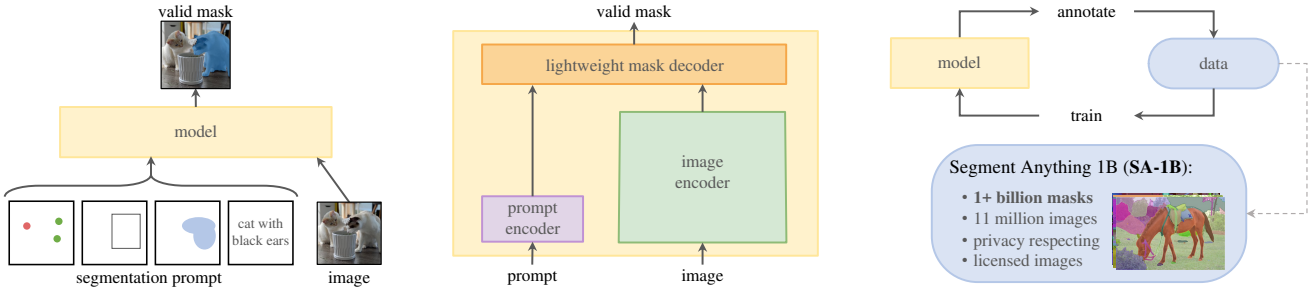
Segment Anything

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) **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.

Abstract

We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our efficient model in a data collection loop, we built the largest segmentation dataset to date (by far), with over 1 ***billion*** *masks on 11M licensed and privacy respecting images. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. We evaluate its capabilities on numerous tasks and find that its zero-shot performance is impressive – often competitive with or even superior to prior fully supervised results. We are releasing the Segment Anything Model (SAM) and corresponding dataset (SA-1B) of 1B masks and 11M images at https://segment-anything.com to foster research into foundation models for computer vision.*

1. Introduction

Large language models pre-trained on web-scale datasets are revolutionizing NLP with strong zero-shot and few-shot generalization [10]. These "foundation models" [8] can generalize to tasks and data distributions beyond those seen during training. This capability is often implemented with *prompt engineering* in which hand-crafted text is used to prompt the language model to generate a valid textual response for the task at hand. When scaled and trained with abundant text corpora from the web, these models' zero and few-shot performance compares surprisingly well to (even matching in some cases) fine-tuned models [10, 21]. Empirical trends show this behavior improving with model scale, dataset size, and total training compute [56, 10, 21, 51].

Foundation models have also been explored in computer vision, albeit to a lesser extent. Perhaps the most prominent illustration aligns paired text and images from the web. For example, CLIP [82] and ALIGN [55] use contrastive learning to train text and image encoders that align the two modalities. Once trained, engineered text prompts enable zero-shot generalization to novel visual concepts and data distributions. Such encoders also compose effectively with other modules to enable downstream tasks, such as image generation (*e.g.*, DALL·E [83]). While much progress has been made on vision and language encoders, computer vision includes a wide range of problems beyond this scope, and for many of these, abundant training data does not exist.

In this work, our goal is to build *a foundation model for image segmentation*. That is, we seek to develop a promptable model and pre-train it on a broad dataset using a task that enables powerful generalization. With this model, we aim to solve a range of downstream segmentation problems on new data distributions using prompt engineering.

The success of this plan hinges on three components: **task**, **model**, and **data**. To develop them, we address the following questions about image segmentation:

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

These questions are entangled and require a comprehensive solution. We start by defining a *promptable segmentation* **task** that is general enough to provide a powerful pre-training objective and to enable a wide range of downstream applications. This task requires a **model** that supports flexible prompting and can output segmentation masks in real-time when prompted to allow for interactive use. To train our model, we need a diverse, large-scale source of **data**. Unfortunately, there is no web-scale data source for segmentation; to address this, we build a "data engine", *i.e*., we iterate between using our efficient model to assist in data collection and using the newly collected data to improve the model. We introduce each interconnected component next, followed by the dataset we created and the experiments that demonstrate the effectiveness of our approach.

