Segment Anything

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson,
Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, Ross Girshick

STRUCT Group Seminar
Presenter: Yexiang Cheng
2023.04.16

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

BACKGROUND: Transformer

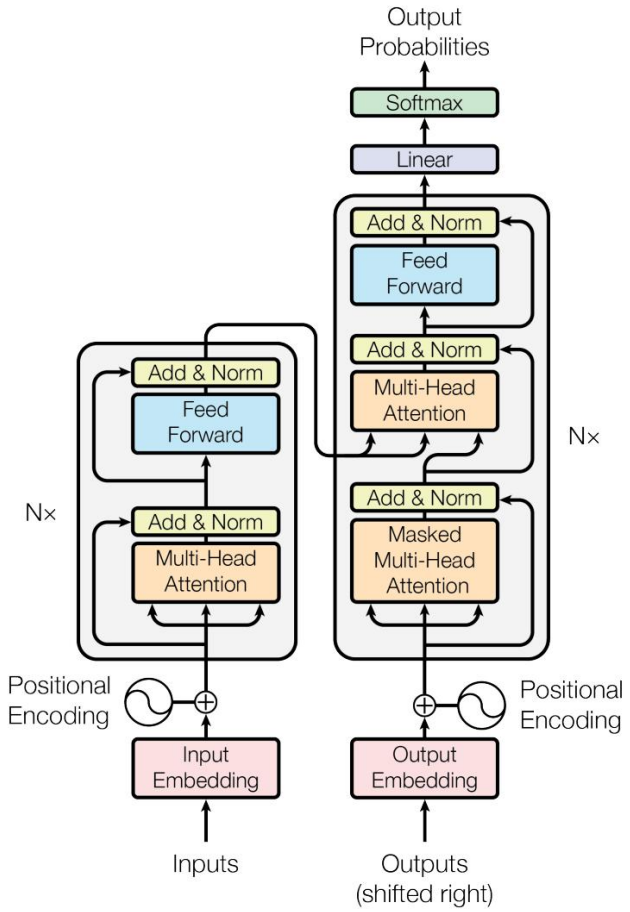
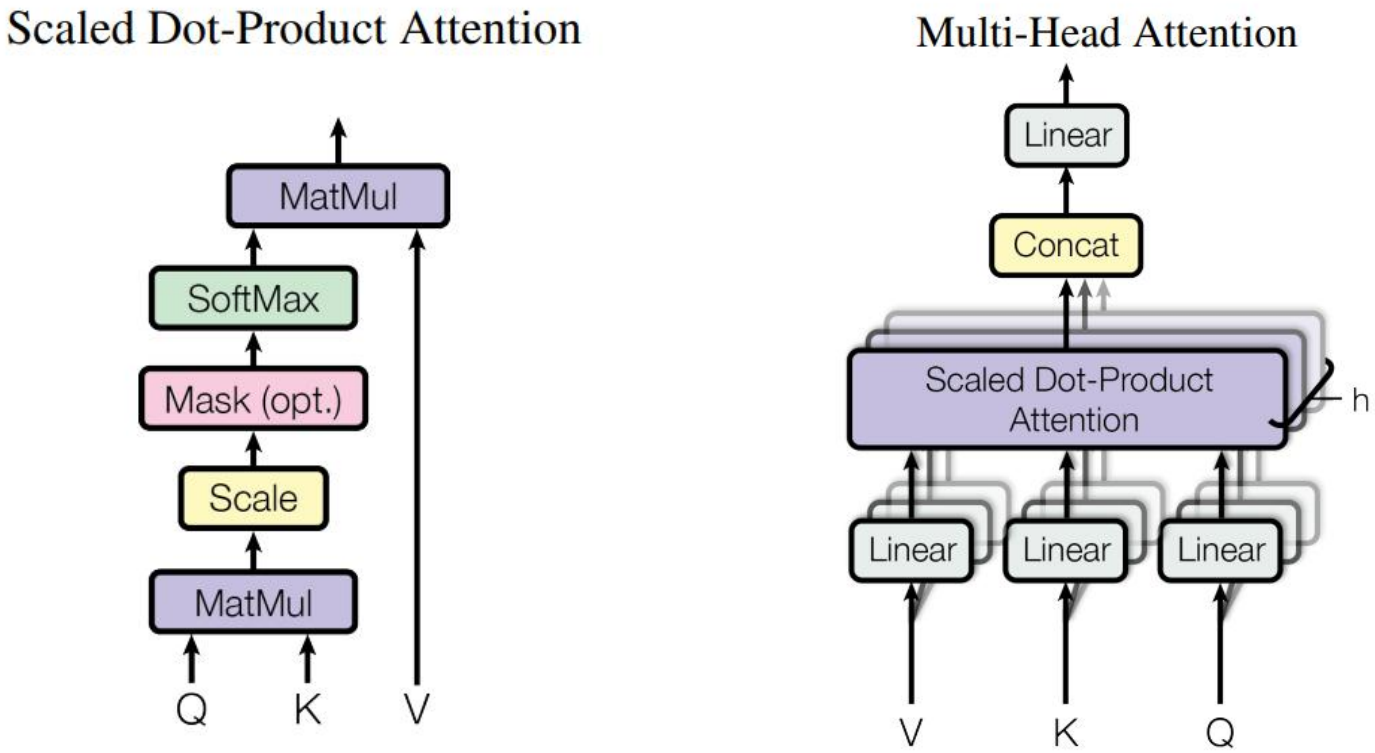


