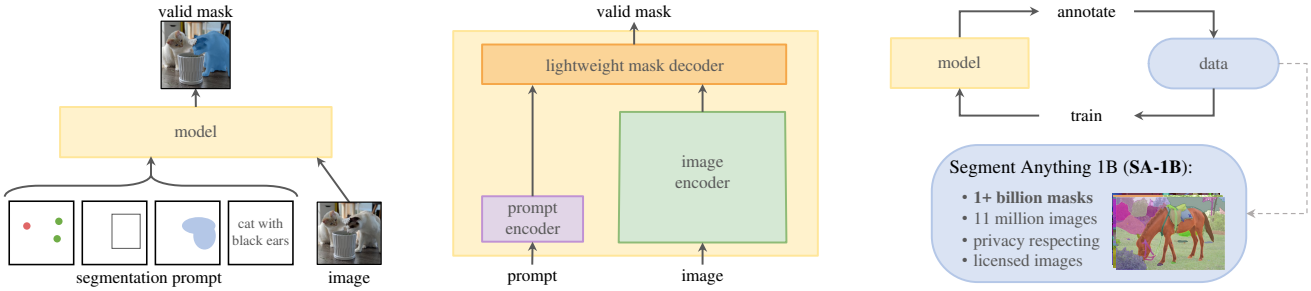
Segment Anything

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.

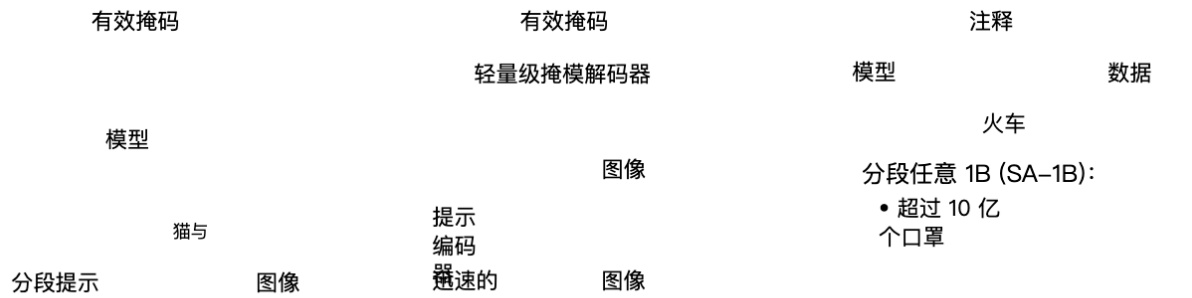
In this work, our goal is to build *a foundation model for image segmentation*. That is, we seek to develop a promptable model and pre-train it on a broad dataset using a task that enables powerful generalization. With this model, we aim to solve a range of downstream segmentation problems on new data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

分割任何东西

亚历山大· 基里洛夫 埃里克· 明顿 尼基拉· 拉维 汉字毛 克洛伊· 罗兰 劳拉· 古斯塔夫森
肖特特 斯宾塞· 怀特海德 亚历山大· 伯格 罗万彦 皮奥特· 多尔 罗斯· 吉尔希克
[1]项目负责人 [2]共同第一作者 [3]平等贡献 [4]定向引导
元人工智能研究，FAIR



(a) 任务：及时分割 (b) 模型：分段任意模型（SAM） (c) 数据：数据引擎（上）和数据集（下）

图 1：我们的目标是通过引入三个相互关联的组件来构建分割的基础模型：可提示的分割任务、支持数据注释并通过即时工程零样本传输到一系列任务的分割模型 (SAM)，以及用于收集 SA- 1B 的数据引擎，SA- 1B 是我们包含超过 10 亿个口罩的数据集。

抽象的

我们介绍 *Segment Anything (SA)* 项目：用于图像分割的新任务、模型和数据集。在数据收集循环中使用我们的高效模型，我们构建了迄今为止（迄今为止）最大的分割数据集，在 *1100* 万张许可且尊重隐私的图像上包含超过 *10* 亿个掩模。该模型的设计和训练具有快速性，因此它可以将零样本转移到新的图像分布和任务。我们评估了它在众多任务上的能力，发现它的零样本性能令人印象深刻——通常可以与之前完全监督的结果相媲美甚至优于之前的结果。我们将在 *https://segment-anything.com* 上发布 *Segment Anything Model (SAM)* 以及相应的 *1B* 掩模和 *11M* 图像数据集 *(SA-1B)*，以促进对计算机视觉基础模型的研究。

一、简介

在网络规模数据集上预训练的大型语言模型正在通过强大的零样本和少样本泛化彻底改变 NLP [10]。这些“基础模型”[8] 可以推广到训练期间所见之外的任务和数据分布。此功能通常通过提示工程来实现，其中使用手工制作的文本来提示语言模型为手头的任务生成有效的文本响应。当使用来自网络的丰富文本语料库进行扩展和训练时，这些模型的零样本和少样本性能与（甚至在某些情况下匹配）微调模型 [10, 21]。经验趋势表明，这种行为随着模型规模、数据集大小和总训练计算而改善 [56,10,21,51]。

计算机视觉领域也对基础模型进行了探索，尽管程度较小。也许最突出的插图对齐了来自网络的配对文本和图像。例如，CLIP [82] 和 ALIGN [55] 使用对比学习来训练对齐两种模式的文本和图像编码器。经过训练后，设计的文本提示可以零样本概括新的视觉概念和数据分布。这种编码器还可以与其他模块有效组合，以实现下游任务，例如图像生成（例如，DALL·E [83]）。尽管在视觉和语言编码器方面取得了很大进展，但计算机视觉还包含超出此范围的广泛问题，并且对于其中许多问题，并不存在丰富的训练数据。在这项工作中，我们的目标是建立图像分割的基础模型。也就是说，我们寻求开发一个可提示的模型，并使用能够实现强大泛化的任务在广泛的数据集上对其进行预训练。通过该模型，我们的目标是使用即时工程解决新数据分布上的一系列下游分割问题。

该计划的成功取决于三个组成部分：任务、模型和数据。为了开发它们，我们解决了以下有关图像分割的问题：

1. 什么任务可以实现零样本泛化？ 2. 对应的模型架构是什么？ 3. 哪些数据可以为该任务和模型提供支持？

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.

