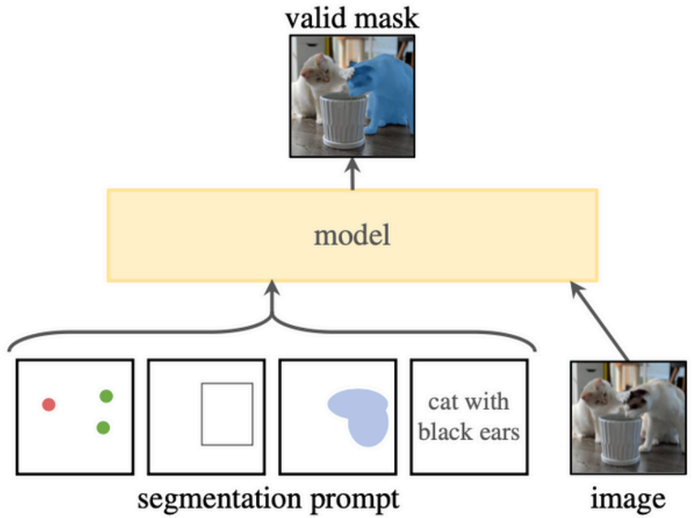
Segment Anything (SAM)

- R.M.K.C. Jayathissa 210258J
- L.A.S. Liyanaarachchi 210341H

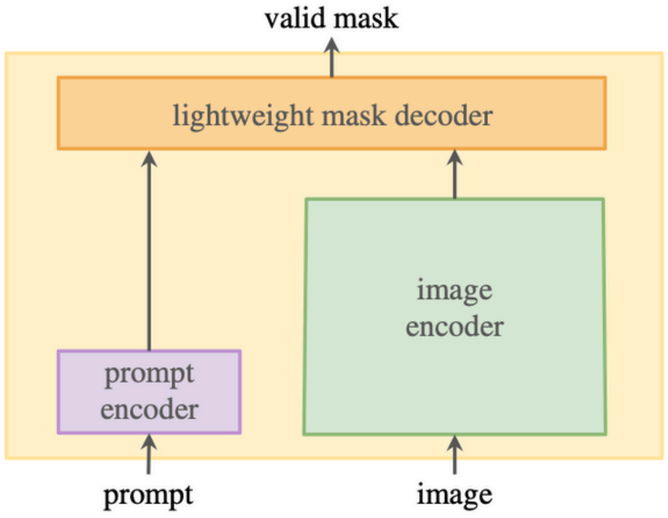
Image Credit : A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1497-1506.

Segment Anything Model (SAM): A Foundation Model for Image Segmentation

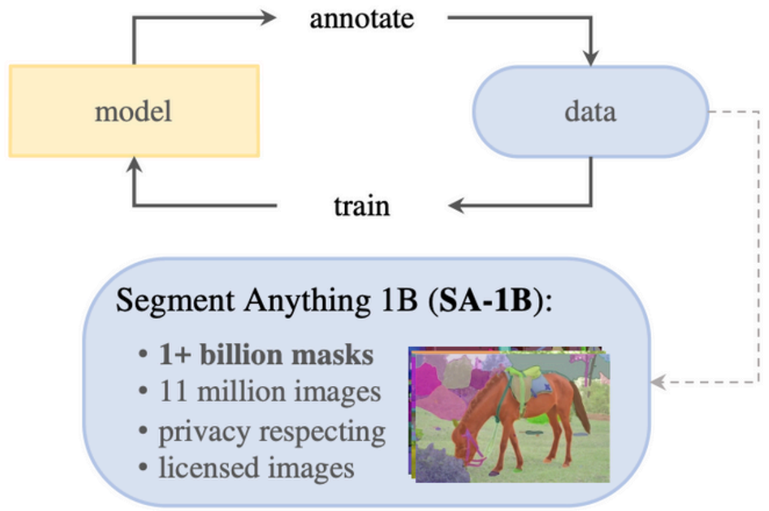
- Introduced by Meta AI Research in ICCV 2023.
- Goal: Build a foundation model for image segmentation that supports promptable segmentation and generalizes zero-shot to new tasks and datasets.
- Contributions:



A novel promptable segmentation task.



