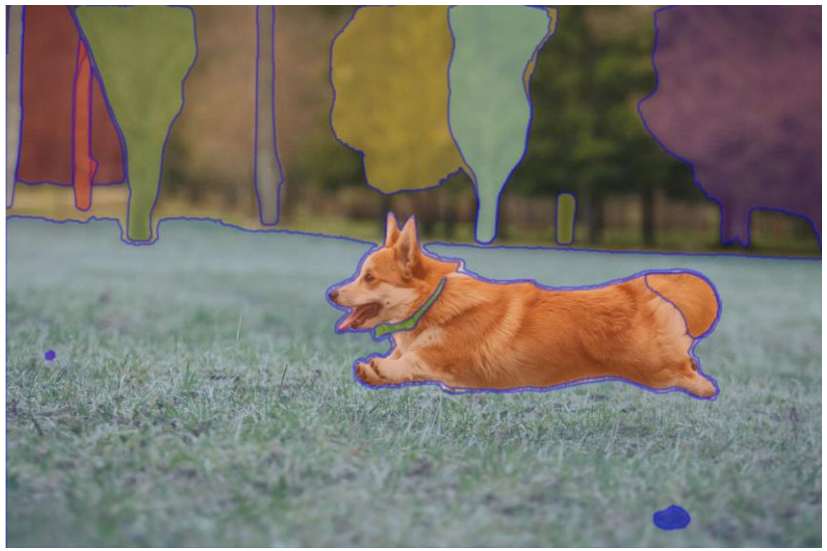
NANYANG TECHNOLOGICAL UNIVERSITY

CE7491 Research Paper Presentations

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

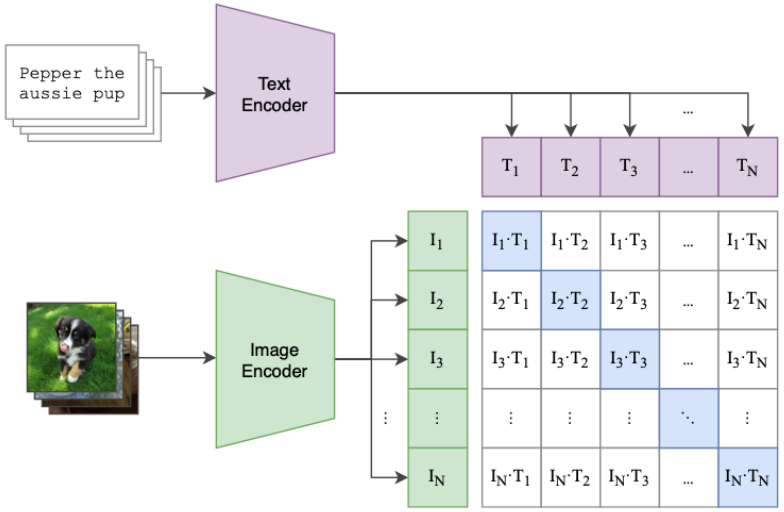


A Foundation Model for Image Segmentation, Marr Prize ICCV 2023

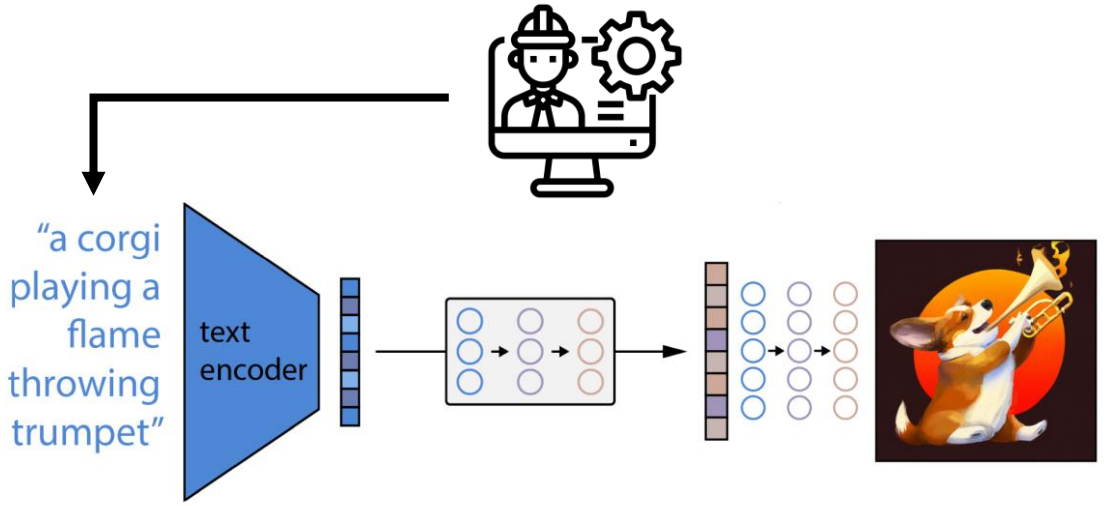
Background Context (General Problem)

- In NLP: large language models pre-trained on web-scale datasets demonstrate strong zero/few-shot capability (i.e., foundation models)
- In CV: most prominent illustration aligns paired text and images from the web (e.g., CLIP and ALIGN)

Upstream (CLIP)

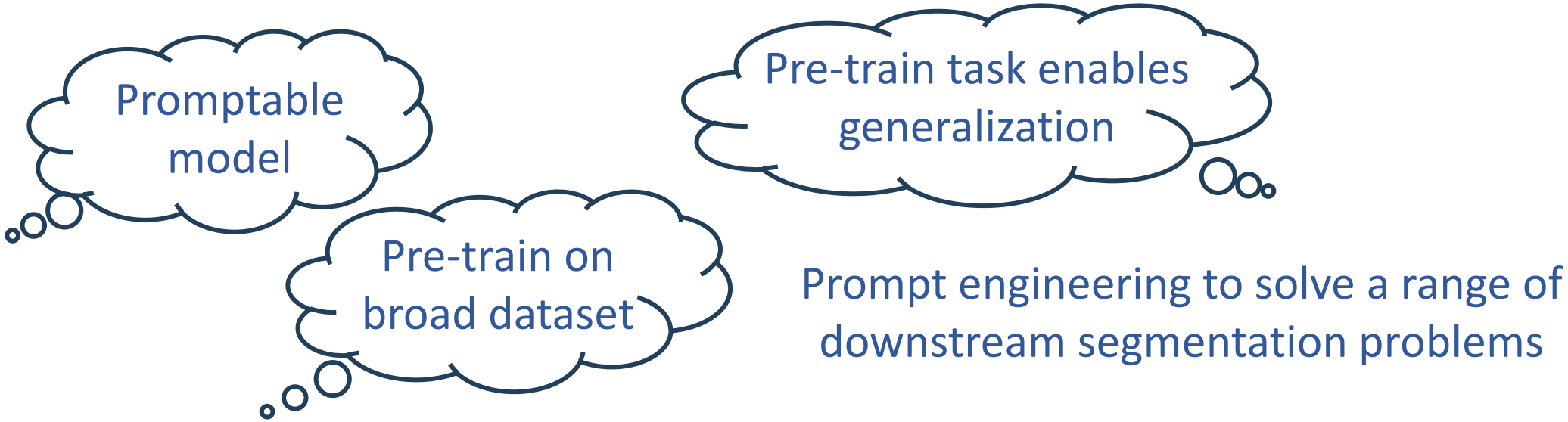


Downstream (DALL·E)