这些问题错综复杂，需要综合解决。我们首先定义一个可提示的分割任务，该任务足够通用，可以提供强大的预训练目标并支持广泛的下游应用程序。该任务需要一个模型支持灵活的提示，并能在提示时实时输出分段掩码以允许交互使用。为了训练我们的模型，我们需要多样化的大规模数据源。不幸的是，没有用于分割的网络规模数据源；为了解决这个问题，我们构建了一个“数据引擎”，即我们在使用高效模型协助数据收集和使用新收集的数据改进模型之间进行迭代。接下来我们介绍每个互连的组件，然后是我们创建的数据集以及证明我们方法有效性的实验。

任务（**§2**）。在 **NLP** 和最近的计算机视觉中，基础模型是一种很有前景的发展，通常可以通过使用“提示”技术对新数据集和任务执行零样本和少样本学习。受这一工作的启发，我们提出了提示分割任务，其目标是在给定任何分割提示的情况下返回有效的分割掩码（见图**1a**）。提示只是指定在图像中分割什么，例如，提示可以包括识别对象的空间或文本信息。有效输出掩码的要求意味着，即使提示不明确并且可能引用多个对象（例如，衬衫上的点可能表示衬衫或穿着它的人），输出也应该是合理的掩码至少其中一个对象。我们使用提示分割任务作为预训练目标，并通过提示工程解决一般下游分割任务。

模型（**§3**）。及时的分割任务和实际使用的目标对模型架构施加了约束。特别是，该模型必须支持灵活的提示，需要实时摊销计算掩码以允许交互使用，并且必须具有模糊性意识。令人惊讶的是，我们发现一个简单的设计满足了所有三个约束：强大的图像编码器计算图像嵌入，提示编码器嵌入提示，然后将两个信息源组合在预测分割掩模的轻量级掩模解码器中。我们将该模型称为分段任意模型**(SAM)**（见图 **1b**）。通过将 **SAM** 分成图像编码器和快速提示编码器/ 掩模解码器，可以通过不同的提示重复使用相同的图像嵌入（及其成本摊销）。给定图像嵌入，提示编码器和掩码解码器在网络浏览器中的**~50** 毫秒内根据提示预测掩码。我们专注于点、框和蒙版提示，并且还使用自由格式的文本提示来呈现初步结果。为了使 **SAM** 能够感知歧义，我们将其设计为预测单个提示的多个掩码，从而使 **SAM** 能够自然地处理歧义，例如衬衫与人的示例。

数据引擎（**§4**）。为了实现对新数据分布的强泛化，我们发现有必要在大量且多样化的掩模上训练 **SAM**，超出任何现有的分割数据集。虽然基础模型的典型方法是在线获取数据[**82**]，但掩码并不丰富，因此我们需要一种替代策略。我们的解决方案是构建一个“数据引擎”，即我们与模型在环数据集注释共同开发我们的模型（见图**1c**）。我们的数据引擎分为三个阶段：辅助手动、半自动和全自动。在第一阶段，**SAM** 协助注释者注释掩模，类似于经典的交互式分割设置。在第二阶段，**SAM** 可以通过提示可能的对象位置来自动为对象子集生成掩码，而注释器则专注于注释其余对象，从而帮助增加掩码多样性。在最后阶段，我们使用前景点的规则网格提示 **SAM**，平均每张图像产生约 **1**00 个高质量掩模。

数据集（**§5**）。我们的最终数据集 **SA- 1B** 包含来自 **11**00 万张许可和隐私保护图像的超过 **1B** 个掩模（见图 **2**）。**SA- 1B** 使用我们数据引擎的最后阶段完全自动收集，其掩码比任何现有分割数据集 [**66,44,117,6**0] 多 **4**00 倍，并且正如我们广泛验证的那样，掩码具有高质量和多样性。除了用于训练 **SAM** 使其变得稳健和通用之外，我们希望 **SA- 1B** 成为旨在构建新基础模型的研究的宝贵资源。

负责任的人工智能（**§6**）。我们研究并报告使用 **SA- 1B** 和 **SAM** 时潜在的公平性问题和偏见。 **SA- 1B** 中的图像跨越了地理和经济上不同的国家，我们发现 **SAM** 在不同人群中的表现相似。我们共同希望这将使我们的工作在现实世界的用例中更加公平。我们在附录中提供模型和数据集卡。

实验（**§7**）。我们广泛评估 **SAM**。首先，使用由 **23** 个分割数据集组成的多样化新套件，我们发现 **SAM** 从单个前景点生成高质量掩模，通常仅略低于手动注释的地面事实。其次，我们使用即时工程在零镜头传输协议下的各种下游任务上发现了一致的强大定量和定性结果，包括边缘检测、对象建议生成、实例分割以及文本到掩模预测的初步探索。这些结果表明，**SAM** 可以通过快速工程开箱即用，解决涉及 **SAM** 训练数据之外的对象和图像分布的各种任务。尽管如此，正如我们在第 **8** 节中讨论的那样，改进的空间仍然存在。