Segment Anything Model (SAM)

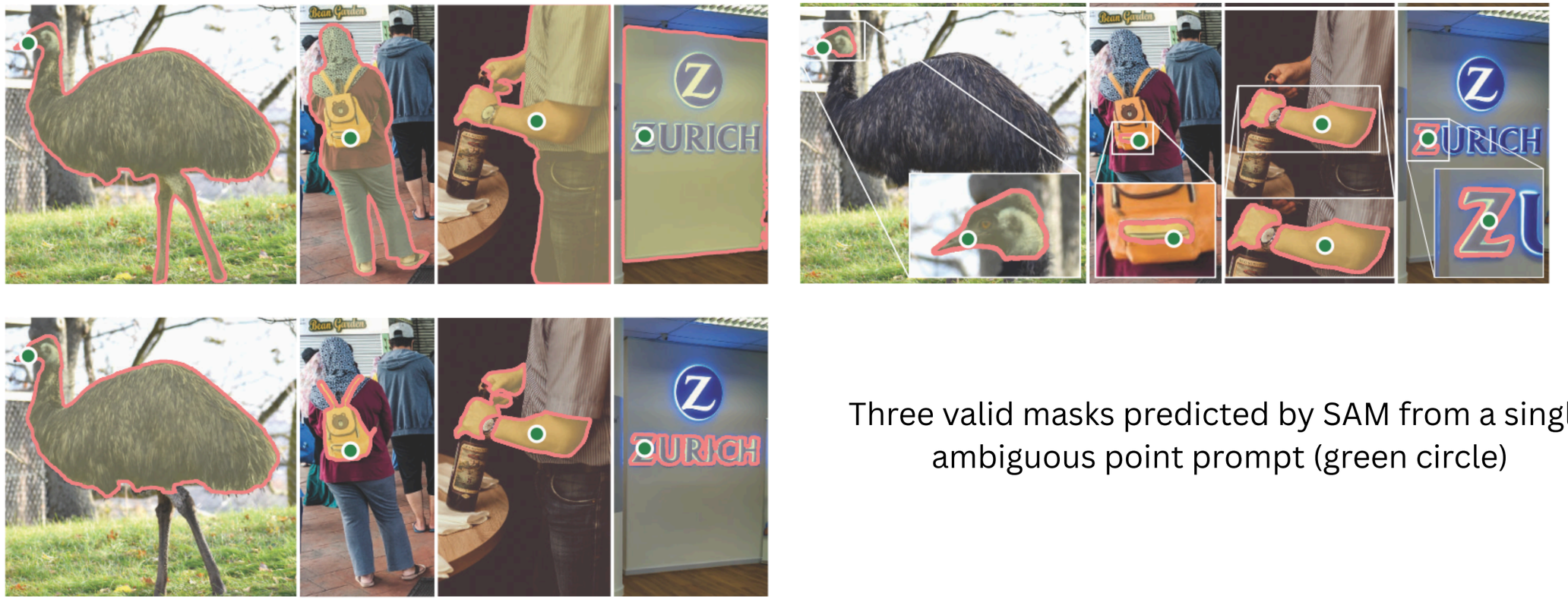


SA-1B dataset with 1.1 billion masks on 11 million images.

Image Credit : A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1497-1506.

The Promptable Segmentation Task

- **The task:** Given any segmentation prompt (point, box, mask, or text), predict a valid segmentation mask.
- **Handles ambiguity:** outputs multiple valid masks if the prompt is ambiguous (e.g., a point on a shirt may reflect the shirt or the person).
- Acts as a pre-training objective enabling zero-shot transfer to many downstream tasks by prompt engineering.



Three valid masks predicted by SAM from a single ambiguous point prompt (green circle)

Image Credit : A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1497-1506.