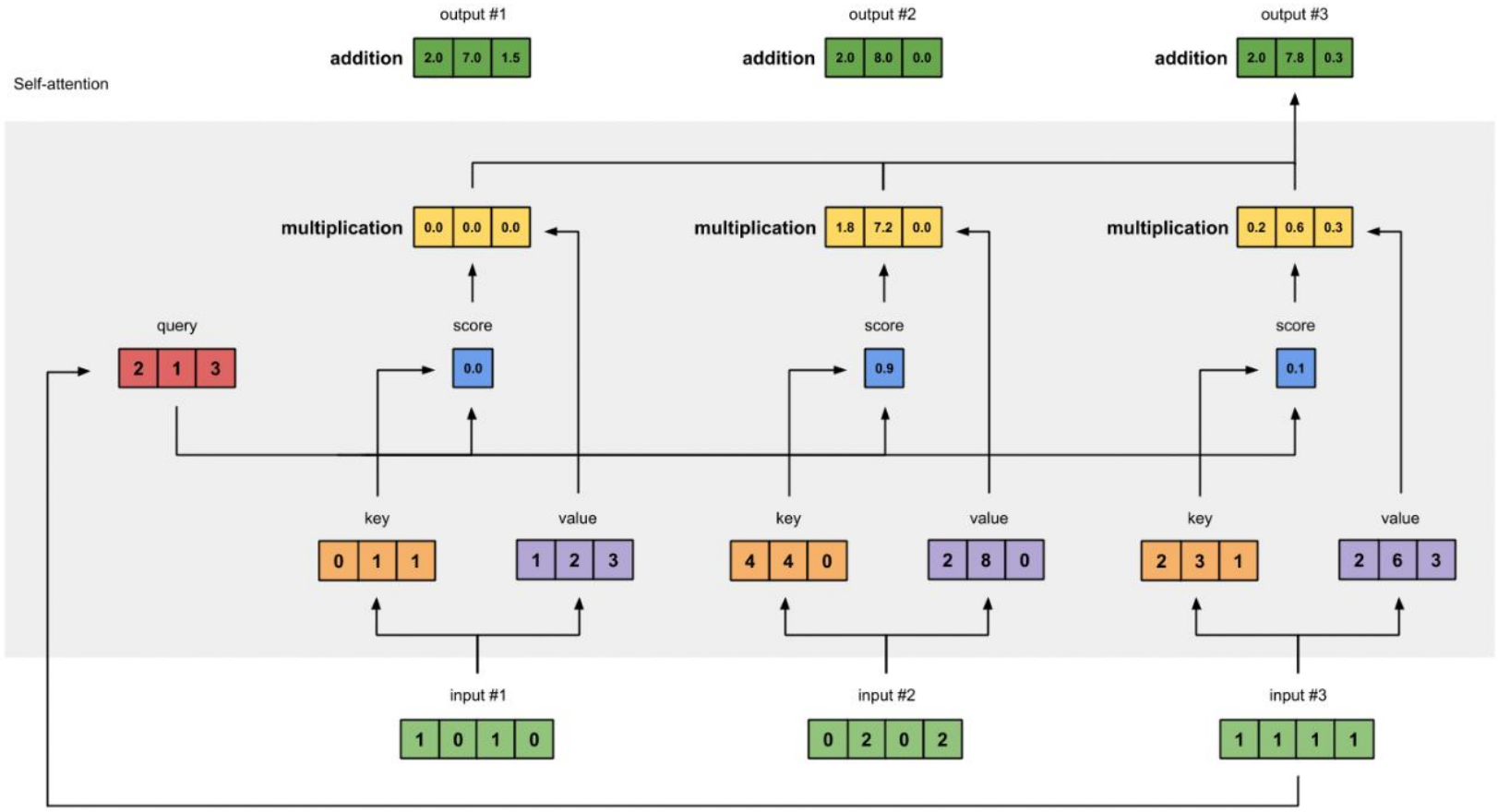
Figure 1: The Transformer - model architecture.

BACKGROUND: Transformer



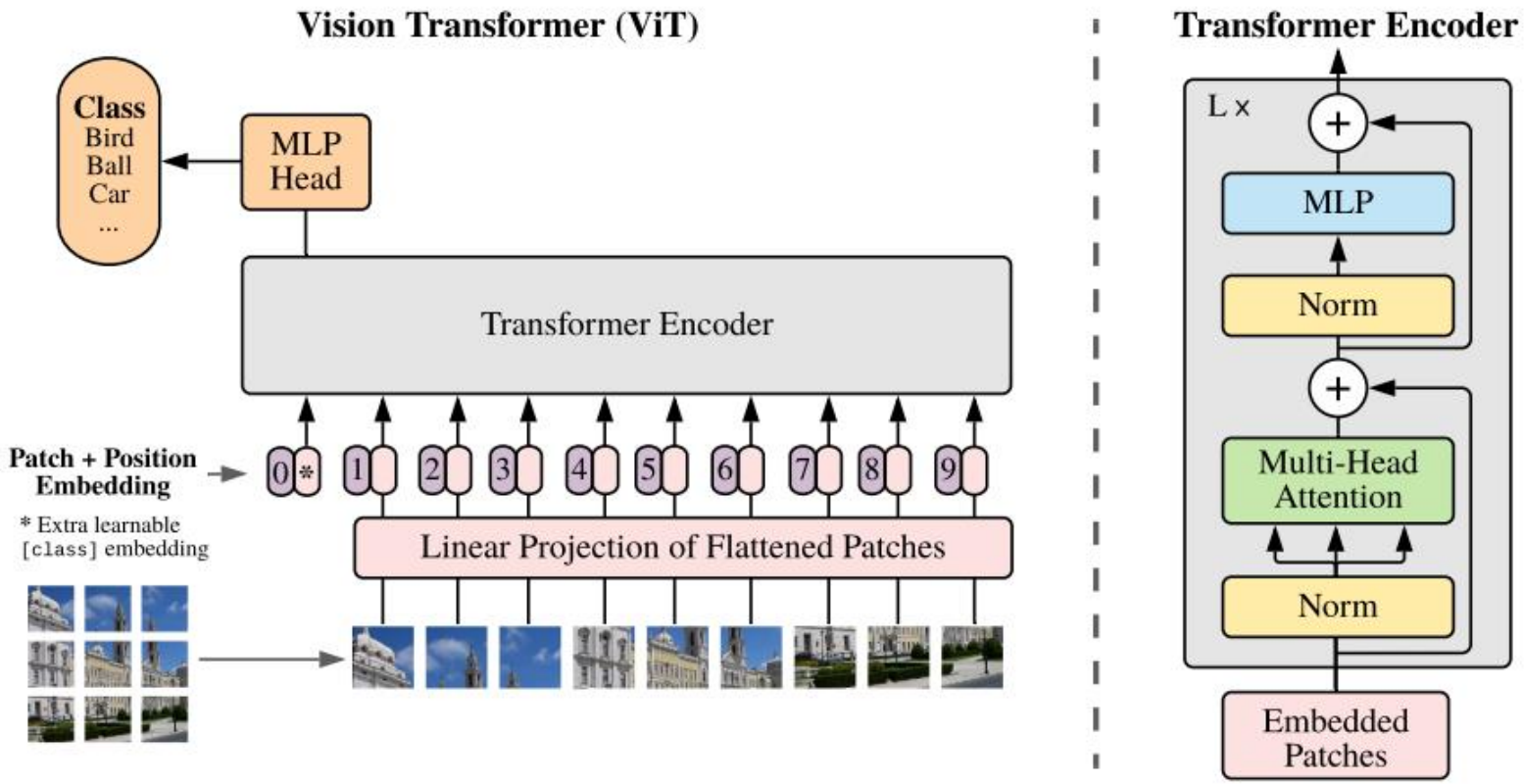
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