发布。我们将出于研究目的发布 **SA- 1B** 数据集，并在许可的情况下提供 **SAM**。我们还通过在线演示展示 **SAM** 的功能。

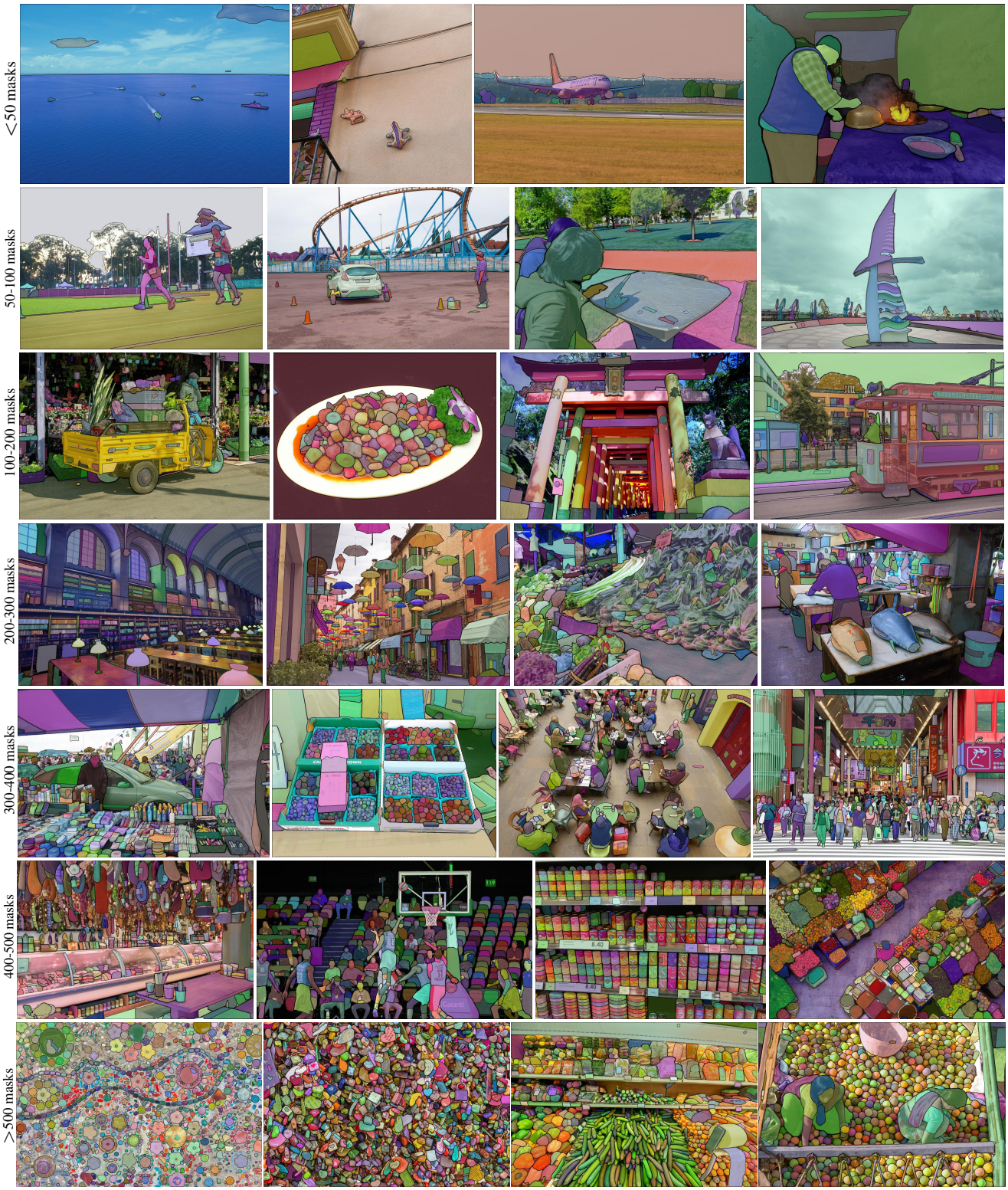


Figure 2: Example images with overlaid masks from our newly introduced dataset, **SA-1B**. SA-1B contains 11M diverse, high-resolution, licensed, and privacy protecting images and 1.1B high-quality segmentation masks. These masks were annotated *fully automatically* by SAM, and as we verify by human ratings and numerous experiments, are of high quality and diversity. We group images by number of masks per image for visualization (there are ∼100 masks per image on average).

图 2：来自我们新引入的数据集 SA- 1B 的带有叠加掩模的示例图像。 SA- 1B包含11M多样化、高分辨率、许可和隐私保护图像和1.1B高质量分割掩模。这些掩模由 SAM 完全自动注释，并且正如我们通过人工评级和大量实验验证的那样，它们具有高质量和多样性。我们根据每个图像的掩模数量对图像进行分组以进行可视化（平均每个图像有 ~100 个掩模）。

2. Segment Anything Task

We take inspiration from NLP, where the next token prediction task is used for foundation model pre-training *and* to solve diverse downstream tasks via prompt engineering [10]. To build a foundation model for segmentation, we aim to define a task with analogous capabilities.

Task. We start by translating the idea of a prompt from NLP to segmentation, where a prompt can be a set of foreground / background points, a rough box or mask, free-form text, or, in general, any information indicating what to segment in an image. The *promptable segmentation task*, then, is to return a *valid* segmentation mask given any *prompt*. The requirement of a "valid" mask simply means that even when a prompt is *ambiguous* and could refer to multiple objects (*e.g.*, recall the shirt *vs.* person example, and see Fig. 3), the output should be a reasonable mask for at least *one* of those objects. This requirement is similar to expecting a language model to output a coherent response to an ambiguous prompt. We choose this task because it leads to a natural pre-training algorithm *and* a general method for zero-shot transfer to downstream segmentation tasks via prompting.