Segment Anything Model Architecture

- SAM consists of:
  - **Image encoder** (MAE-pretrained ViT) producing an image embedding.
  - **Prompt encoder** embedding sparse (points, boxes, text) and dense (masks) prompts.
  - **Lightweight mask decoder** that outputs segmentation masks in ~50ms.
- Designed for efficiency and interactive real-time use.
- Predicts multiple masks per prompt with confidence scores to resolve ambiguity.

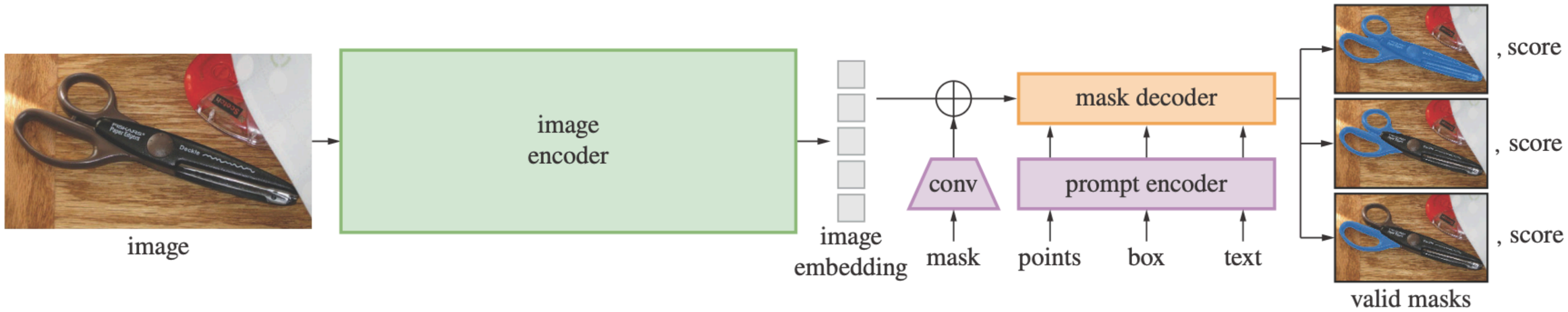


Figure 4: SAM architecture overview. The image encoder outputs an embedding that is queried by the prompt encoder and mask decoder to predict segmentation masks at amortized real-time speed. Multiple masks and confidence scores handle ambiguous prompts

Image Credit : A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1497-1506.

Data Engine and SA-1B Dataset

- Three annotation stages:
  - Assisted-manual with human annotators guided by SAM.
  - Semi-automatic with automatic mask pre-filling and human refinement.
  - Fully automatic with SAM generating dense masks using grid prompts.
- SA-1B dataset: 11M high-resolution images with 1.1B high-quality masks, mostly automatically generated.
- Validated mask quality: 94% of auto masks have >90% IoU with professionally corrected masks.

Figure 2: Example images from SA-1B with overlaid masks grouped by mask count per image.

Image Credit : A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1497-1506.