BACKGROUND: Transformer

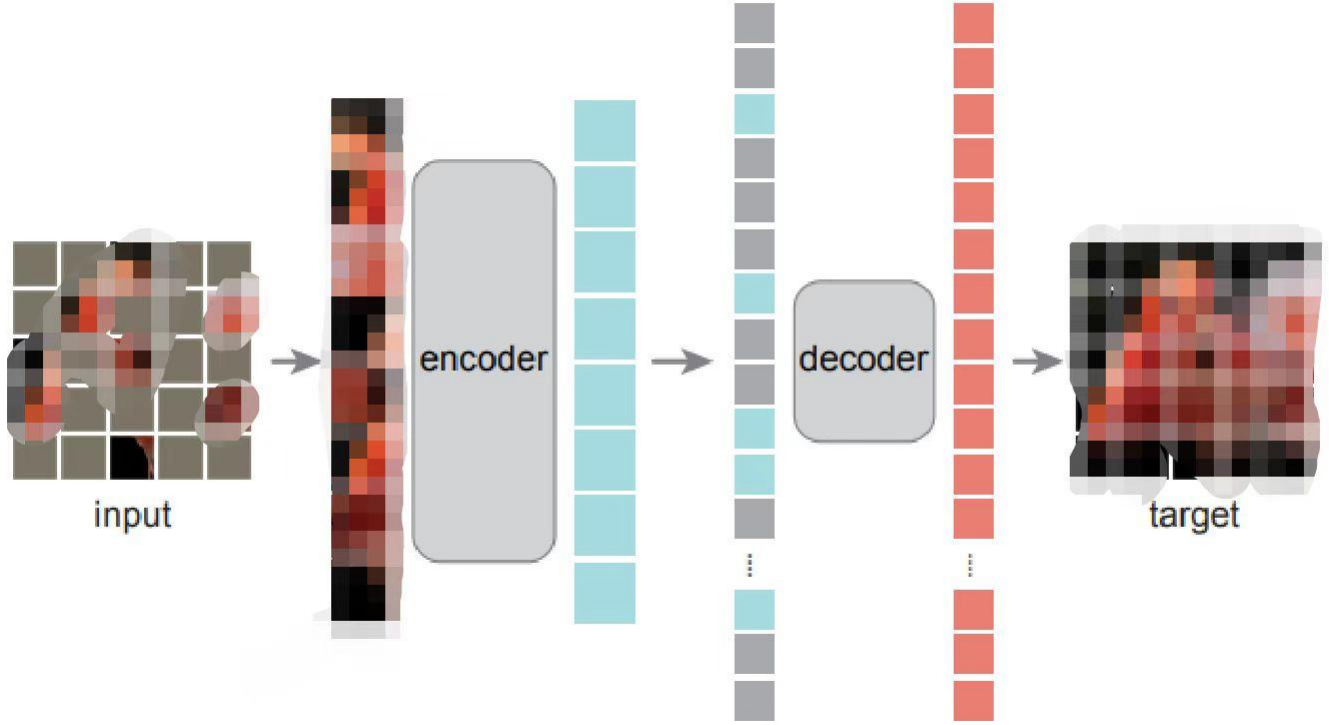


From [Illustrated: Self-Attention. A step-by-step guide to self-attention… | by Raimi Karim | Towards Data Science]

BACKGROUND: Vision Transformer



BACKGROUND: Masked Autoencoder



BACKGROUND: Prompt engineering

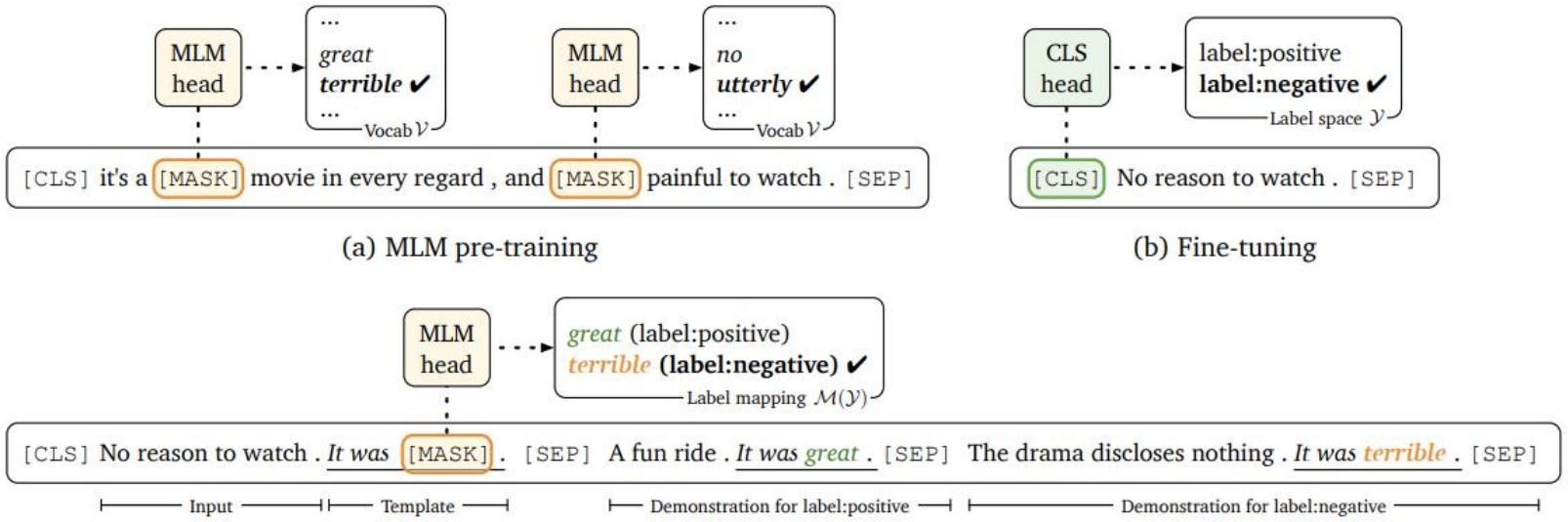
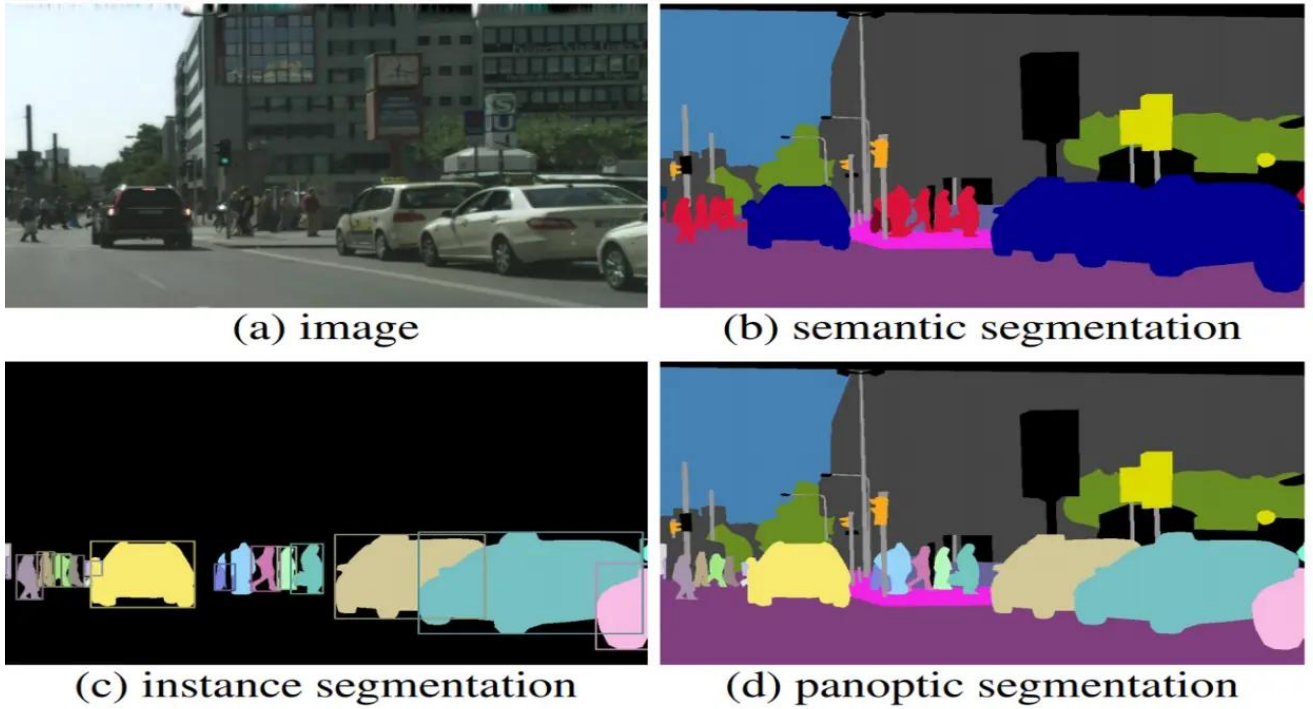