Pre-training. The promptable segmentation task suggests a natural pre-training algorithm that simulates a sequence of prompts (*e.g.*, points, boxes, masks) for each training sample and compares the model's mask predictions against the ground truth. We adapt this method from interactive segmentation [109, 70], although unlike interactive segmentation whose aim is to eventually predict a valid mask after enough user input, our aim is to always predict a *valid mask* for *any prompt* even when the prompt is *ambiguous*. This ensures that a pre-trained model is effective in use cases that involve ambiguity, including automatic annotation as required by our data engine §4. We note that performing well at this task is challenging and requires specialized modeling and training loss choices, which we discuss in §3.

Zero-shot transfer. Intuitively, our pre-training task endows the model with the ability to respond appropriately to any prompt at inference time, and thus downstream tasks can be solved by engineering appropriate prompts. For example, if one has a bounding box detector for cats, cat instance segmentation can be solved by providing the detector's box output as a prompt to our model. In general, a wide array of practical segmentation tasks can be cast as prompting. In addition to automatic dataset labeling, we explore five diverse example tasks in our experiments in §7.

Related tasks. Segmentation is a broad field: there's interactive segmentation [57, 109], edge detection [3], super pixelization [85], object proposal generation [2], foreground segmentation [94], semantic segmentation [90], instance segmentation [66], panoptic segmentation [59], *etc.* The goal of our promptable segmentation task is to produce a broadly capable model that can adapt to *many* (though not all) existing and *new* segmentation tasks via prompt engineering. This capability is a form of task generalization [26]. Note that this is different than previous work on multi-task segmentation systems. In a multi-task system, a single model performs a *fixed* set of tasks, *e.g.*, joint semantic, instance, and panoptic segmentation [114, 19, 54], but the training and test tasks are the same. An important distinction in our work is that a model trained for promptable segmentation can perform a new, different task at inference time by acting as a *component* in a larger system, *e.g.*, to perform instance segmentation, a promptable segmentation model is *combined* with an existing object detector.

Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

Discussion. Prompting and composition are powerful tools that enable a single model to be used in extensible ways, potentially to accomplish tasks unknown at the time of model design. This approach is analogous to how other foundation models are used, *e.g.*, how CLIP [82] is the text-image alignment component of the DALL·E [83] image generation system. We anticipate that composable system design, powered by techniques such as prompt engineering, will enable a wider variety of applications than systems trained specifically for a fixed set of tasks. It's also interesting to compare promptable and interactive segmentation through the lens of composition: while interactive segmentation models are designed with human users in mind, a model trained for promptable segmentation can also be composed into a larger algorithmic system as we will demonstrate.

2. 对任何任务进行分段

我们从 NLP 中获得灵感，其中下一个令牌预测任务用于基础模型预训练，并通过即时工程解决各种下游任务 [10]。为了构建分割的基础模型，我们的目标是定义具有类似功能的任务。

任务。我们首先将提示的概念从 NLP 转化为分割，其中提示可以是一组前景/背景点、一个粗略的框或掩模、自由格式的文本，或者一般来说，指示要分割的内容的任何信息一个图像。那么，可提示的分割任务是在给定任何提示的情况下返回有效的分割掩码。“有效”掩码的要求仅仅意味着，即使提示不明确并且可能引用多个对象（例如，回想一下衬衫与人的示例，请参见图 3），输出也应该是一个合理的掩码这些物体中至少有一个。此要求类似于期望语言模型对不明确的提示输出连贯的响应。我们选择这个任务是因为它产生了一种自然的预训练算法和一种通过提示零样本转移到下游分割任务的通用方法。

预训练。可提示的分割任务提出了一种自然的预训练算法，该算法模拟每个训练样本的一系列提示（例如，点、框、掩模），并将模型的掩模预测与真实情况进行比较。我们从交互式分割[109, 70]中改编了这种方法，尽管与交互式分割的目标是在足够的用户输入后最终预测有效掩码不同，我们的目标是始终为任何提示预测有效掩码，即使提示不明确。这确保了预训练模型在涉及歧义的用例中是有效的，包括我们的数据引擎§4所需的自动注释。我们注意到，在这项任务中表现良好具有挑战性，需要专门的建模和训练损失选择，我们将在第 3 节中讨论。

零射击转移。直观地说，我们的预训练任务赋予模型在推理时对任何提示做出适当响应的能力，因此可以通过设计适当的提示来解决下游任务。例如，如果有一个针对猫的边界框检测器，则可以通过提供检测器的框输出作为我们模型的提示来解决猫实例分割问题。一般来说，大量的实际分割任务都可以作为提示。除了自动数据集标记之外，我们还在第 7 节的实验中探索了五个不同的示例任务。

相关任务。分割是一个广阔的领域：有交互式分割[57，109]、边缘检测[3]、超像素化[85]、对象建议生成[2]、前景分割[94]、语义分割[90]、实例分割[66]]、全景分割[59]等

我们的提示分割任务的目标是生成图 3：每列显示 SAM 根据单个模糊点提示（绿色圆圈）生成的 3 个有效掩码。

一个功能广泛的模型，可以通过即时工程适应许多（尽管不是全部）现有的和新的分割任务。这种能力是任务泛化的一种形式[26]。请注意，这与之前关于多任务分割系统的工作不同。在多任务系统中，单个模型执行一组固定的任务，例如联合语义、实例和全景分割[114,19,54]，但训练和测试任务是相同的。我们工作中的一个重要区别是，为可提示分割训练的模型可以通过充当更大系统中的组件来在推理时执行新的、不同的任务，例如，为了执行实例分割，可提示分割模型与现有的分割模型相结合。物体探测器。

