 1.1B high-quality segmentation masks. These masks were annotated *fully automatically* by SAM, and as we verify by human ratings and numerous experiments, are of high quality and

Prompt Ambiguity

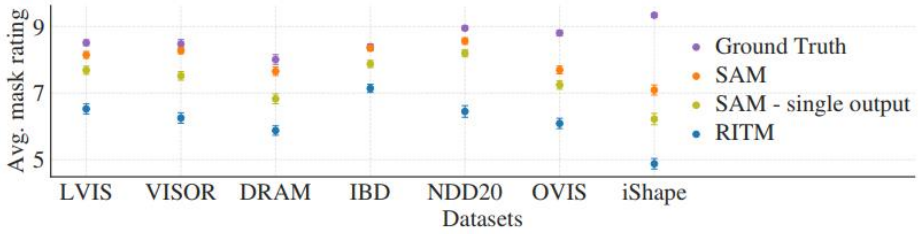
- 对于单个提示，存在歧义性。因此预测3个mask。

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

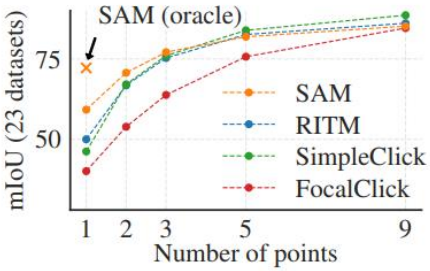
Point to mask



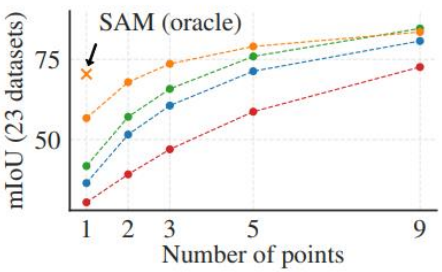
(a) SAM *vs.* RITM [92] on 23 datasets



(b) Mask quality ratings by human annotators



(c) Center points (default)



(d) Random points

Figure 9: Point to mask evaluation on 23 datasets. (a) Mean IoU of SAM and the strongest single point segmenter, RITM [92]. Due to ambiguity, a single mask may not match ground truth; circles show "oracle" results of the most relevant of SAM's 3 predictions. (b) Per-dataset comparison of mask quality ratings by annotators from 1 (worst) to 10 (best). All methods use the ground truth mask center as the prompt. (c, d) mIoU with varying number of points. SAM significantly outperforms prior interactive segmenters with 1 point and is on par with more points. Low absolute mIoU at 1 point is the result of ambiguity.

Box to mask

	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

Table 5: Instance segmentation results. SAM is prompted with ViTDet boxes to do zero-shot segmentation. The fully-supervised ViTDet outperforms SAM, but the gap shrinks on the higher-quality LVIS masks. Interestingly, SAM outperforms ViTDet according to human ratings (see Fig. 11).