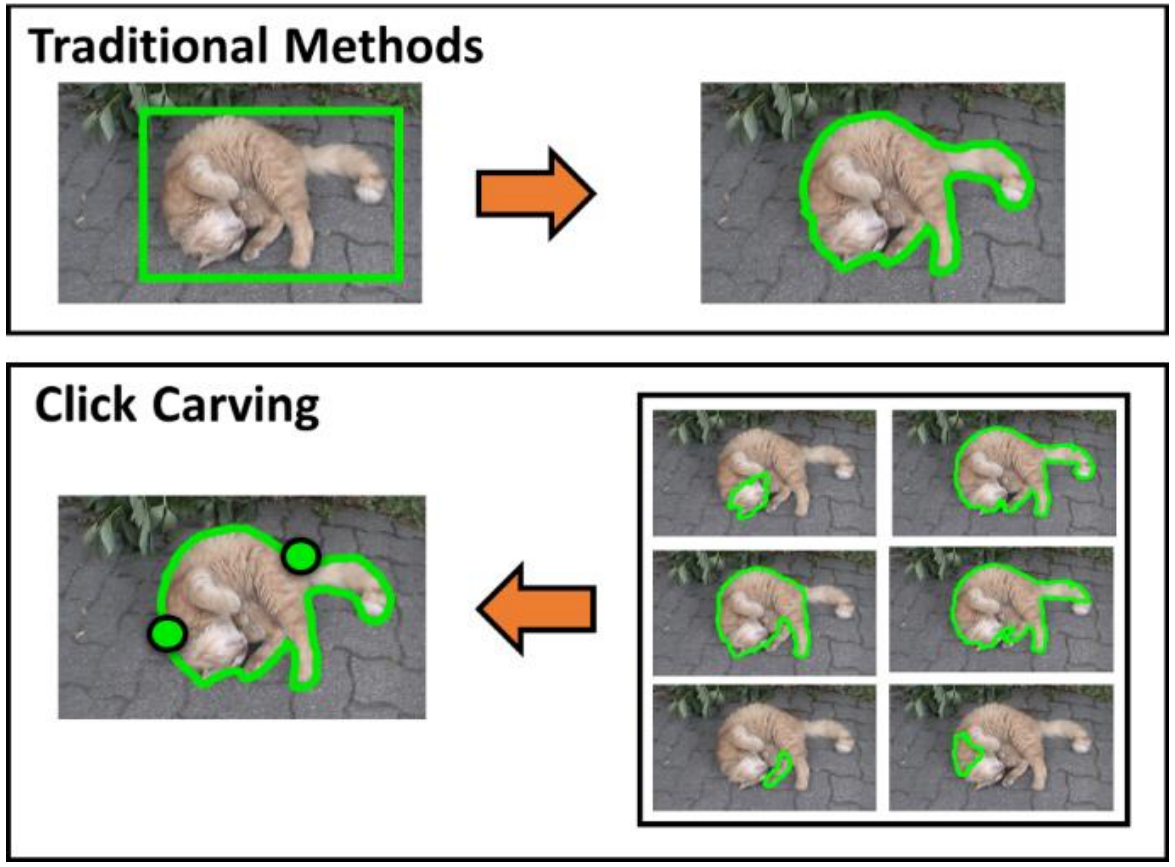
Background Context (General Problem)

- In NLP: large language models pre-trained on web-scale datasets demonstrate strong zero/few-shot capability (i.e., foundation models)
- In CV: most prominent illustration aligns paired text and images from the web (e.g., CLIP and ALIGN)
- In this work, the goal is a foundation model for image segmentation



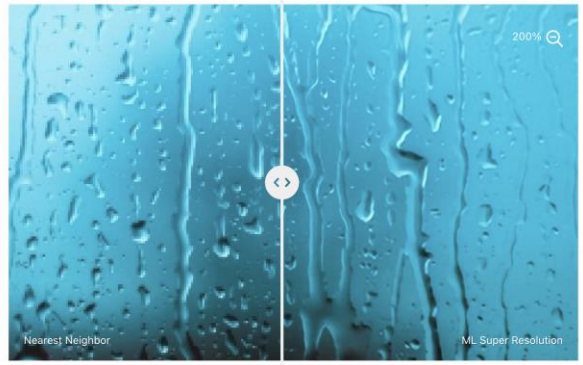
Background Context (Historical Evolution)

- Segmentation is a broad field:
  - Interactive segmentation
  - Edge detection
  - Super pixelization
  - Object proposal generation
  - Foreground segmentation
  - Semantic segmentation
  - Instance segmentation
  - Panoptic segmentation

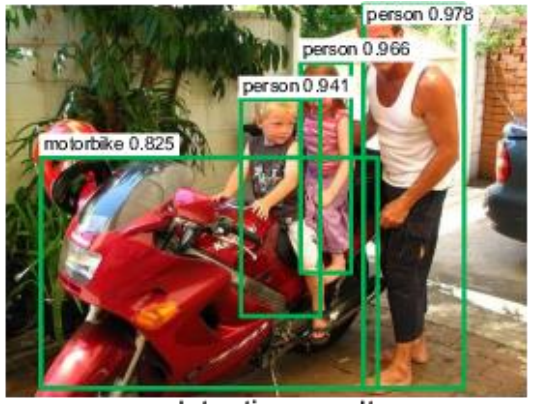


Background Context (Historical Evolution)

- Segmentation is a broad field:
  - Interactive segmentation
  - Edge detection
  - Super pixelization
  - Object proposal generation
  - Foreground segmentation
  - Semantic segmentation
  - Instance segmentation
  - Panoptic segmentation




region proposals



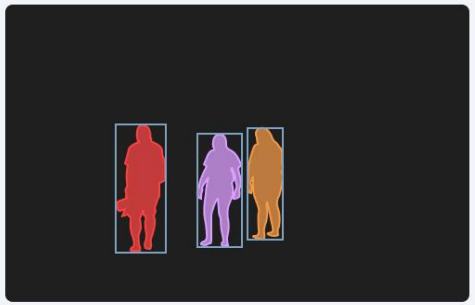
detection results

Background Context (Historical Evolution)

- Segmentation is a broad field:
  - Interactive segmentation
  - Edge detection
  - Super pixelization
  - Object proposal generation
  - Foreground segmentation
  - Semantic segmentation
  - Instance segmentation
  - Panoptic segmentation

(a) Image

(b) Semantic Segmentation



(c) Instance Segmentation

(d) Panoptic Segmentation

Background Context (Historical Evolution)

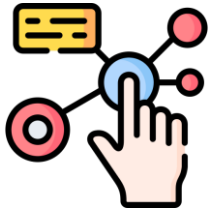
- Segmentation is a broad field
- A broadly capable model that can adapt to many existing and new segmentation tasks via prompt engineering
- Different from previous work on multi-task segmentation systems, where a single model performs a fixed set of tasks, the training and test tasks are the same
- Perform a new, different task at inference time by acting as a component in a larger system
- While interactive segmentation models are designed with human users in mind, a promptable segmentation model can also be composed into a larger algorithmic.

Background Context (Key Challenges)

- To enable zero-shot generalization, the promptable segmentation task to be defined needs to be general enough to support a wide range of downstream applications.
- The task requires a model that supports flexible prompting and can output segmentation masks in real-time for interactive use.
- To achieve strong generalization to new data distributions, it is necessary to train on a large and diverse dataset, geographically and economically.

Promptable Task

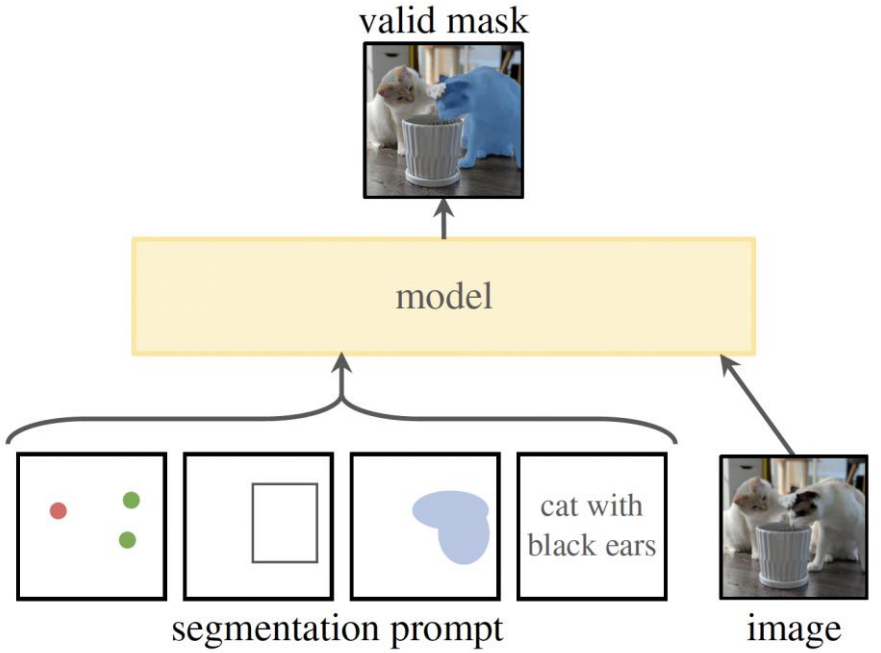
Flexible Model



Data Engine

The big idea – Promptable Segmentation

- Different types of prompts: points, rough box, mask, free-form text
- Task: return a valid segmentation mask given any prompt
- Valid means when the prompt is ambiguous, the output should be a reasonable mask for at least one of the objects