Figure 1: An illustration of (a) masked language model (MLM) pre-training, (b) standard fine-tuning, and (c) our proposed LM-BFF using prompt-based fine-tuning with demonstrations. The underlined text is the task-specific *template*, and colored words are *label words*.

BACKGROUND: Image Segmentation

The process of partitioning a digital image into multiple image segments

- Semantic segmentation
- Instance segmentation
- Panoptic segmentation

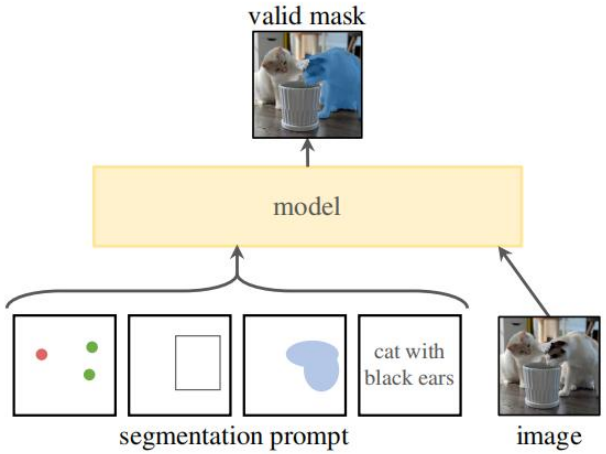


OUTLINE

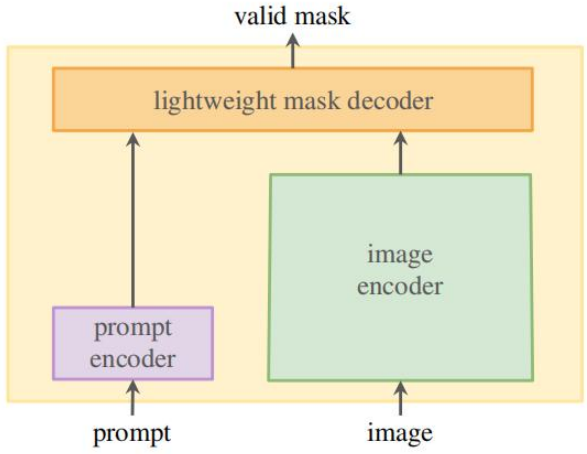
- Authorship
- Background
- Method
- Experiments
- Conclusion