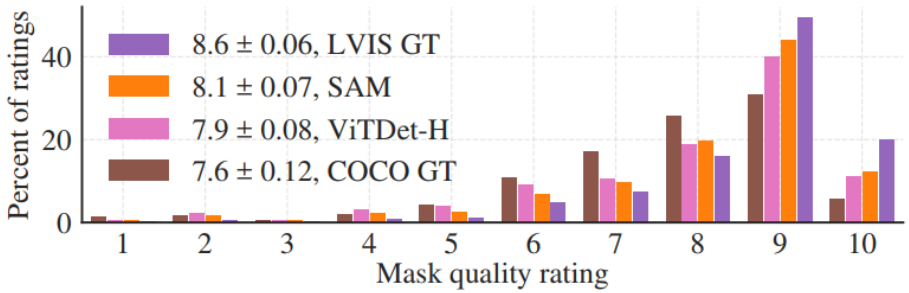


Figure 11: Mask quality rating distribution from our human

Text to mask

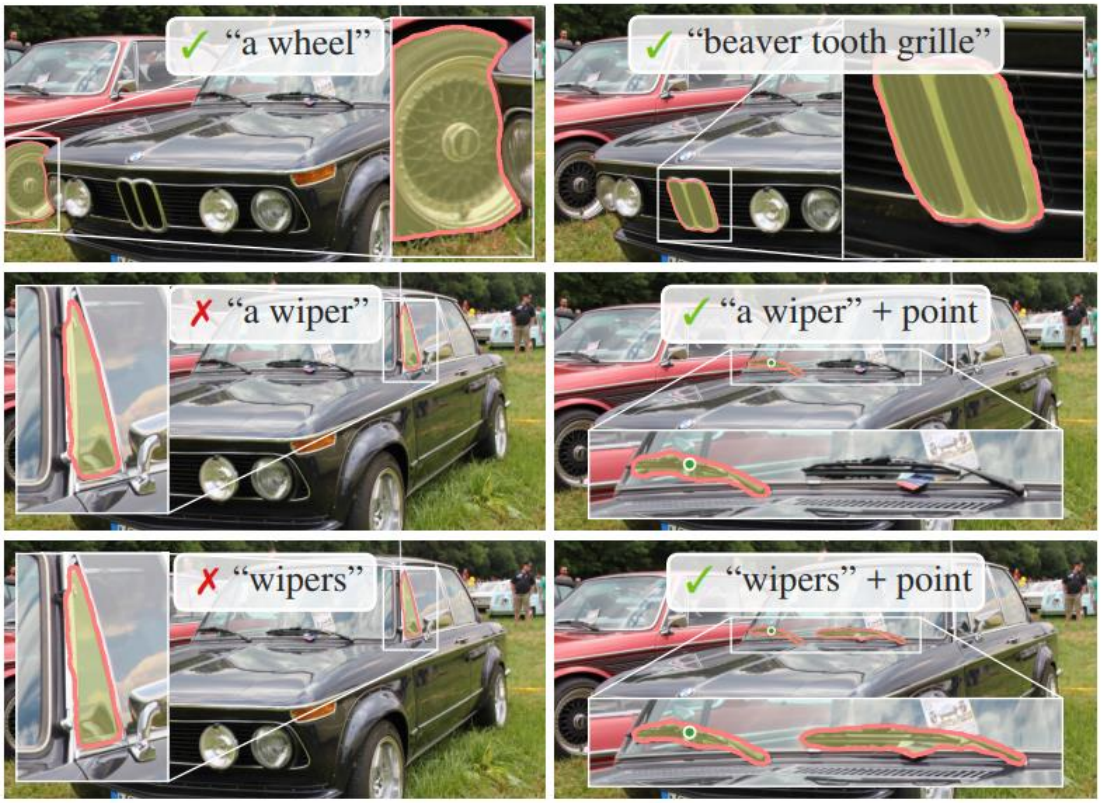


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Edge detection

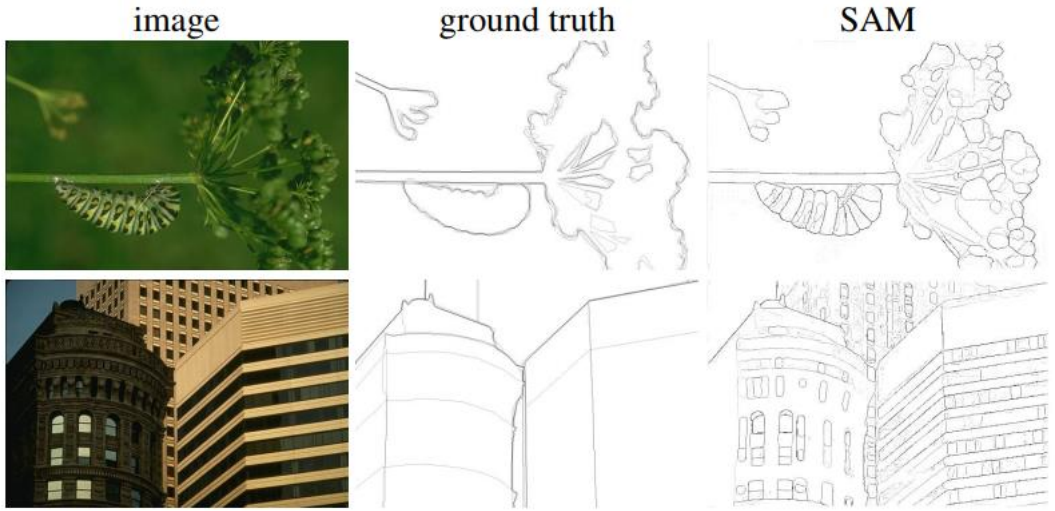


Figure 10: Zero-shot edge prediction on BSDS500. SAM was not trained to predict edge maps nor did it have access to BSDS images or annotations during training.

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.

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

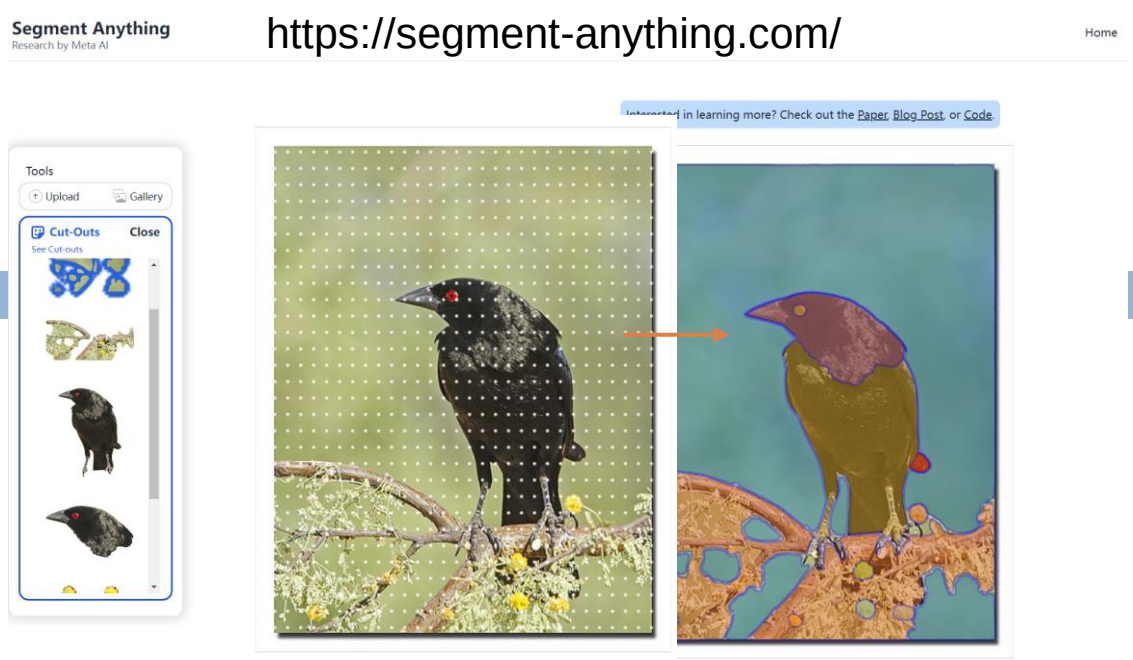
总结

- 通用的提示分割框架，泛化性好
- 人类-模型协同标注，渐进式训练
- 没有开放text prompt，暂时还不能直接分类

- 我们虽然很难训练大模型，但是应该用起来
  - 用来（精细）标注数据
  - foundation model + prompt的范式
  - Text-to-mask

多部位分割

- demo



```
print(len(masks))
print(masks[0].keys())
```

```
65
dict_keys(['segmentation', 'area', 'bbox', 'predicted_iou', 'point_coords', 'stability_score', 'crop_box'])
```

Text encoder没有开源

- 需要与其他模型配合实现text-to-mask分割
- IDEA做了demo，把SAM与其他模型组合使用

Grounded-SAM + Stable-Diffusion Inpainting: Data-Factory, Generating New Data!

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

Thanks!