Data Engine and SA-1B Dataset

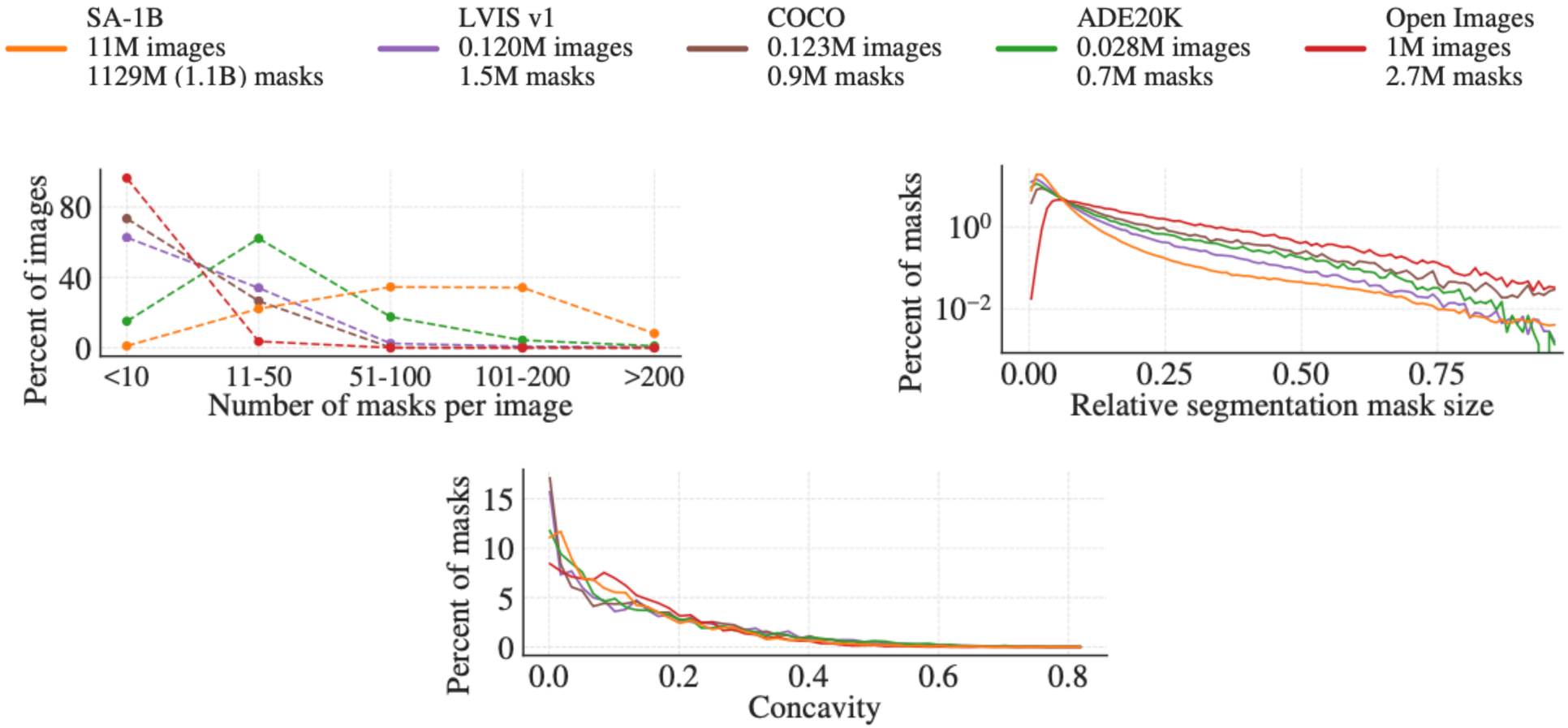


Figure 6: SA-1B mask statistics versus existing datasets, highlighting scale and diversity advantages.

Image Credit : A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1497-1506.

Results and Evaluation

- Evaluated on 23 diverse segmentation datasets not seen in training, verifying strong zero-shot transfer.
- In single foreground point segmentation, SAM outperforms prior interactive methods like RITM on most datasets.
- Human studies rate SAM masks significantly higher in quality.
- Predicting multiple masks per prompt allows resolving ambiguity, improving performance with oracle mask selection.
- Proof-of-concept zero-shot text-to-mask segmentation shows SAM's ability to segment from free-form text prompts with optional point refinement.

Key Highlights:

- Up to ~47 IoU improvement vs. baselines (oracle scenario).
- 94% auto-masks have >90% IoU with professional corrections.

Figure 7: Samples from 23 diverse evaluation datasets.

Image Credit : A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1497-1506.