讨论。提示和组合是强大的工具，使单个模型能够以可扩展的方式使用，有可能完成模型设计时未知的任务。这种方法类似于其他基础模型的使用方式，例如 CLIP[82] 是 DALL·E[83] 图像生成系统的文本图像对齐组件。我们预计，由即时工程等技术支持的可组合系统设计将比专门为一组固定任务训练的系统实现更广泛的应用。通过组合的角度来比较即时分割和交互式分割也很有趣：虽然交互式分割模型是在考虑人类用户的情况下设计的，但经过训练的即时分割模型也可以组合成更大的算法系统，正如我们将演示的那样。

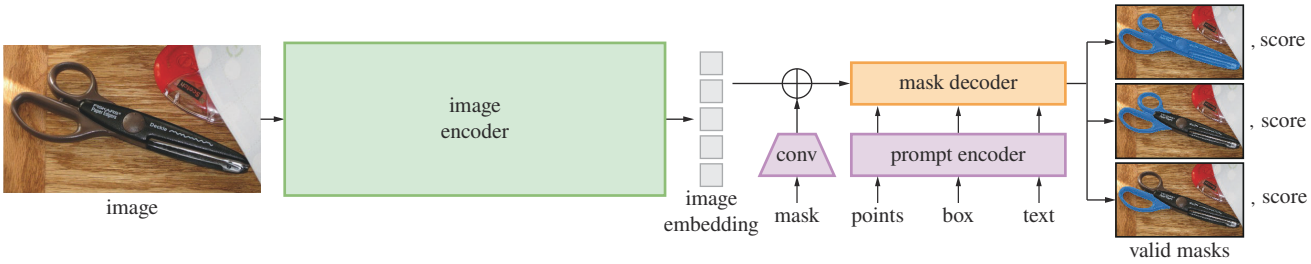


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

3. Segment Anything Model

We next describe the Segment Anything Model (SAM) for promptable segmentation. SAM has three components, illustrated in Fig. 4: an image encoder, a flexible prompt encoder, and a fast mask decoder. We build on Transformer vision models [14, 33, 20, 62] with specific tradeoffs for (amortized) real-time performance. We describe these components at a high-level here, with details in §A.

Image encoder. Motivated by scalability and powerful pre-training methods, we use an MAE [47] pre-trained Vision Transformer (ViT) [33] minimally adapted to process high resolution inputs [62]. The image encoder runs once per image and can be applied prior to prompting the model.

Prompt encoder. We consider two sets of prompts: *sparse* (points, boxes, text) and *dense* (masks). We represent points and boxes by positional encodings [95] summed with learned embeddings for each prompt type and free-form text with an off-the-shelf text encoder from CLIP [82]. Dense prompts (*i.e.*, masks) are embedded using convolutions and summed element-wise with the image embedding.

Mask decoder. The mask decoder efficiently maps the image embedding, prompt embeddings, and an output token to a mask. This design, inspired by [14, 20], employs a modification of a Transformer decoder block [103] followed by a dynamic mask prediction head. Our modified decoder block uses prompt self-attention and cross-attention in two directions (prompt-to-image embedding and vice-versa) to update *all* embeddings. After running two blocks, we upsample the image embedding and an MLP maps the output token to a dynamic linear classifier, which then computes the mask foreground probability at each image location.

Resolving ambiguity. With one output, the model will average multiple valid masks if given an ambiguous prompt. To address this, we modify the model to predict multiple output masks for a single prompt (see Fig. 3). We found 3 mask outputs is sufficient to address most common cases (nested masks are often at most three deep: whole, part, and subpart). During training, we backprop only the minimum loss [15, 45, 64] over masks. To rank masks, the model predicts a confidence score (*i.e.*, estimated IoU) for each mask.

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 runtime performance enables seamless, real-time interactive prompting of our model.

Losses and training. We supervise mask prediction with the linear combination of focal loss [65] and dice loss [73] used in [14]. We train for the promptable segmentation task using a mixture of geometric prompts (for text prompts see §7.5). Following [92, 37], we simulate an interactive setup by randomly sampling prompts in 11 rounds per mask, allowing SAM to integrate seamlessly into our data engine.

4. Segment Anything Data Engine

As segmentation masks are not abundant on the internet, we built a data engine to enable the collection of our 1.1B mask dataset, SA-1B. The data engine has three stages: (1) a model-assisted manual annotation stage, (2) a semi-automatic stage with a mix of automatically predicted masks and model-assisted annotation, and (3) a fully automatic stage in which our model generates masks without annotator input. We go into details of each next.

Assisted-manual stage. In the first stage, resembling classic interactive segmentation, a team of professional annotators labeled masks by clicking foreground / background object points using a browser-based interactive segmentation tool powered by SAM. Masks could be refined using pixel-precise "brush" and "eraser" tools. Our model-assisted annotation runs in real-time directly inside a browser (using precomputed image embeddings) enabling a truly interactive experience. We did not impose semantic constraints for labeling objects, and annotators freely labeled both "stuff" and "things" [1]. We suggested annotators label objects they could name or describe, but did not collect these names or descriptions. Annotators were asked to label objects in order of prominence and were encouraged to proceed to the next image once a mask took over 30 seconds to annotate.

