A review of **Segment Anything (SAM)**, 2023

Zachary Parent

UPC - UB

April 25, 2025

Outline

The Vision: A Foundation Model for Segmentation

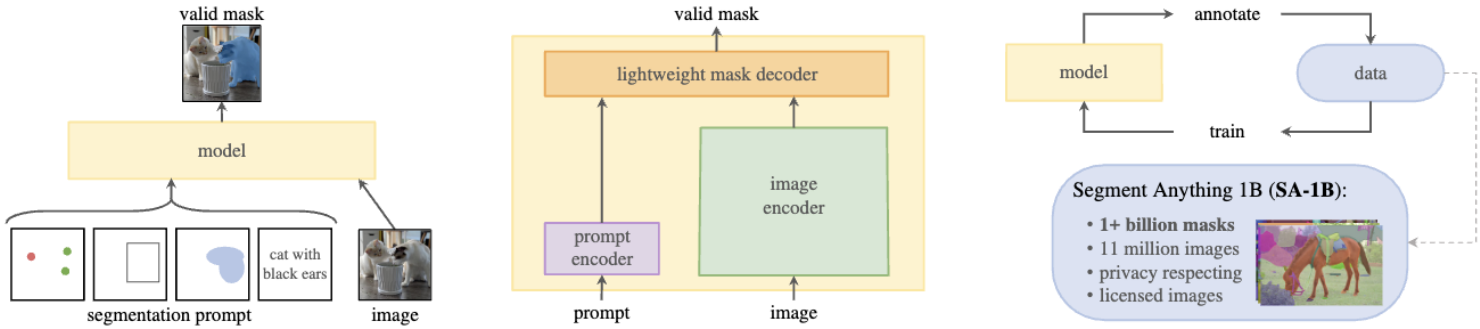
Goal: Build a foundation model analogous to LLMs for vision, specifically for segmentation.

- Like GPT, but for pixels.
- Enables zero-shot generalization via prompt engineering.
- Three core components: Task, Model, Dataset.

Released model (SAM) and dataset (SA-1B) to foster research.

Key Research Questions

1. What task will enable zero-shot generalization?
2. What is the corresponding model architecture?
3. What data can power this task and model?



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure: We aim to build a foundation model for segmentation by introducing three interconnected components: a prompt- able 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.

Defining the Task: Promptable Segmentation

Input: An image and a prompt (point, box, mask, text).
Output: A valid segmentation mask.

- "Valid" means outputting a reasonable mask even if the prompt is ambiguous.
- Inspired by interactive segmentation, but goal is always predicting a valid mask immediately.

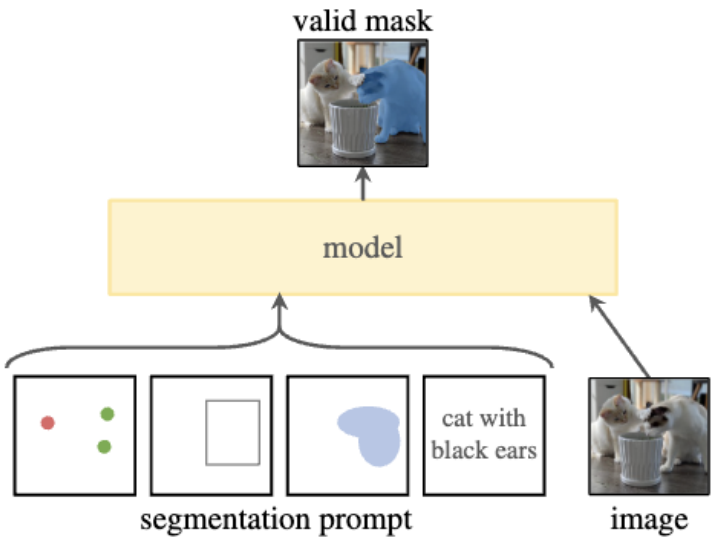


Figure: Prompt types (a) Point prompts (b) Box prompts (c) Mask prompts (d) Text prompts.

Zero-Shot Transfer via Prompting

The core idea: Train a model on the general promptable task, then solve specific downstream tasks by designing the right prompts.

- Example: Use bounding box outputs from a cat detector as prompts to get cat instance segmentation.
- SAM acts as a composable component in a larger system.
- Can adapt to many (but not all) existing and new tasks.

Model Requirements & Design

Requirements:

- Flexible Prompts: Handle various input types.
- Real-time: Amortized real-time inference (∼50ms) for interactivity.
- Ambiguity-aware: Output multiple masks for ambiguous prompts.

Design: Decoupled architecture for efficiency.

1. Heavyweight Image Encoder (ViT-H MAE pre-trained) - 1x per image.
2. Lightweight Prompt Encoder (Positional embeddings, CLIP for text).
3. Lightweight Mask Decoder (Transformer) - combines image and prompt embeddings.