Model Architecture

- 3 components: image encoder, prompt encoder, and mask decoder
- Separate image and prompt encoder to reuse image embeddings

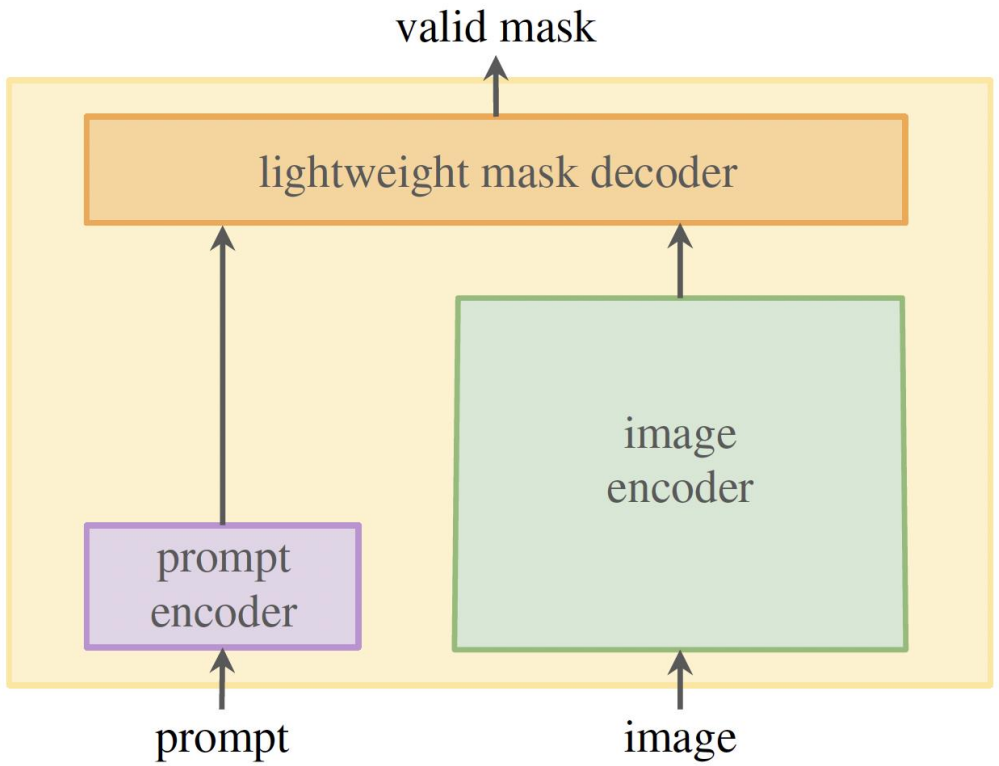
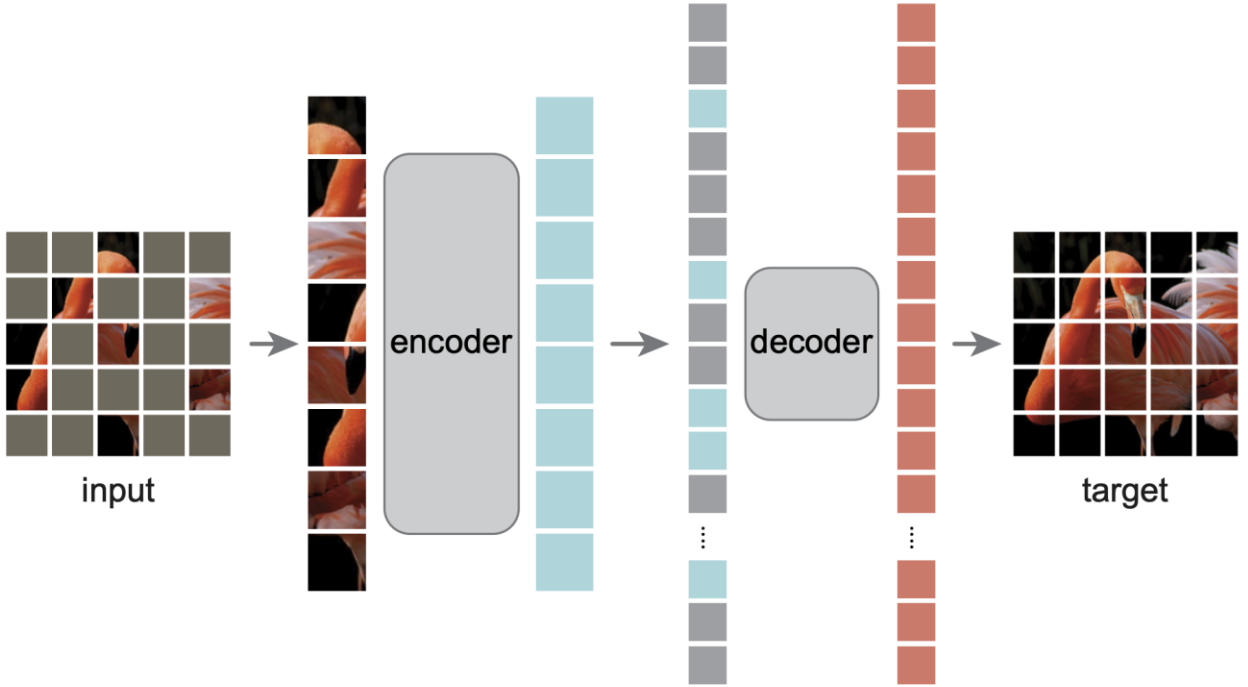


Image encoder

- A MAE pre-trained Vision Transformer (ViT) minimally adapted to process high resolution inputs
- The image encoder runs once per image and can be applied prior to prompting the model



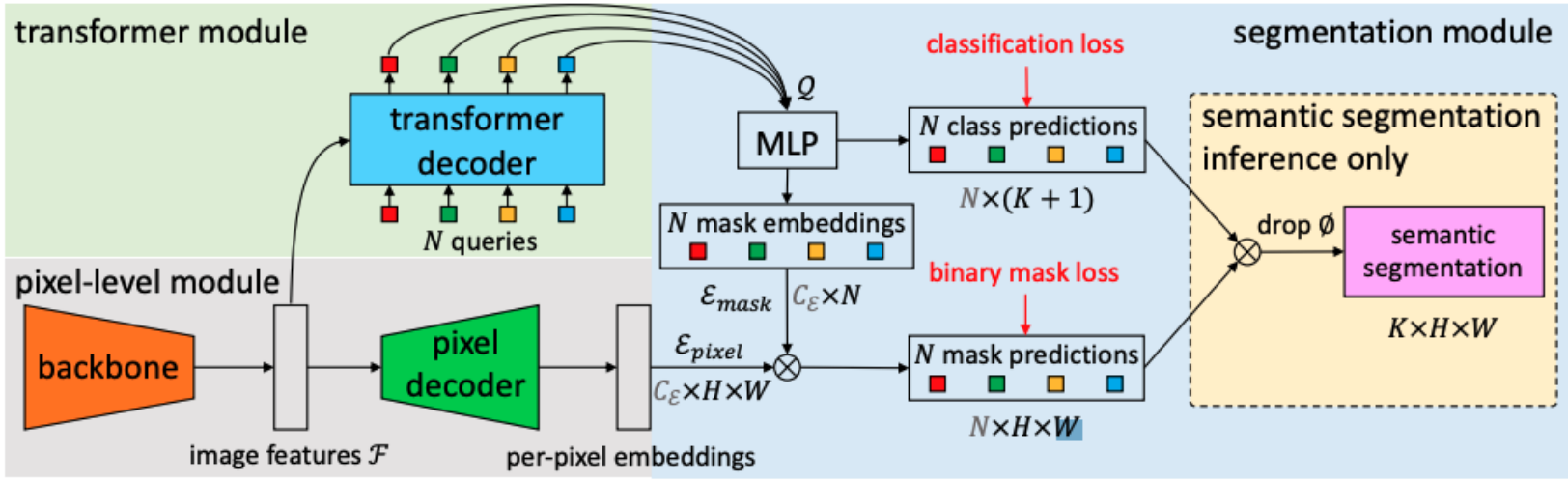
Prompt Encoder

- Two sets of prompts: sparse (point, box, text) and dense (mask)
- Points and boxes: positional encodings [1] summed with learned embedding for each prompt type
- Text: off-the-shelf text encoder from CLIP
- Mask: convolutions and summed element-wise with the image embedding