Results and Evaluation

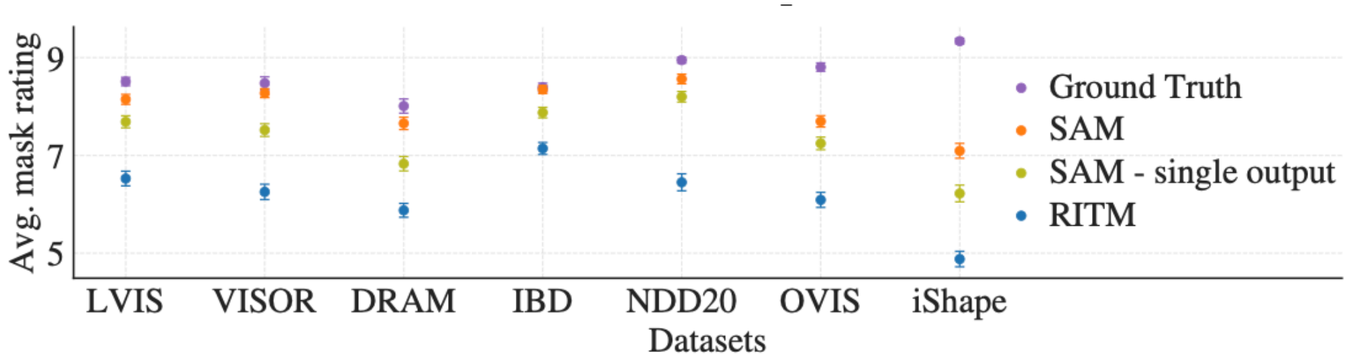


Figure 7: Human mask quality ratings favor SAM. (d,e) mIoU as prompt points increase, showing SAM's strong single-point performance.

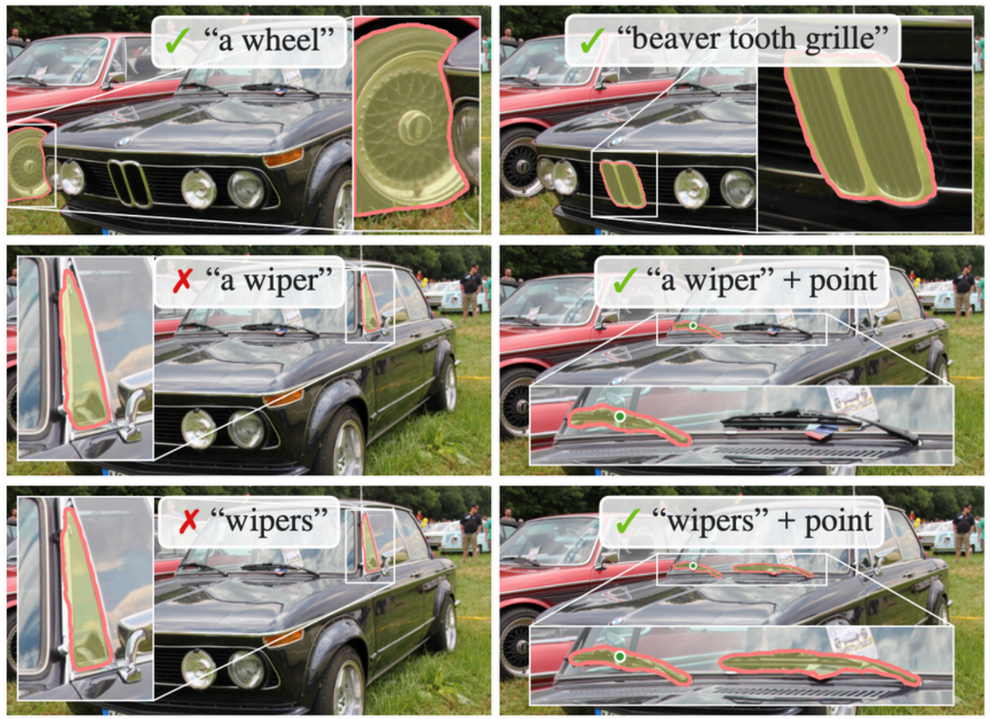


Figure 8: Zero-shot text-to-mask prediction examples with SAM segmenting objects from text prompts and improved with point prompts.

Image Credit : A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1497-1506.

Thank You