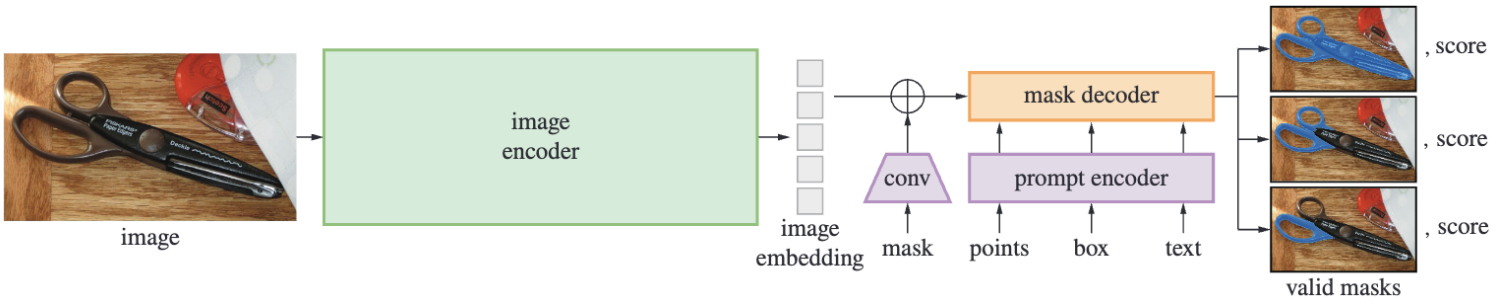


Figure: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple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

Encoder/Decoder Details

Image Encoder: MAE pre-trained ViT. **Prompt Encoder:**

- Sparse (Points, Boxes, Text): Positional encodings + learned embeddings per prompt type. Text uses CLIP.
- Dense (Masks): CNN downscales mask, combines with image embedding.

Mask Decoder:

- Transformer decoder updates prompt tokens via self/cross-attention.
- Upscales image embedding.
- Predicts mask logits via MLP + dynamic linear classifier using output tokens.

Loss: Linear combination of Focal Loss and Dice Loss.

The Need for Data & The Data Engine

Problem: No web-scale segmentation data exists. Solution: Build a "Data Engine" - iterate between model-assisted annotation and model improvement. **Three Stages:**

1. **Assisted-Manual:** Annotators label interactively with SAM's help. Model retrained iteratively.
2. **Semi-Automatic:** SAM proposes masks for likely objects; annotators focus on the rest.
3. **Fully Automatic:** Prompt SAM with a regular grid (32x32) of points $\rightarrow$ 100 masks/image.

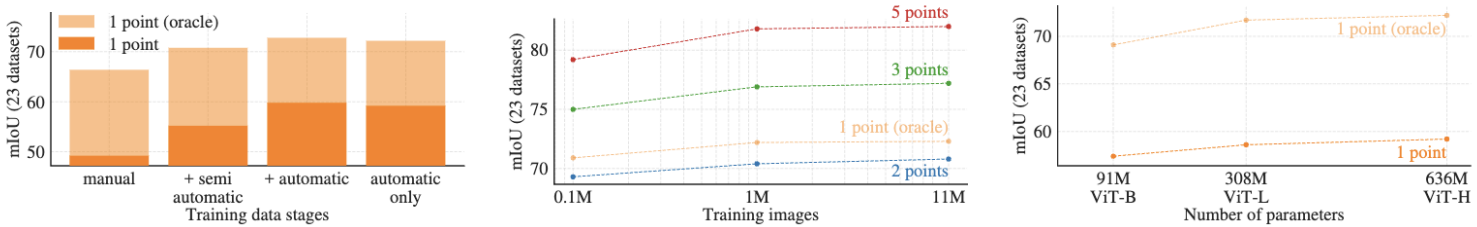


Figure: Ablation studies of our data engine stages, image encoder scaling, and training data scaling. (Left) Each data engine stage leads to improvements on our 23 dataset suite, and training with only the automatic data (our default) yields similar results to using data from all three stages. (Middle)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

SA-1B Dataset

Result of the fully automatic stage.

- **1.1 Billion** high-quality masks.
- **11 Million** diverse, high-resolution images (licensed from provider).
- **400x** more masks than previous largest datasets (COCO, LVIS).
- Masks automatically generated and filtered for quality/stability (IoU, stability threshold).
- Geographically and economically diverse image sources.