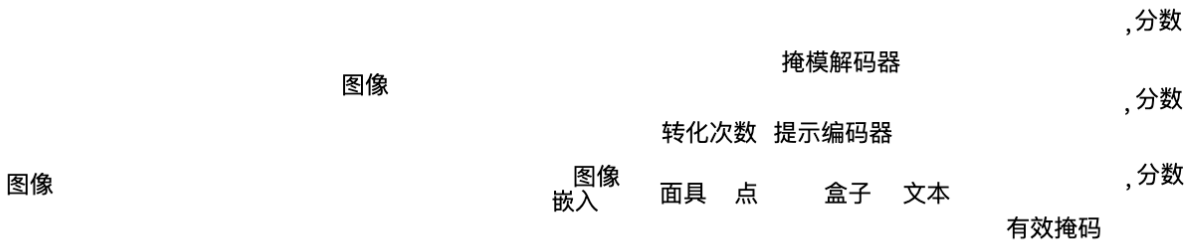


图 4：分段任意模型 (SAM) 概述。重量级图像编码器输出图像嵌入，然后可以通过各种输入提示有效地查询该图像嵌入，以摊销实时速度生成对象蒙版。对于对应于多个对象的不明确提示，SAM 可以输出多个有效掩码和相关的置信度分数。

3. 分割任何模型

接下来我们描述用于快速分割的分段任意模型（SAM）。SAM 具有三个组件，如图 4 所示：图像编码器、灵活提示编码器和快速掩模解码器。我们建立在 Transformer 视觉模型 [14,33,20,62] 的基础上，并针对（摊销的）实时性能进行了特定的权衡。我们在这里概括地描述了这些组件，并在§A 中提供了详细信息。

图像编码器。受可扩展性和强大的预训练方法的推动，我们使用 **MAE** [**47**] 预训练的视觉变换器 **(ViT)** [**33**] 最低限度地适应处理高分辨率输入 [**62**]。图像编码器每张图像运行一次，并且可以在提示模型之前应用。

提示编码器。我们考虑两组提示：稀疏（点、框、文本）和密集（掩模）。我们代表

通过位置编码 [95] 计算点和框，并使用 CLIP [82] 中现成的文本编码器对每种提示类型和自由格式文本进行学习嵌入。使用卷积嵌入密集提示（即掩码），并与图像嵌入按元素求和。

掩码解码器。掩码解码器有效地将图像嵌入、提示嵌入和输出标记映射到掩码。该设计受到 [**14, 2**0] 的启发，采用了 **Transformer** 解码器块 [**103**] 的修改，后跟动态掩模预测头。我们修改后的解码器块使用两个方向的即时自注意力和交叉注意力（即时图像嵌入，反之亦然）来更新所有嵌入。运行两个块后，我们对图像嵌入进行上采样，并且 **MLP** 将输出标记映射到动态线性分类器，然后计算每个图像位置的掩模前景概率。

解决歧义。对于一个输出，如果给出不明确的提示，模型将对多个有效掩码进行平均。为了解决这个问题，我们修改模型以预测单个提示的多个输出掩码（见图 **3**）。我们发现 **3** 个掩码输出足以解决最常见的情况（嵌套掩码通常最多三层深度：整体、部分和子部分）。在训练中

我们仅反向传播掩模上的最小损失[15,45,64]。为了对掩模进行排名，模型预测每个掩模的置信度得分（即估计的 IoU）。

效率。整体模型设计很大程度上是出于效率的考虑。给定预先计算的图像嵌入，提示编码器和掩码解码器在 **Web** 浏览器中的 **CPU** 上运行，耗时约 **5**0 毫秒。这种运行时性能可以为我们的模型提供无缝、实时的交互式提示。

损失和培训。我们使用[**14**]中使用的焦点损失[**65**]和骰子损失[**73**]的线性组合来监督掩模预测。我们使用几何提示的混合来训练可提示的分割任务（有关文本提示，请参见第 **7.5** 节）。遵循 [9**2, 37**]，我们通过在每个掩码 **11** 轮中随机采样提示来模拟交互式设置，从而使 **SAM** 能够无缝集成到我们的数据引擎中。

4. 分段任何数据引擎

由于互联网上的分割掩模并不丰富，我们构建了一个数据引擎来收集我们的 1.1B 掩模数据集 SA- 1B。数据引擎分为三个阶段：（1）模型辅助手动注释阶段，（2）混合自动预测掩模和模型辅助注释的半自动阶段，以及（3）全自动阶段模型无需注释器输入即可生成掩模。接下来我们将详细介绍每一项。

辅助手动阶段。在第一阶段，类似于经典的交互式分割，专业注释者团队使用由 **SAM** 提供支持的基于浏览器的交互式分割工具，通过单击前景/ 背景对象点来标记蒙版。可以使用像素精确的“画笔”和“橡皮擦”工具来细化蒙版。我们的模型辅助注释直接在浏览器内实时运行（使用预先计算的图像嵌入），从而实现真正的交互式体验。我们没有对标记对象施加语义约束，注释者可以自由地标记“东西”和“事物”[**1**]。我们建议注释者标记他们可以命名或描述的对象，但不收集这些名称或描述。注释者被要求按照突出的顺序标记对象，并被鼓励在蒙版注释时间超过 **3**0 秒后继续处理下一张图像。

At the start of this stage, SAM was trained using common public segmentation datasets. After sufficient data annotation, SAM was retrained using only newly annotated masks. As more masks were collected, the image encoder was scaled from ViT-B to ViT-H and other architectural details evolved; in total we retrained our model 6 times. Average annotation time per mask decreased from 34 to 14 seconds as the model improved. We note that 14 seconds is 6.5× faster than mask annotation for COCO [66] and only 2× slower than bounding-box labeling with extreme points [76, 71]. As SAM improved, the average number of masks per image increased from 20 to 44 masks. Overall, we collected 4.3M masks from 120k images in this stage.

Semi-automatic stage. In this stage, we aimed to increase the *diversity* of masks in order to improve our model's ability to segment anything. To focus annotators on less prominent objects, we first automatically detected confident masks. Then we presented annotators with images prefilled with these masks and asked them to annotate any additional unannotated objects. To detect confident masks, we trained a bounding box detector [84] on all first stage masks using a generic "object" category. During this stage we collected an additional 5.9M masks in 180k images (for a total of 10.2M masks). As in the first stage, we periodically retrained our model on newly collected data (5 times). Average annotation time per mask went back up to 34 seconds (excluding the automatic masks) as these objects were more challenging to label. The average number of masks per image went from 44 to 72 masks (including the automatic masks).