METHOD

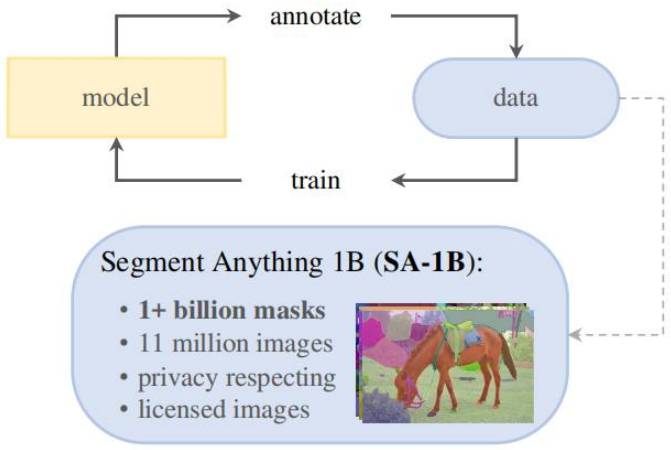
Overview



(a) **Task**: promptable segmentation



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



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

METHOD

Overview

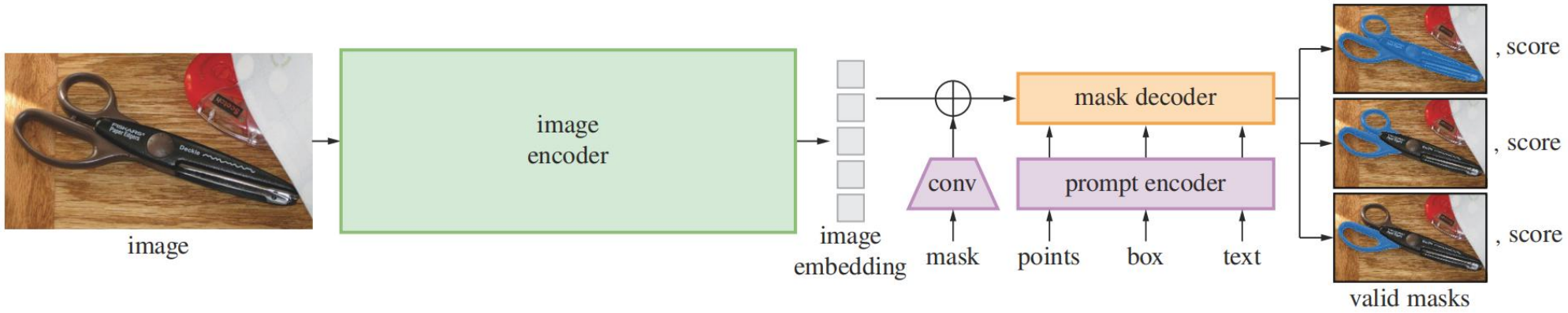


Image encoder: MAE pre-trained ViT
Prompt encoder:
(1) Points and boxes: positional encodings + learned embeddings for each prompt type
(2) Text: text encoder from CLIP
(3) Dense prompts(such as masks): embedded using convolutions
Mask decoder: A modification of a Transformer decode block followed by a dynamic mask prediction head

METHOD

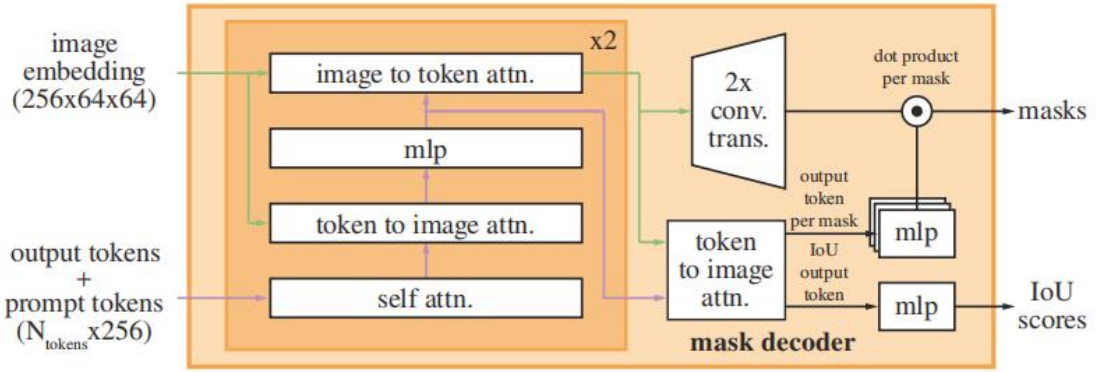


Figure 14: Details of the lightweight mask decoder. A two-layer decoder updates both the image embedding and prompt tokens via cross-attention. Then the image embedding is upscaled, from which the updated output tokens are used to dynamically predict masks. (Not illustrated for figure clarity: At every attention layer, positional encodings are added to the image embedding, and the entire original prompt token (including position encoding) is re-added to the token queries and keys.)

METHOD

Dataset Engine

- Assisted-manual stage

  A team of professional annotators do interactive segmentation.

- Semi-automatic stage

  Annotators annotate the objections that were not detected by the model.

- Fully automatic stage

  Model generates masks without annotator input.