[1] *Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains*, NeurIPS'20

Mask decoder

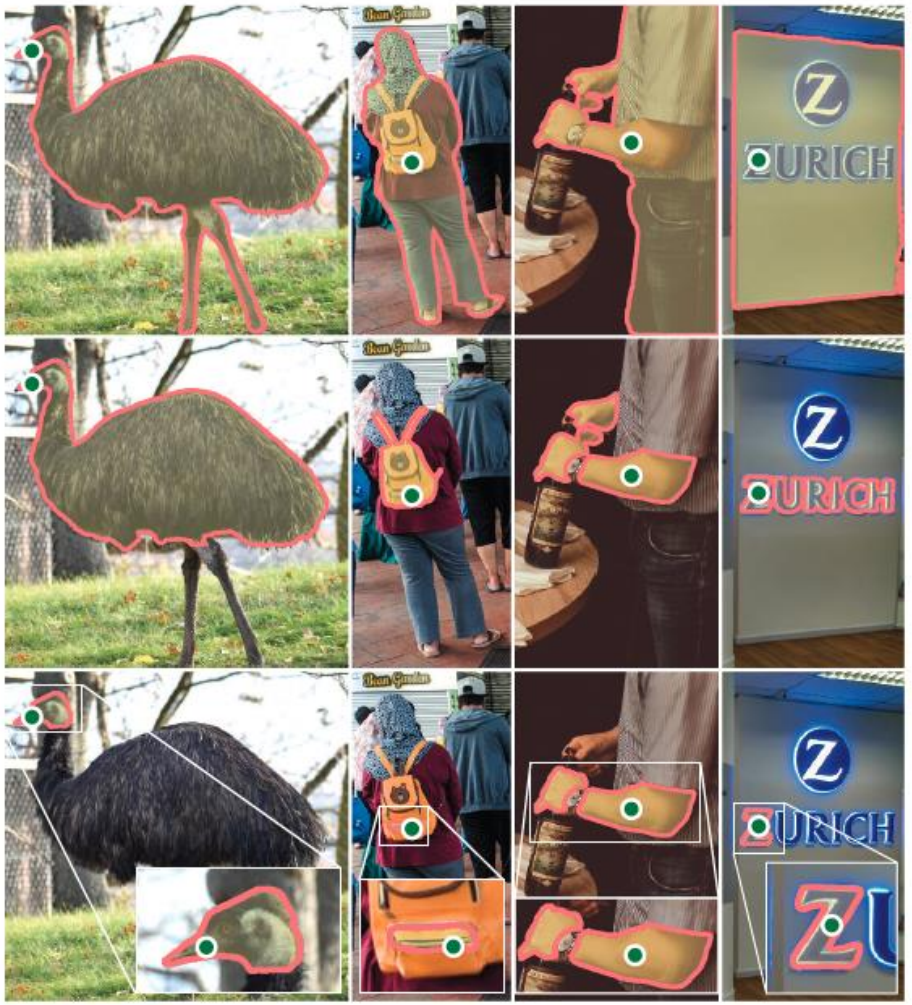
- Employs a modification of a Transformer decoder block [1]
- Use prompt self-attention and cross-attention in two directions to update all embeddings (prompt-to-image & image-to-image)
- Upsample the image embedding and use a dynamic linear classifier to computes the mask foreground probability at each pixel



[1] *Per-Pixel Classification is Not All You Need for Semantic Segmentation*, NeurIPS'21

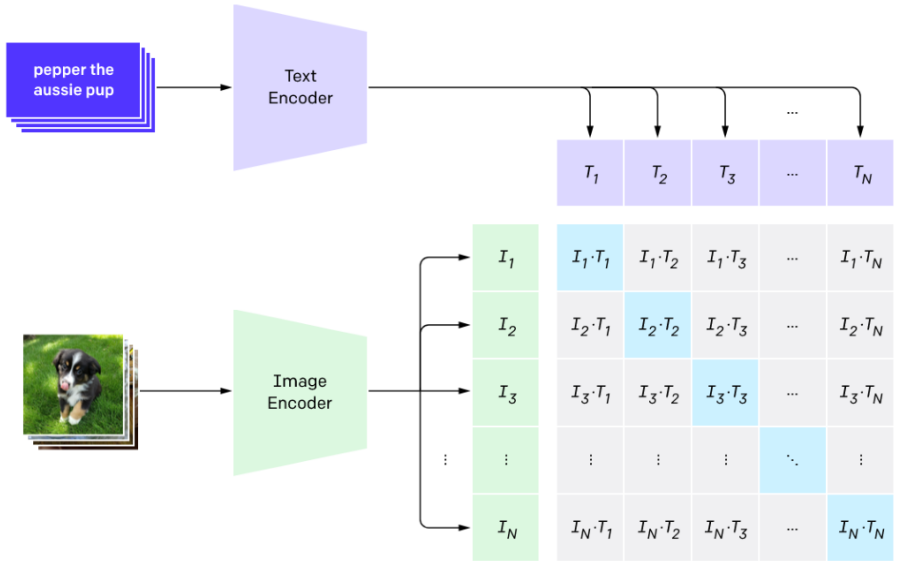
Resolving ambiguity

- With one output, the model will average multiple masks given ambiguous prompt
- SAM predict multiple output masks for a single prompt to resolve ambiguity. 3 outputs is found to be sufficient (whole, part, subpart)
- During training, only backprop the minimum loss over masks



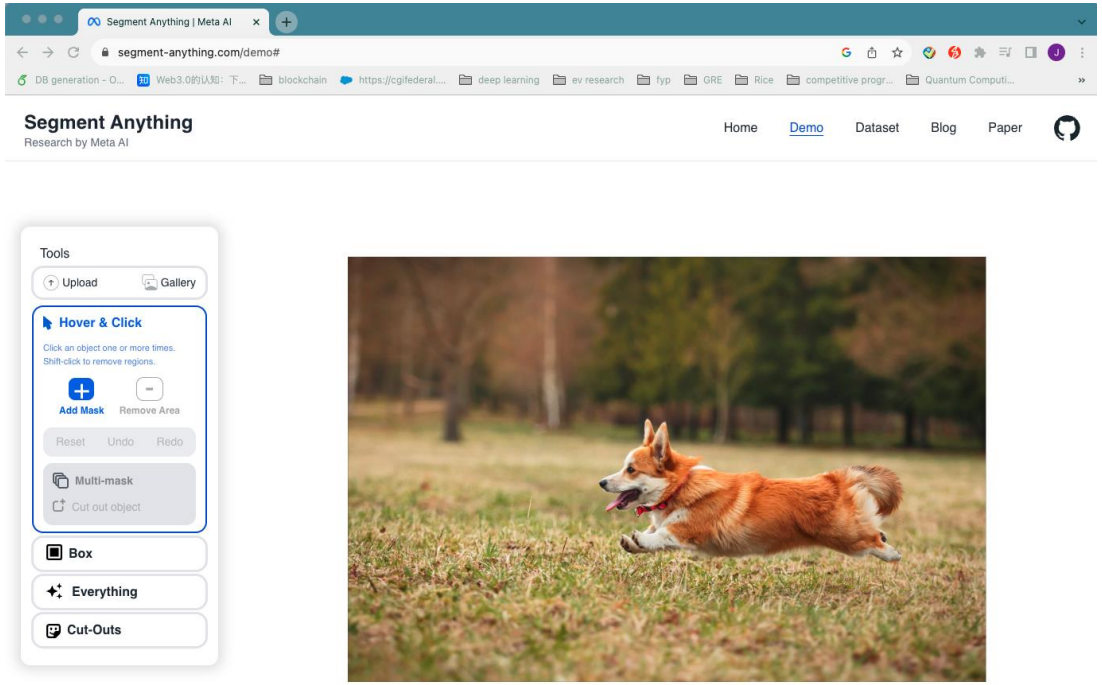
Loss and training

- Supervise mask prediction using focal loss and dice loss
- Training using a mixture of geometric prompts
- To train for text prompts, for each mask with area larger than 100^2, the CLIP image embedding is feed into SAM as the text embedding
- Key observation: CLIP's image embedding are trained to align with the text embeddings