Fully automatic stage. In the final stage, annotation was *fully automatic*. This was feasible due to two major enhancements to our model. First, at the start of this stage, we had collected enough masks to greatly improve the model, including the diverse masks from the previous stage. Second, by this stage we had developed the ambiguity-aware model, which allowed us to predict valid masks even in ambiguous cases. Specifically, we prompted the model with a 32×32 regular grid of points and for each point predicted a set of masks that may correspond to valid objects. With the ambiguity-aware model, if a point lies on a part or subpart, our model will return the subpart, part, and whole object. The IoU prediction module of our model is used to select *confident* masks; moreover, we identified and selected only *stable* masks (we consider a mask stable if thresholding the probability map at $0.5 - \delta$ and $0.5 + \delta$ results in similar masks). Finally, after selecting the confident and stable masks, we applied non-maximal suppression (NMS) to filter duplicates. To further improve the quality of smaller masks, we also processed multiple overlapping zoomed-in image crops. For further details of this stage, see §B. We applied fully automatic mask generation to all 11M images in our dataset, producing a total of 1.1B high-quality masks. We describe and analyze the resulting dataset, SA-1B, next.

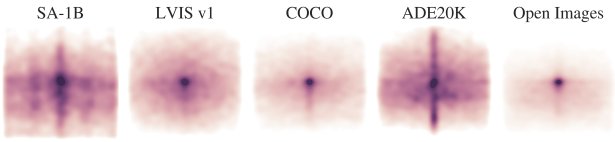


Figure 5: Image-size normalized mask center distributions.

5. Segment Anything Dataset

Our dataset, SA-1B, consists of 11M diverse, high-resolution, licensed, and privacy protecting images and 1.1B high-quality segmentation masks collected with our data engine. We compare SA-1B with existing datasets and analyze mask quality and properties. We are releasing SA-1B to aid future development of foundation models for computer vision. We note that SA-1B will be released under a favorable license agreement for certain research uses and with protections for researchers.

Images. We licensed a new set of 11M images from a provider that works directly with photographers. These images are high resolution (3300×4950 pixels on average), and the resulting data size can present accessibility and storage challenges. Therefore, we are releasing downsampled images with their shortest side set to 1500 pixels. Even after downsampling, our images are significantly higher resolution than many existing vision datasets (*e.g.*, COCO [66] images are ~480×640 pixels). Note that most models today operate on much lower resolution inputs. Faces and vehicle license plates have been blurred in the released images.

Masks. Our data engine produced 1.1B masks, 99.1% of which were generated fully automatically. Therefore, the quality of the automatic masks is centrally important. We compare them directly to professional annotations and look at how various mask properties compare to prominent segmentation datasets. Our main conclusion, as borne out in the analysis below and the experiments in §7, is that our automatic masks are high quality and effective for training models. Motivated by these findings, SA-1B *only includes automatically generated masks.*

Mask quality. To estimate mask quality, we randomly sampled 500 images (~50k masks) and asked our professional annotators to improve the quality of all masks in these images. Annotators did so using our model and pixel-precise "brush" and "eraser" editing tools. This procedure resulted in pairs of automatically predicted and professionally corrected masks. We computed IoU between each pair and found that 94% of pairs have greater than 90% IoU (and 97% of pairs have greater than 75% IoU). For comparison, prior work estimates inter-annotator consistency at 85-91% IoU [44, 60]. Our experiments in §7 confirm by human ratings that mask quality is high relative to a variety of datasets and that training our model on automatic masks is nearly as good as using all masks produced by the data engine.

在此阶段开始时，使用常见的公共分割数据集对 SAM 进行训练。经过足够的数据注释后，仅使用新注释的掩模重新训练 SAM。随着收集到的掩模越来越多，图像编码器从 ViT- B 扩展到 ViT- H，并且其他架构细节也不断发展；我们总共重新训练了模型 6 次。 AV-

每个掩码的平均注释时间从 34 减少到 14

模型改进后的秒数。我们注意到，14 秒比 COCO [66] 的掩模注释快 6.5 倍，仅比带有极值点的边界框标记 [76，71] 慢 2 倍。随着 SAM 的改进，平均数量

每个图像的蒙版从 20 个增加到 44 个蒙版。总的来说，我们在此阶段从 12 万张图像中收集了 430 万个掩模。

半自动阶段。在这个阶段，我们的目标是增加面具的多样性，以提高我们的模型分割任何东西的能力。为了将注释者集中在不太突出的对象上，我们首先自动检测置信蒙版。然后，我们向注释者提供了预先填充了这些蒙版的图像，并要求他们注释任何其他未注释的对象。为了检测置信的掩模，我们使用通用的“对象”类别在所有第一阶段掩模上训练了边界框检测器[**84**]。在此阶段，我们在 **180k** 图像中额外收集了 **5**90 万个掩模（总共 **1**020 万个掩模）。与第一阶段一样，我们定期根据新收集的数据重新训练我们的模型（**5** 次）。平均注释-

每个掩模的处理时间回升至 34 秒（不包括自动掩模），因为这些物体标记起来更具挑战性。每张图像的平均掩模数量

44 至 72 个面罩（包括自动面罩）。

全自动舞台。在最后阶段，注释是全自动的。由于我们的模型有两项重大改进，这是可行的。首先，在这个阶段开始时，我们收集了足够的掩模来极大地改进模型，包括上一阶段的各种掩模。其次，到这个阶段，我们已经开发了歧义感知模型，即使在歧义情况下，它也使我们能够预测有效的掩码。具体来说，我们使用 **32** ×**32** 规则点网格提示模型，并为每个点预测一组可能对应于有效对象的掩模。使用歧义感知模型，如果一个点位于部分或子部分上，我们的模型将返回子部分、部分和整个对象。我们模型的 **IoU** 预测模块用于选择置信掩码；此外，我们仅识别并选择稳定的掩模（如果阈值- 我们认为掩模是稳定的）