METHOD

Responsible AI analysis

	# 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%

	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

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

EXPERIMENTS

Zero-Shot Single Point Valid Mask Evaluation Task

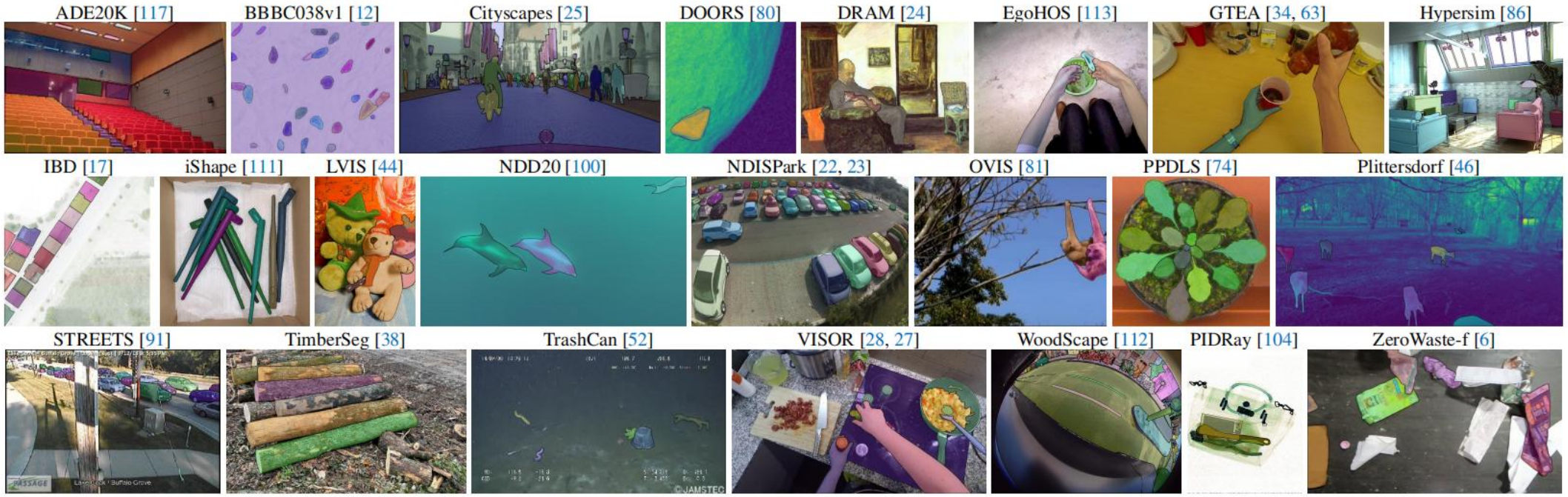
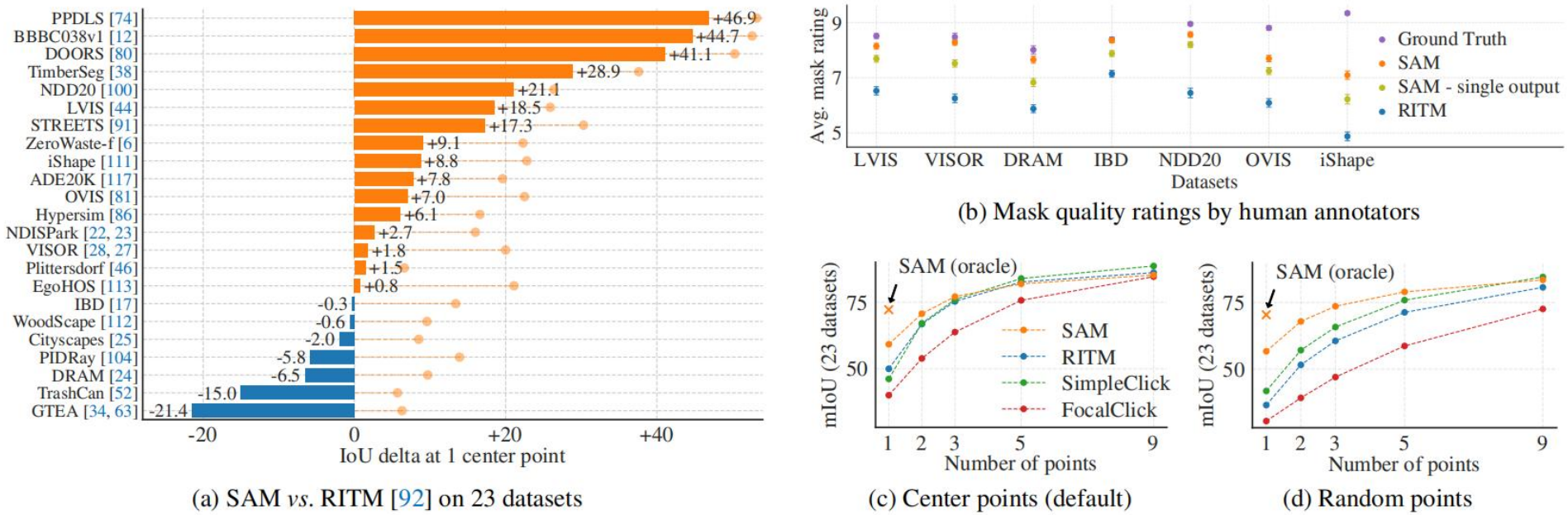


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.

EXPERIMENTS

Zero-Shot Single Point Valid Mask Evaluation Task



(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.

EXPERIMENTS

Zero-Shot 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.

EXPERIMENTS

Zero-Shot Object Proposals

	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.

EXPERIMENTS

Zero-Shot Instance Segmentation

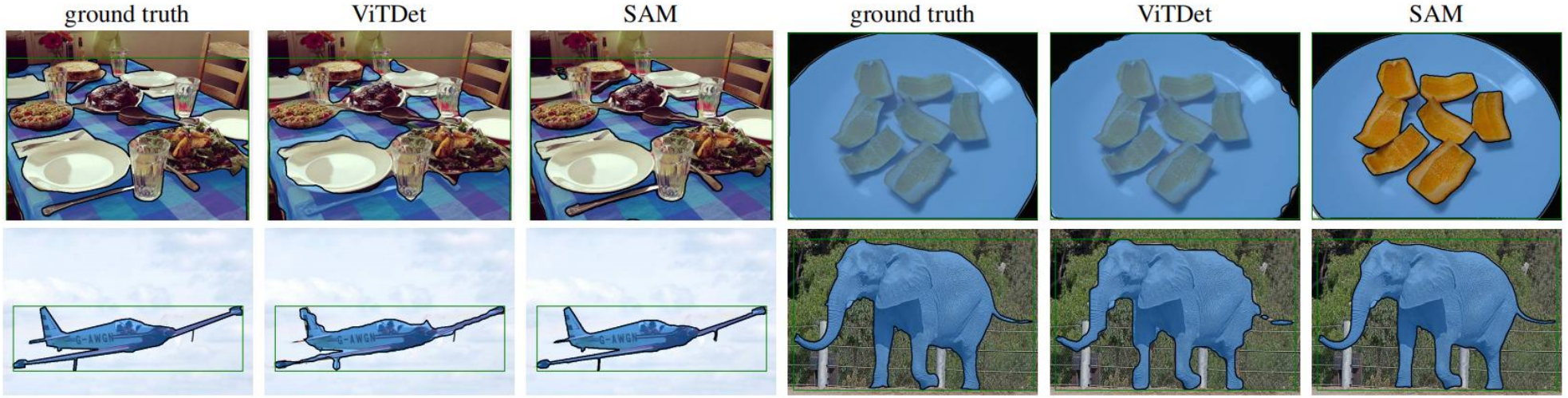


Figure 16: Zero-shot instance segmentation on LVIS v1. SAM produces higher quality masks than ViTDet. As a zero-shot model, SAM does not have the opportunity to learn specific training data biases; see top-right as an example where SAM makes a modal prediction, whereas the ground truth in LVIS is amodal given that mask annotations in LVIS have no holes.

EXPERIMENTS

Zero-Shot Instance Segmentation

	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).

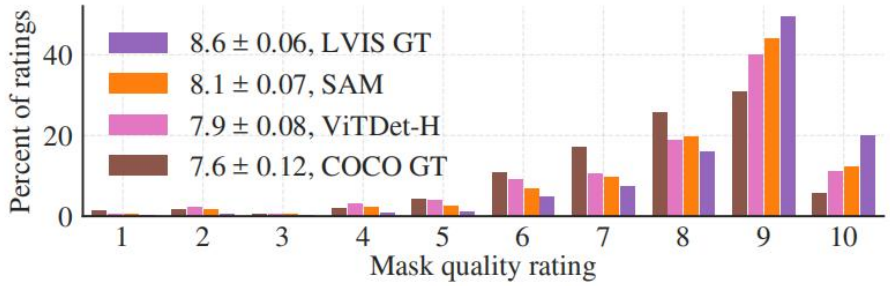


Figure 11: Mask quality rating distribution from our human study for ViTDet and SAM, both applied to LVIS ground truth boxes. We also report LVIS and COCO ground truth quality. The legend shows rating means and 95% confidence intervals. Despite its lower AP (Table 5), SAM has higher ratings than ViTDet, suggesting that ViTDet exploits biases in the COCO and LVIS training data.