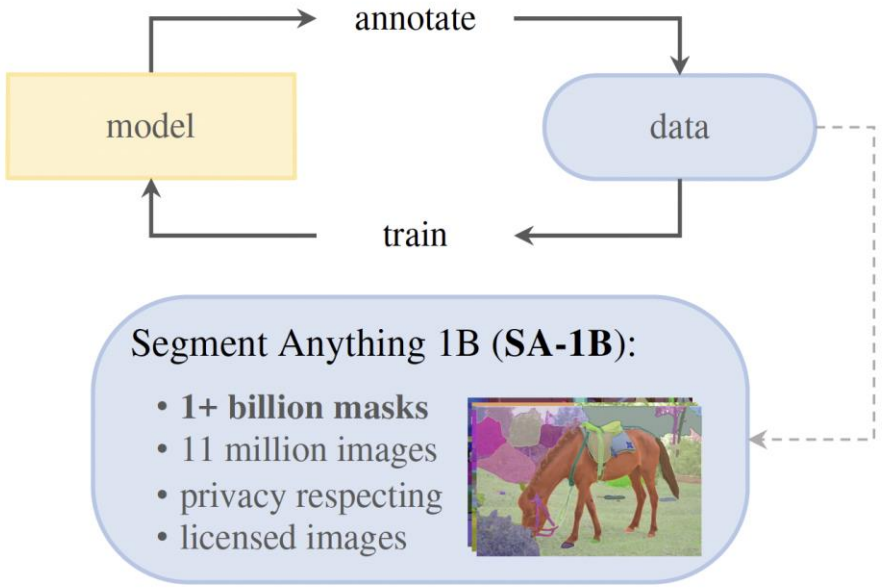
Efficiency

- The overall model design is largely motivated by efficiency
- Given a precomputed image embedding, the prompt encoder and mask decoder run in a web browser, on CPU, in ~50ms.
- This enables seamless, real-time interactive prompting



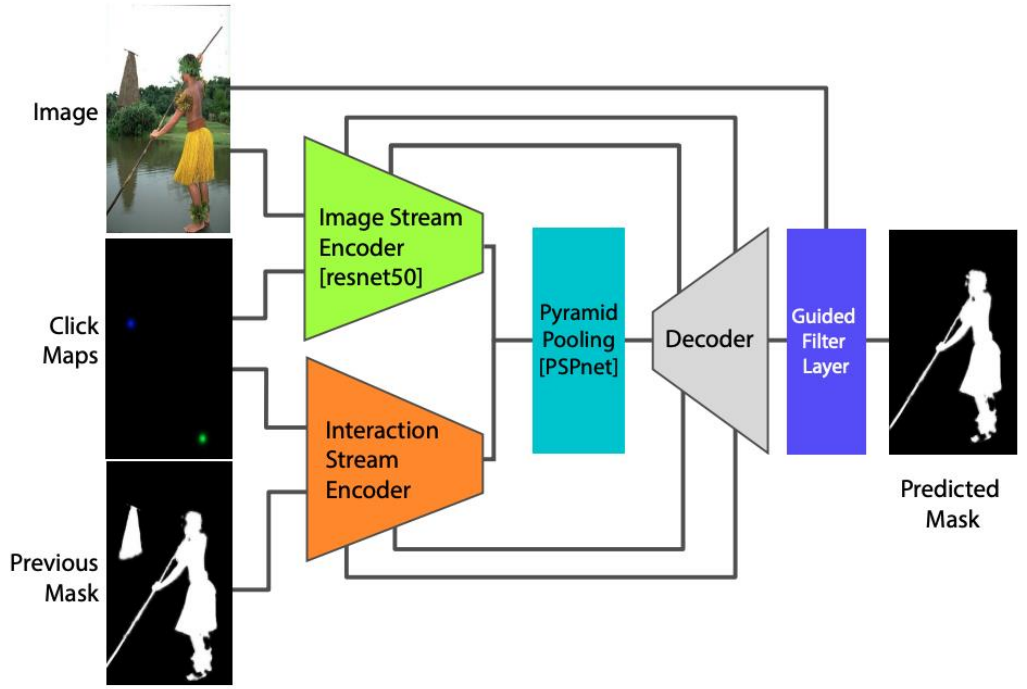
Data Engine

- To achieve strong generalization capability, a large and diverse set of training data is needed
- However, mask data are not naturally abundant, i.e. hard to obtain large amount of data from the Internet
- Solution: model-in-the-loop dataset annotation



Data Engine – assisted-manual stage

- Interactive segmentation with the assistance of a team of professional annotators
- The model-assisted annotation runs a real-time directly insider a browser (using precomputed image embeddings)



Data Engine: semi-automatic stage

- Aim to increase the diversity of masks and focus on less prominent objects
- First automatically detect confident masks
- Ask annotators with images prefilled with these masks to annotate additional objects

Data Engine: fully automatic stage

- Training without human supervision
- Prompt the model with a 32X32 regular grid of points
- For each point, predict a set of masks that may correspond to valid objects.
- Select confident and stable masks through IoU prediction module
- Apply non-maximal suppression to filter duplicates.

Results (In-distribution tasks)

As introduced in previous section, the pre-training task for SAM is to return a valid segmentation mask given any prompt.

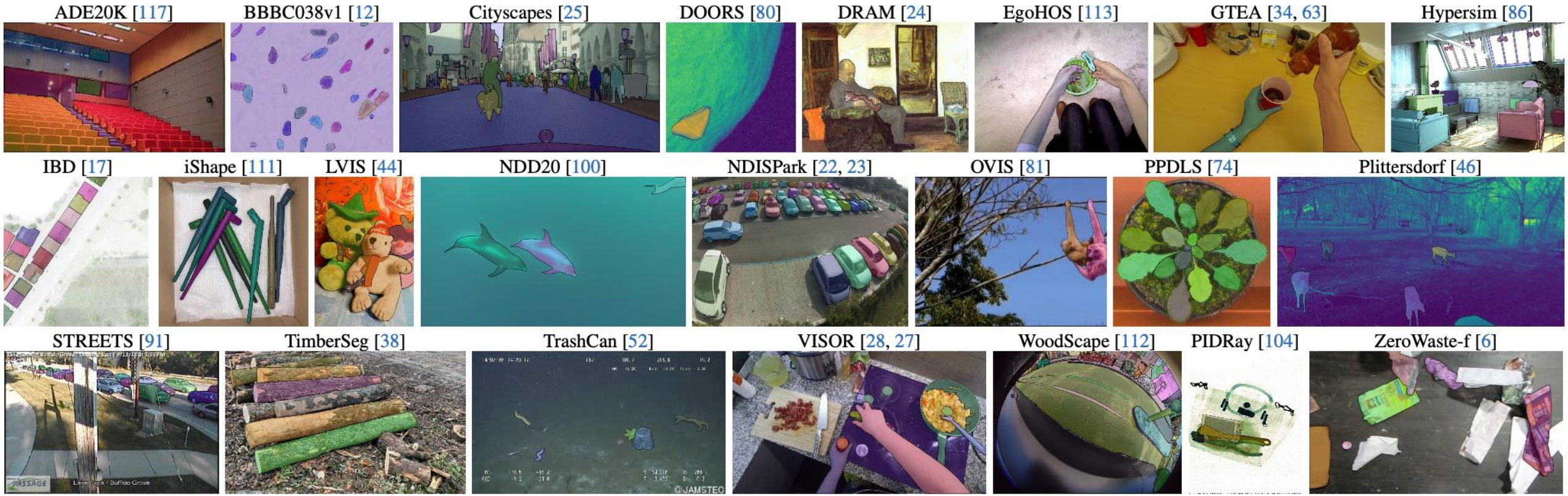
We will first show the visual results of this basic feature (the pretraining tasks)

https://youtu.be/tGOUvbEHb5Q

https://youtu.be/c9EO39unfPQ

Results (Evaluation Dataset)

~10k images sampled from 23 prior segmentation datasets with high variety.