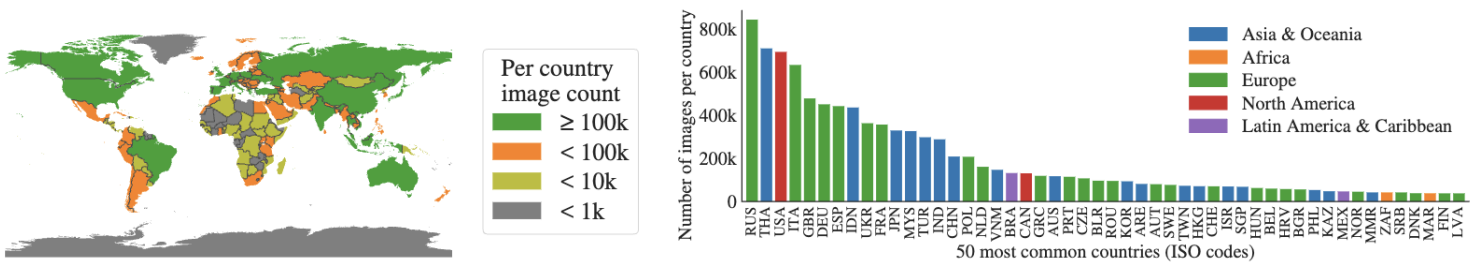
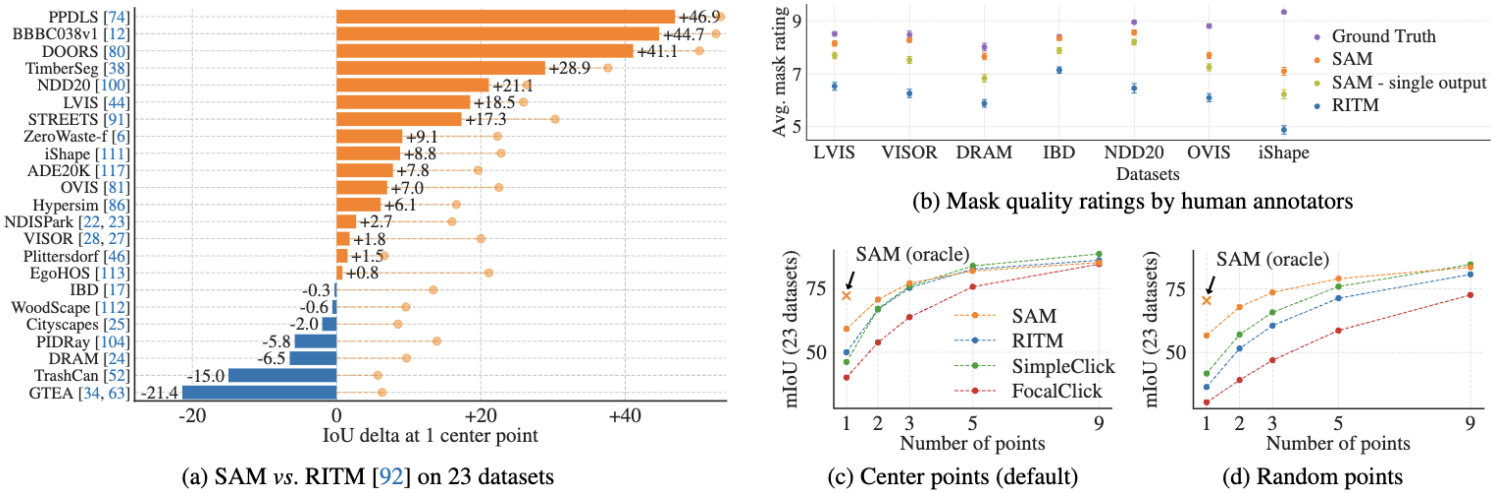


Figure: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.

Zero-Shot Evaluation

Extensive evaluation on 23 diverse segmentation datasets.

- Single point prompt performance often near manually annotated GT.
- Strong results on zero-shot downstream tasks via prompt engineering:
  - Edge Detection
  - Object Proposal Generation
  - Instance Segmentation
  - Text-to-Mask (preliminary)



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

(b) Mask quality ratings by human annotators

(c) Center points (default)

(d) Random points

Figure: 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.

Strengths

- **Foundation Model Vision:** Pioneering promptable, general segmentation.
- **Unifying Task:** Promptable segmentation enables flexibility and zero-shot transfer.
- **Massive Dataset (SA-1B):** Unprecedented scale and diversity.
- **Strong Zero-Shot Performance:** Impressive generalization across many tasks/datasets.
- **Efficient Architecture:** Decoupled design allows real-time interaction.

Limitations & Future Work

- **Dataset Quality Assurance:** Transparency needed on automated mask filtering effectiveness.
- **Downstream Implementation:** Lack of practical integration examples.
- **Prompt Engineering:** Methodology underexplored.
- **Shallow Semantics:** Text understanding needs improvement.
- **Bias Analysis:** Surface-level; deeper sociotechnical audit needed.
- **Generalist vs Specialist:** May not beat specialized models in niche domains.

Conclusion

Segment Anything introduces a powerful new paradigm for image segmentation.

- Successfully implements a promptable task, model, and large-scale dataset.
- Demonstrates strong zero-shot capabilities.
- Opens up many avenues for future research in foundation models for vision, prompt engineering, and dataset quality.

A significant step towards general-purpose, interactive image understanding.

Thank You / Q&A

Questions?

Model and dataset available at:
`https://segment-anything.com`