EXPERIMENTS

Zero-Shot Text-to-Mask

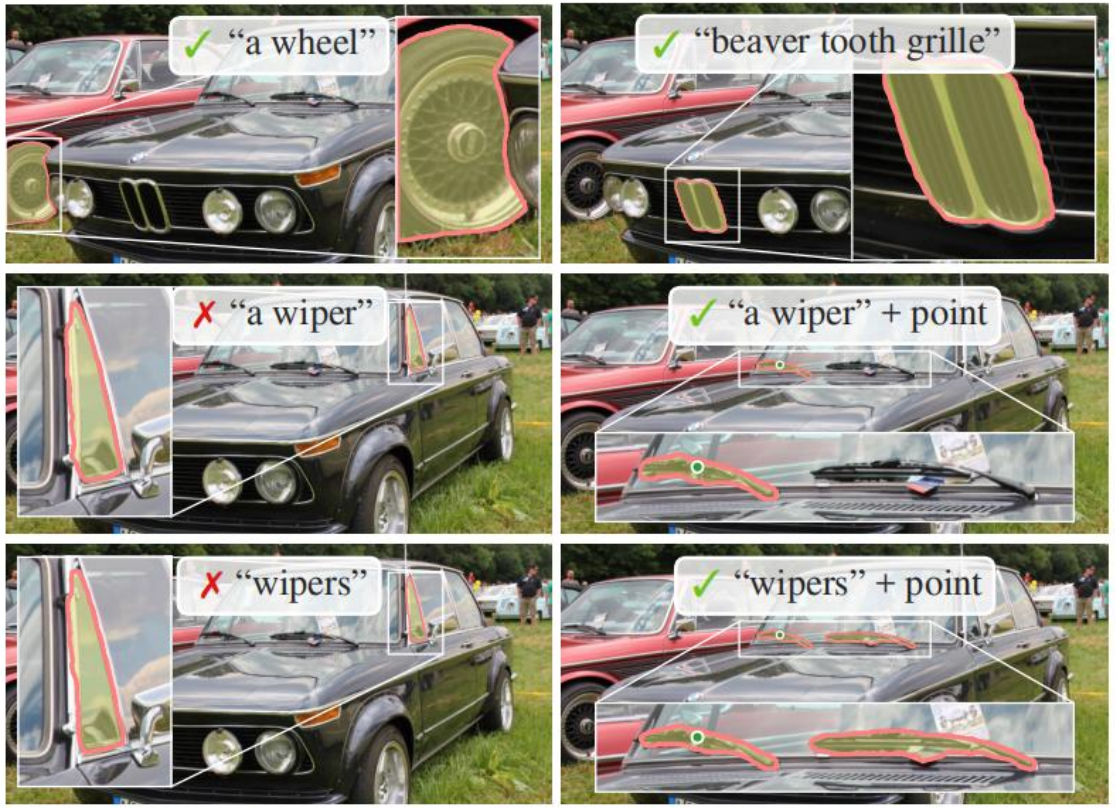


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

EXPERIMENTS

Ablations

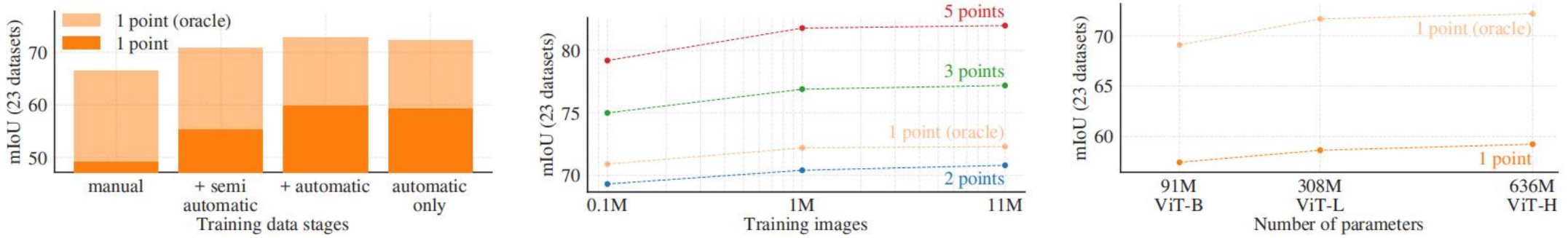


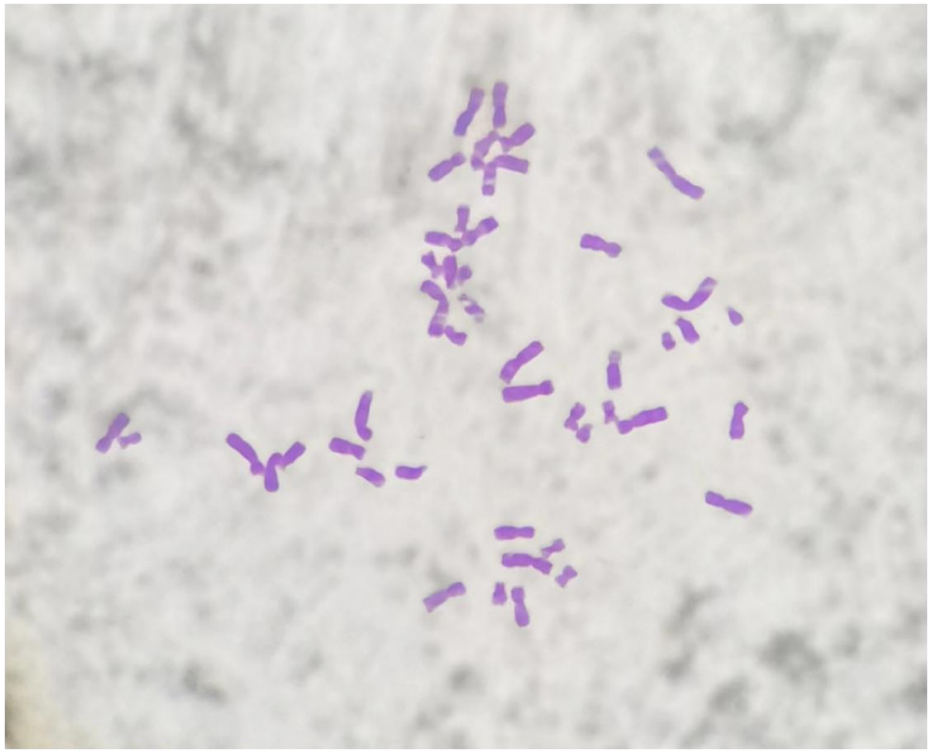
Figure 13: Ablation studies of our data engine stages, image encoder scaling, and training data scaling. (Left) Each data engine stage leads to improvements on our 23 dataset suite, and training with only the automatic data (our default) yields similar results to using data from all three stages. (Middle) 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.