Results (Zero-shot transfer tasks)

Zero-Shot Single Point Valid Mask Evaluation - In-distribution task

Zero-Shot Edge Detection - Low-level task

Zero-Shot Object Proposals - Mid-level task

Zero-Shot Instance Segmentation

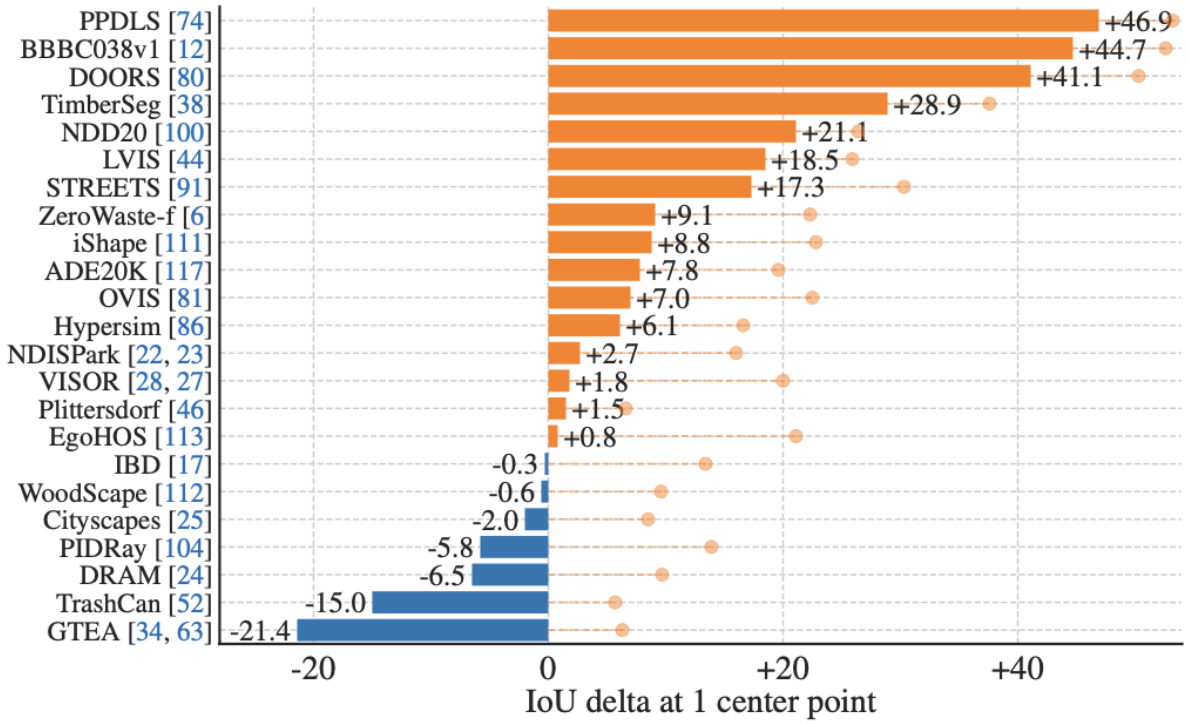
Zero-Shot Text-to-Mask

} Higher-level tasks

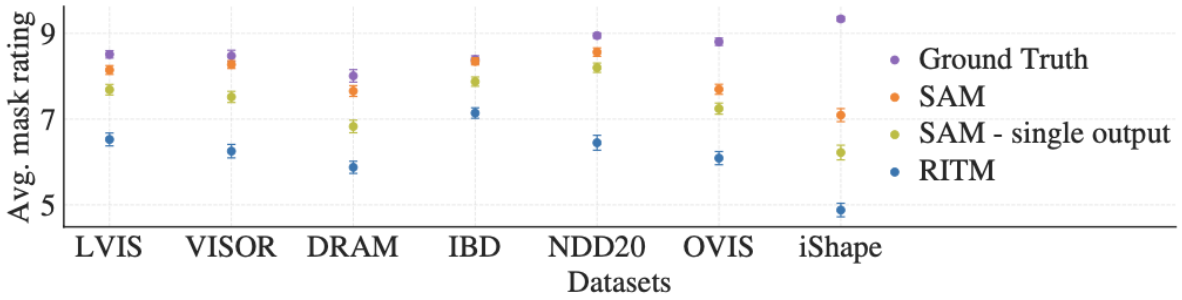
Results (Zero-Shot Single Point Valid Mask Evaluation)

Segmenting an object from a single foreground point is evaluated.

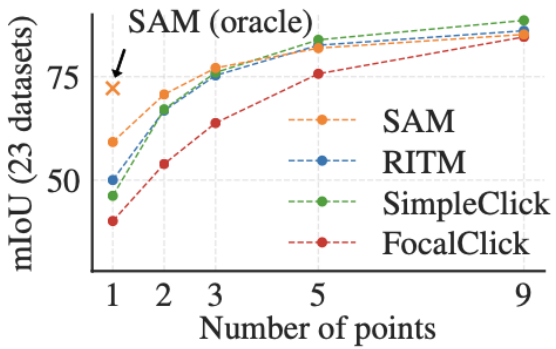
This task is ill-posed as one point can refer to multiple objects