Task (§2). In NLP and more recently computer vision, foundation models are a promising development that can perform zero-shot and few-shot learning for new datasets and tasks often by using "prompting" techniques. Inspired by this line of work, we propose the *promptable segmentation task*, where the goal is to return a *valid* segmentation mask given any segmentation *prompt* (see Fig. 1a). A prompt simply specifies what to segment in an image, *e.g*., a prompt can include spatial or text information identifying an object. The requirement of a valid output mask means that even when a prompt is ambiguous and could refer to multiple objects (for example, a point on a shirt may indicate either the shirt or the person wearing it), the output should be a reasonable mask for at least one of those objects. We use the promptable segmentation task as both a pre-training objective and to solve general downstream segmentation tasks via prompt engineering.

Model (§3). The promptable segmentation task and the goal of real-world use impose constraints on the model architecture. In particular, the model must support *flexible prompts*, needs to compute masks in amortized *real-time* to allow interactive use, and must be *ambiguity-aware*. Surprisingly, we find that a simple design satisfies all three constraints: a powerful image encoder computes an image embedding, a prompt encoder embeds prompts, and then the two information sources are combined in a lightweight mask decoder that predicts segmentation masks. We refer to this model as the Segment Anything Model, or SAM (see Fig. 1b). By separating SAM into an image encoder and a fast prompt encoder / mask decoder, the same image embedding can be reused (and its cost amortized) with different prompts. Given an image embedding, the prompt encoder and mask decoder predict a mask from a prompt in ∼50ms in a web browser. We focus on point, box, and mask prompts, and also present initial results with free-form text prompts. To make SAM ambiguity-aware, we design it to predict multiple masks for a single prompt allowing SAM to naturally handle ambiguity, such as the shirt *vs*. person example.

Data engine (§4). To achieve strong generalization to new data distributions, we found it necessary to train SAM on a large and diverse set of masks, beyond any segmentation dataset that already exists. While a typical approach for foundation models is to obtain data online [82], masks are not naturally abundant and thus we need an alternative strategy. Our solution is to build a "data engine", *i.e*., we co-develop our model with model-in-the-loop dataset annotation (see Fig. 1c). Our data engine has three stages: *assisted-manual*, *semi-automatic*, and *fully automatic*. In the first stage, SAM assists annotators in annotating masks, similar to a classic interactive segmentation setup. In the second stage, SAM can automatically generate masks for a subset of objects by prompting it with likely object locations and annotators focus on annotating the remaining objects, helping increase mask diversity. In the final stage, we prompt SAM with a regular grid of foreground points, yielding on average ∼100 high-quality masks per image.

Dataset (§5). Our final dataset, SA-1B, includes more than *1B* masks from *11M* licensed and privacy-preserving images (see Fig. 2). SA-1B, collected fully automatically using the final stage of our data engine, has 400× more masks than any existing segmentation dataset [66, 44, 117, 60], and as we verify extensively, the masks are of high quality and diversity. Beyond its use in training SAM to be robust and general, we hope SA-1B becomes a valuable resource for research aiming to build new foundation models.

Responsible AI (§6). We study and report on potential fairness concerns and biases when using SA-1B and SAM. Images in SA-1B span a geographically and economically diverse set of countries and we found that SAM performs similarly across different groups of people. Together, we hope this will make our work more equitable for real-world use cases. We provide model and dataset cards in the appendix.

Experiments (§7). We extensively evaluate SAM. First, using a diverse new suite of 23 segmentation datasets, we find that SAM produces high-quality masks from a single foreground point, often only slightly below that of the manually annotated ground truth. Second, we find consistently strong quantitative and qualitative results on a variety of downstream tasks under a zero-shot transfer protocol using prompt engineering, including edge detection, object proposal generation, instance segmentation, and a preliminary exploration of text-to-mask prediction. These results suggest that SAM can be used out-of-the-box with prompt engineering to solve a variety of tasks involving object and image distributions beyond SAM's training data. Nevertheless, room for improvement remains, as we discuss in §8.

Release. We are releasing the SA-1B dataset for research purposes and making SAM available under a permissive open license (Apache 2.0) at https://segment-anything.com. We also showcase SAM's capabilities with an online demo.