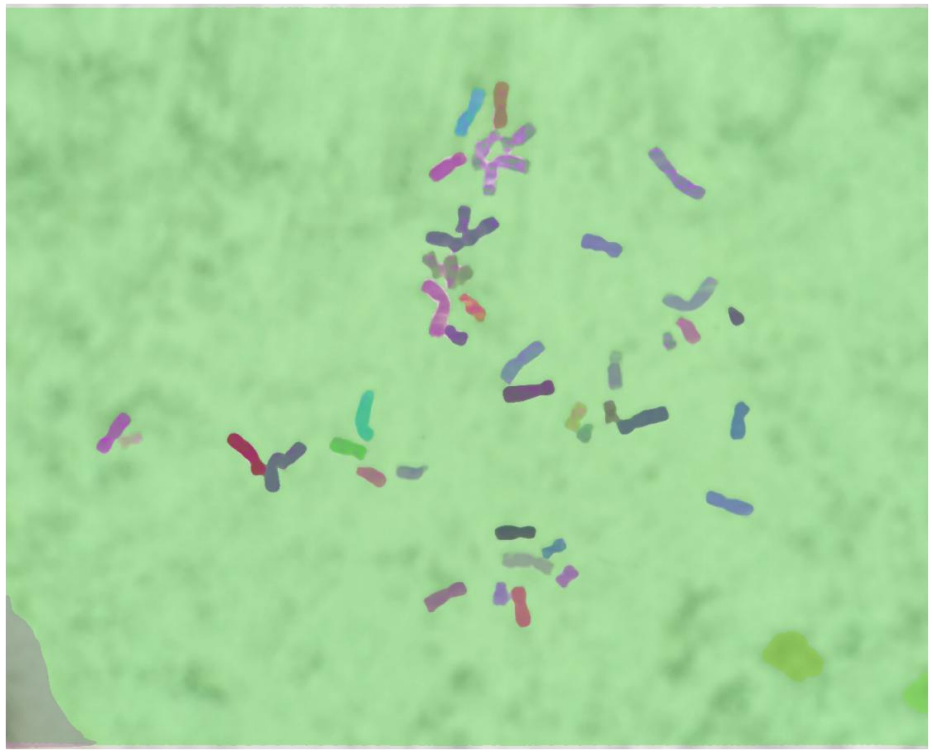
EXPERIMENTS

Other attempts

EXPERIMENTS

Other attempts





EXPERIMENTS

Other attempts

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

EXPERIMENTS

Other attempts

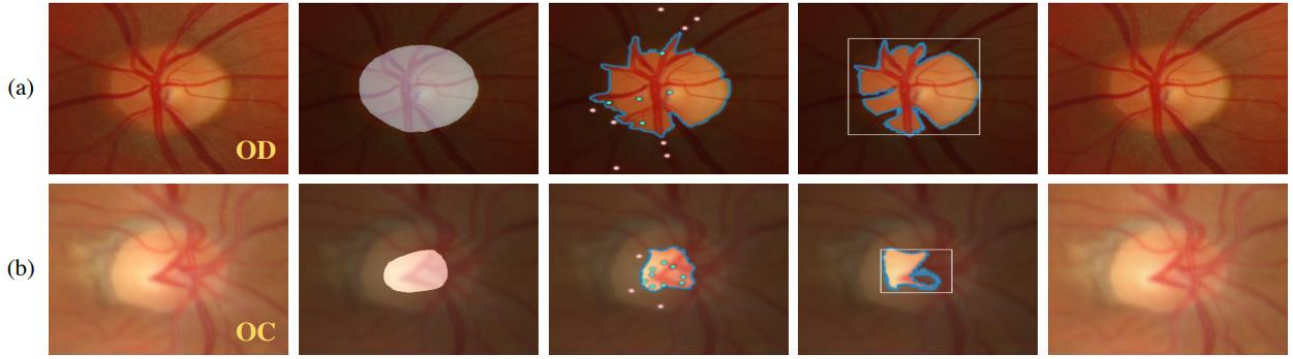


Figure 12. Application on **joint optical disc (OD) and optical cup (OC) segmentation**, where SAM$^{1/2/3}$ mean using Click, Box, and Everything modes respectively. Here SAM^{3} does not generate any results on these cases.

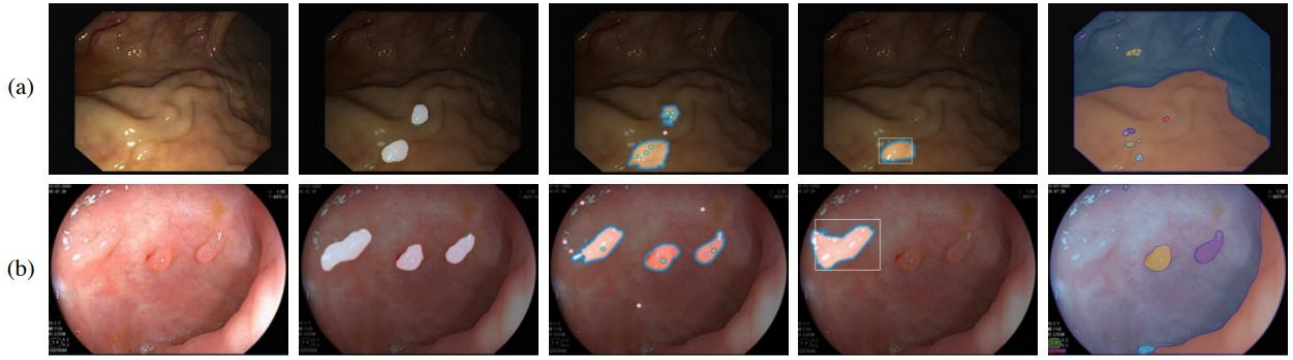


Figure 13. Application on **polyp segmentation**, where SAM$^{1/2/3}$ mean using Click, Box, and Everything modes respectively. The * indicates the SAM results within a box prompt.

From arXiv:2304.05750

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

CONCLUSION

- Lift image segmentation into the era of foundation models
- A new task: promptable segmentation
- A new dataset: SA-1B
- Wide range of downstream tasks

Thanks for listening!