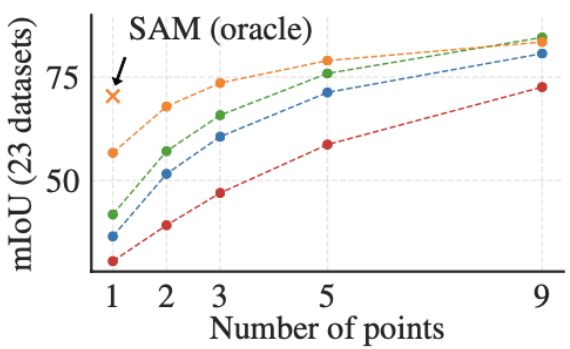
(a) SAM *vs.* RITM [92] on 23 datasets



(b) Mask quality ratings by human annotators

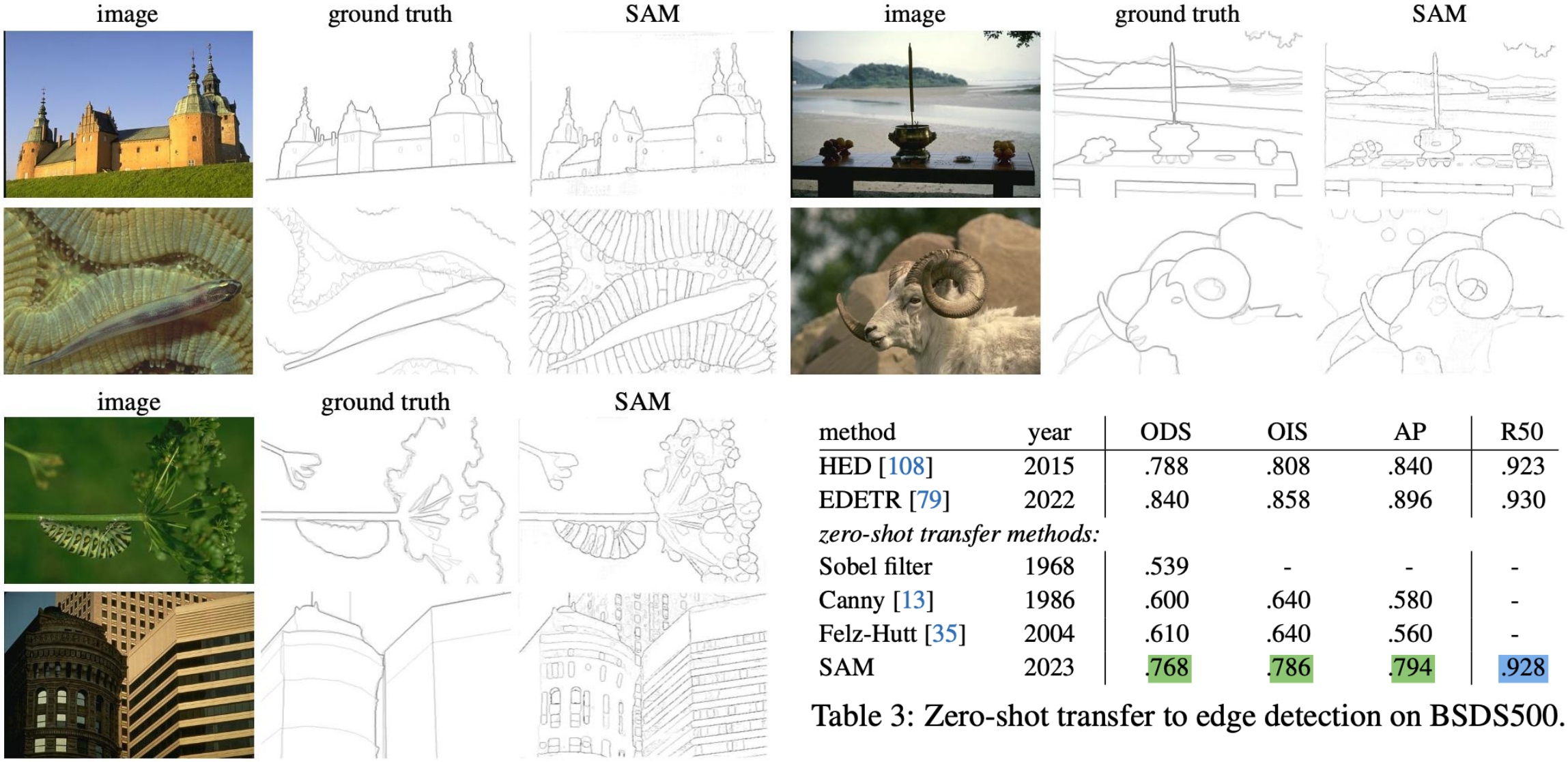


(c) Center points (default)



(d) Random points

Results (Zero-Shot Edge Detection)



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.

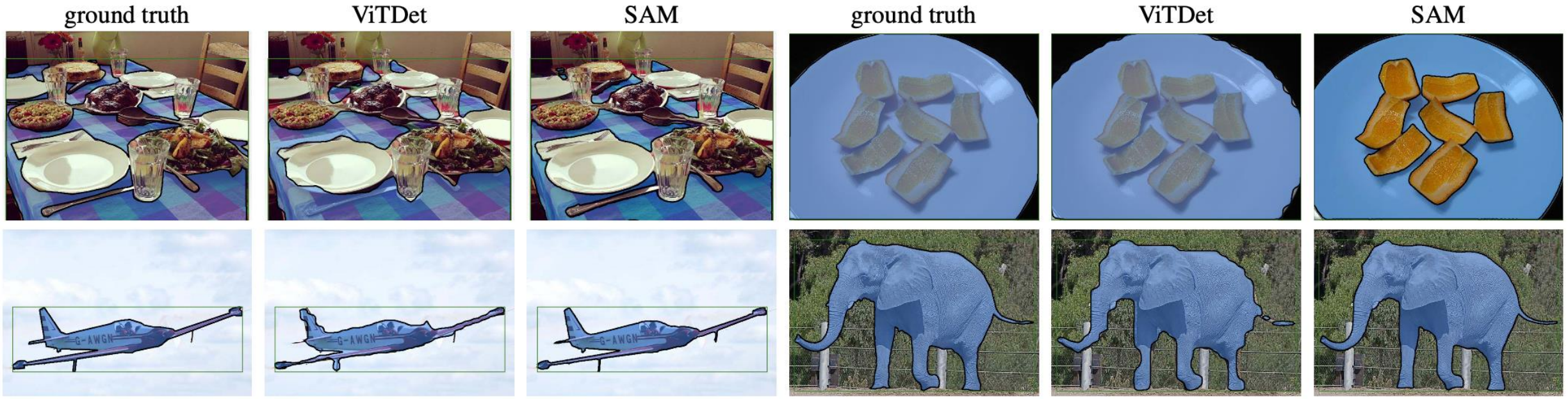
Results (Zero-Shot Object Proposals)

	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.

SAM does remarkably well on proposing masks for medium and large objects, as well as rare and common objects.

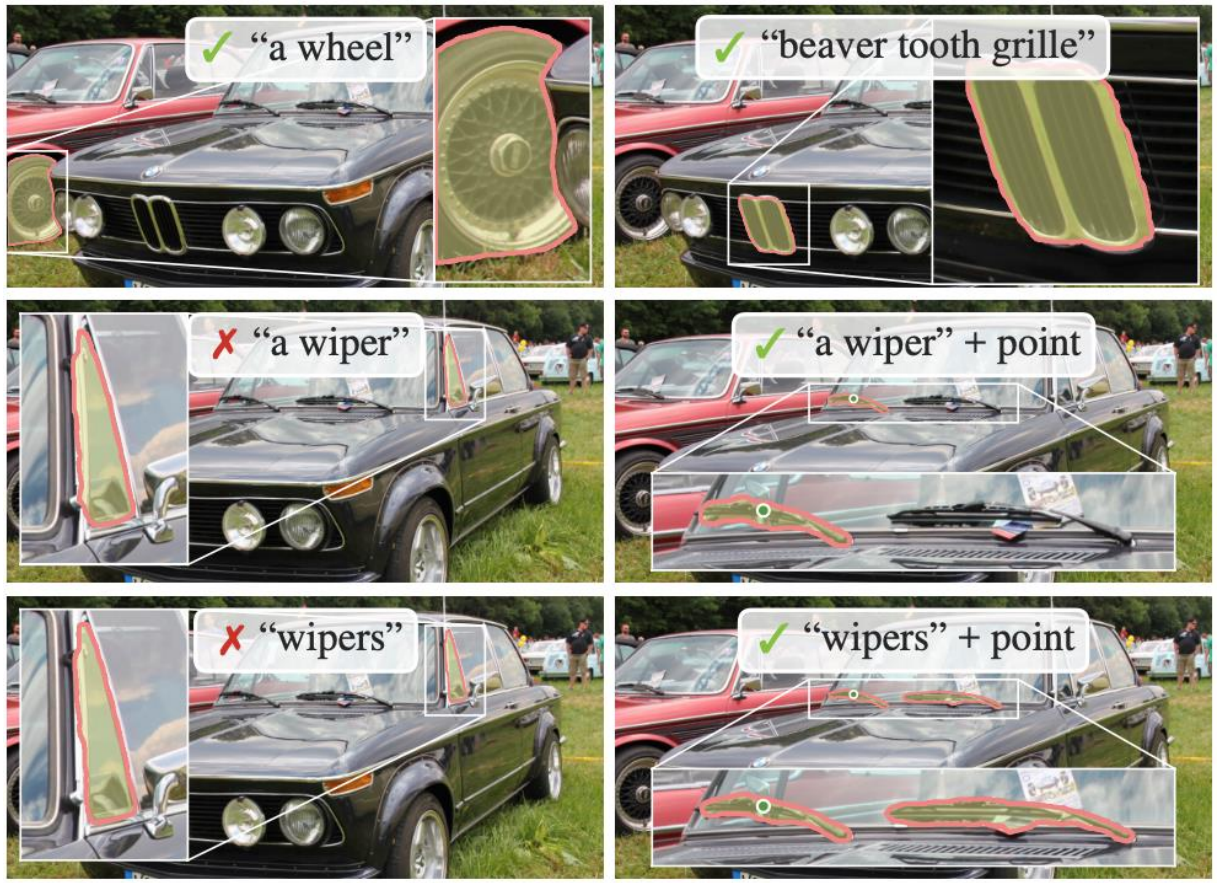
Results (Zero-Shot Instance Segmentation)



Compared to ViTDet, SAM tends to produce higher quality masks with cleaner boundaries.

* LVIS masks cannot contain holes by design so the plate is annotated amodally.

Results (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

Critical Analysis (Key Takeaways)

- This paper makes the attempt to lift image segmentation into the era of foundation models
- The promptable segmentation task, SAM model, and SA-1B dataset make this leap possible
- SAM demonstrates impressive performance and zero-shot transfer capability in various downstream tasks which significantly differ from the promptable segmentation task
- SAM will be utilized as a part of bigger systems, enabling new applications, designed for generality and breadth of user
- While SAM performs well in general, it is not perfect and limited to image segmentation

SEGMENT ANYTHING IN HIGH QUALITY

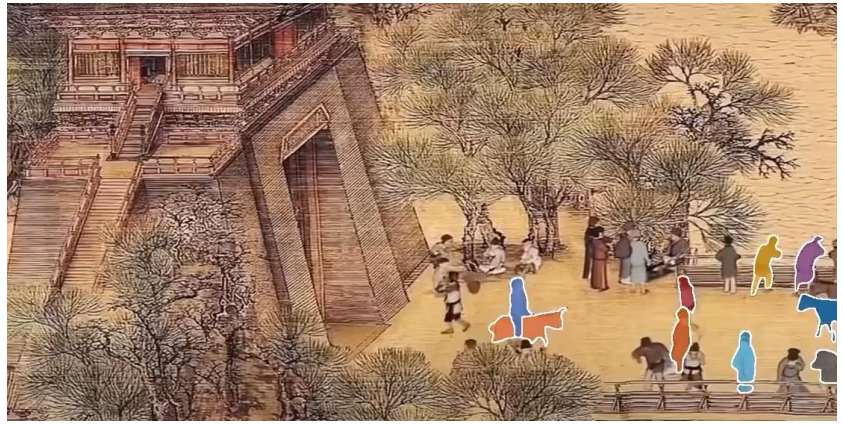
Critical Analysis (Pros and Cons)

- Zero-shot transfer to unseen distributions and diverse downstream tasks
- Real-time, seamless interactive prompting
- Fairness in segmenting people (gender, age, skin…)
- Strong quantitively and qualitative results on a variety of downstream tasks
  - Compositionality to be combined with other components in a large system

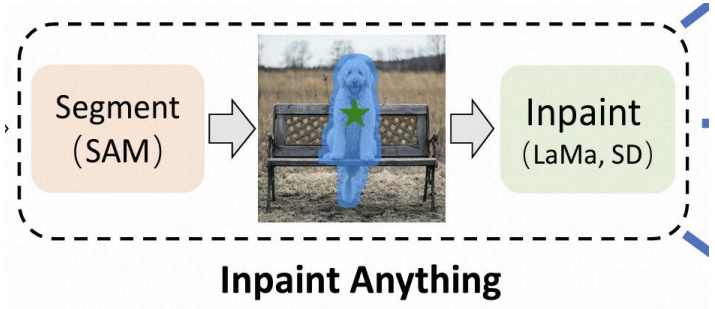
- A foundation model for image segmentation is an inherently limited scope
- Capacity mainly comes from large-scale supervised training rather than self-supervised training
  - It may miss fine structures, hallucinate small components, and make vague boundaries
- Prompts can be processed in real time, but image encoding can be heavy
- Unclear how to design simple prompts for semantic and panoptic segmentation

Critical Analysis (Subsequent Works & Extensions)

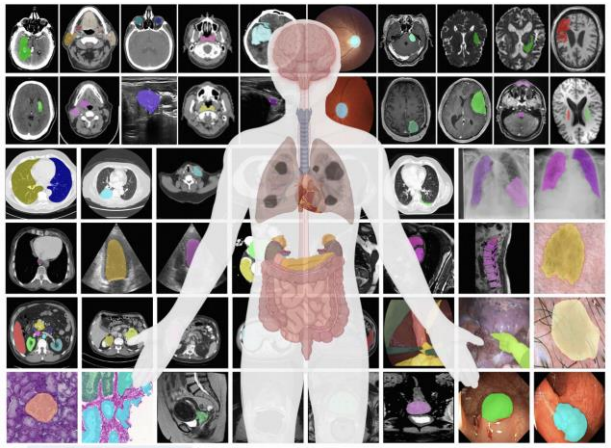
Track Anything



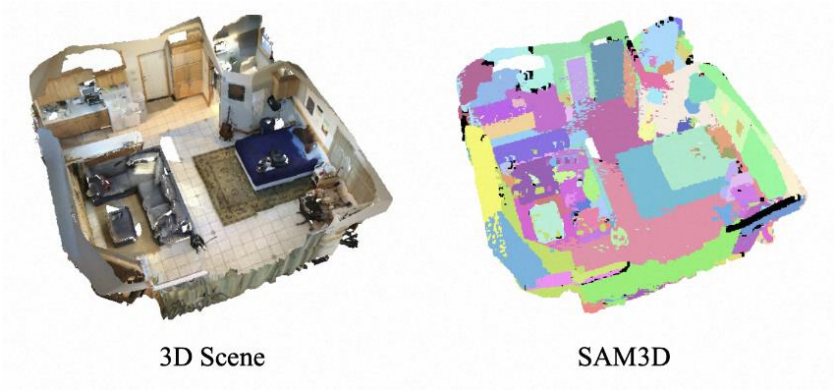
Inpainting

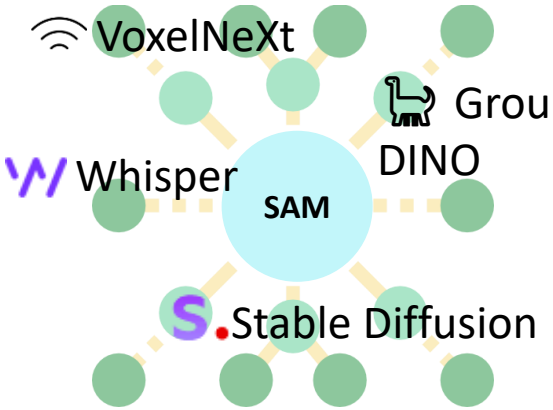


SAM in Medical Images

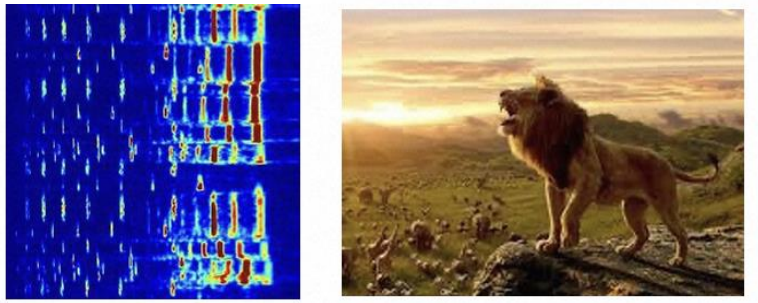


SAM 3D





Audio-Visual SAM



Anomaly Detection,

LiDAR-camera Calibration,

AR/VR video generation …

References

1. Kirillov A, Mintun E, Ravi N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.
2. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
3. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.
4. Yang Y, Wu X, He T, et al. SAM3D: Segment Anything in 3D Scenes[J]. arXiv preprint arXiv:2306.03908, 2023.
5. Ma J, Wang B. Segment anything in medical images[J]. arXiv preprint arXiv:2304.12306, 2023.
6. Yang J, Gao M, Li Z, et al. Track anything: Segment anything meets videos[J]. arXiv preprint arXiv:2304.11968, 2023.
7. Luo Z, Yan G, Li Y. Calib-Anything: Zero-training LiDAR-Camera Extrinsic Calibration Method Using Segment Anything[J]. arXiv preprint arXiv:2306.02656, 2023.
8. Cao Y, Xu X, Sun C, et al. Segment Any Anomaly without Training via Hybrid Prompt Regularization[J]. arXiv preprint arXiv:2305.10724, 2023.

Thank you!

Q&A