将概率图绘制在 0.5 − δ 和 0.5 + δ 处会产生类似的掩模）。最后，在选择置信且稳定的掩模后，我们应用非极大值抑制（NMS）来过滤重复项。为了进一步提高较小蒙版的质量，我们还处理了多个重叠的放大图像裁剪。有关此阶段的更多详细信息，请参阅§B。我们对数据集中的所有 1100 万张图像应用了全自动掩模生成，总共生成了 1.1B 个高质量掩模。接下来我们描述并分析生成的数据集 SA-1B。

图 5：图像大小归一化掩模中心分布。

5. 对任何数据集进行分段

我们的数据集 SA- 1B 由我们的数据引擎收集的 11M个多样化、高分辨率、许可且隐私保护的图像和 1.1B 个高质量分割掩模组成。我们将 SA- 1B 与现有数据集进行比较并分析掩模质量和特性。我们发布 SA- 1B 是为了帮助计算机视觉基础模型的未来开发。我们注意到，SA- 1B 将根据针对某些研究用途的有利许可协议发布，并为研究人员提供保护。

图片。我们从直接与摄影师合作的提供商处授权了一组新的 **11**00万张图像。这些图像具有高分辨率（平均 **33**00 ×**49**50 像素），并且生成的数据大小可能会带来可访问性和存储方面的挑战。因此，我们正在发布下采样

最短边设置为 1500 像素的图像。即使在下采样之后，我们的图像的分辨率也明显高于许多现有的视觉数据集（例如 COCO [66]

图像约为 480 ×640 像素）。请注意，当今大多数模型都在分辨率低得多的输入上运行。在发布的图像中，面孔和车牌已经模糊。

面具。我们的数据引擎生成了 **1.1B** 掩码，其中 99.**1**% 是完全自动生成的。因此，自动口罩的质量至关重要。我们直接将它们与专业注释进行比较，并研究各种掩模属性与突出的分割数据集的比较。正如下面的分析和第 **7** 节中的实验所证实的，我们的主要结论是，我们的自动掩模对于训练模型来说是高质量且有效的。受这些发现的启发，**SA- 1B** 仅包含自动生成的掩模。

面膜质量。为了估计掩模质量，我们随机采样了 **5**00 张图像（~**50k** 掩模），并要求我们的专业注释者提高这些图像中所有掩模的质量。注释者使用我们的模型和像素精确的“画笔”和“橡皮擦”编辑工具来完成此操作。该过程产生了成对的自动预测和专业校正的掩模。我们计算了每对之间的 **IoU**，发现 9**4**% 的对的 **IoU** 大于 90%（9**7**% 的对的 **IoU** 大于 **75**%）。为了进行比较，之前的工作估计注释者间的一致性为 **85 -** 9**1**%

欠条 [44, 60]。我们在第 7 节中的实验通过人类评级证实，相对于各种数据集，掩模质量较高，并且在自动掩模上训练我们的模型几乎与使用数据引擎生成的所有掩模一样好。

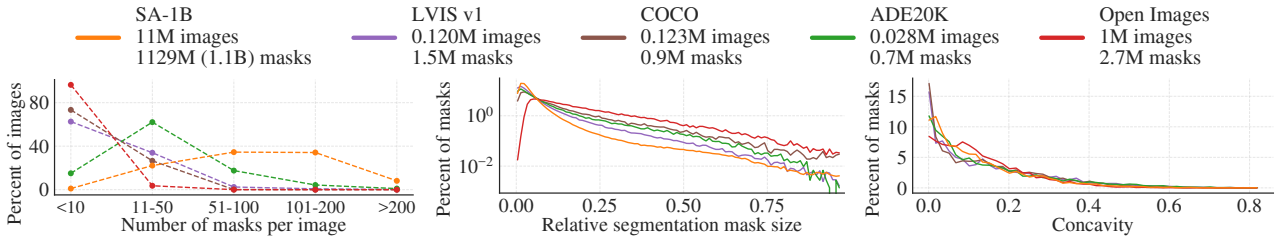


Figure 6: Dataset mask properties. The legend references the number of images and masks in each dataset. Note, that SA-1B has 11× more images and 400× more masks than the largest existing segmentation dataset Open Images [60].

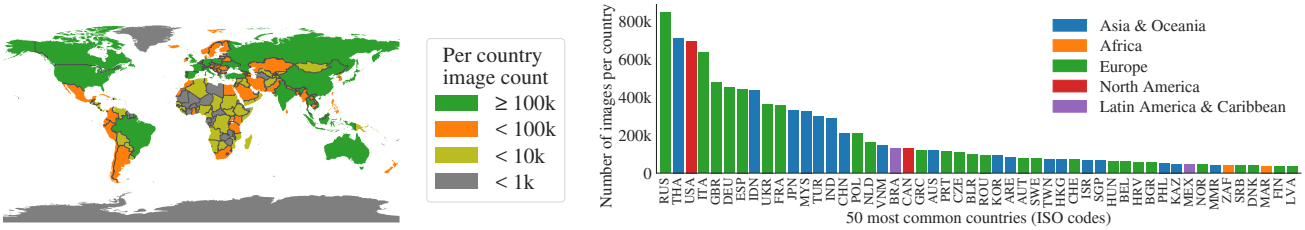


Figure 7: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.

Mask properties. In Fig. 5 we plot the spatial distribution of object centers in SA-1B compared to the largest existing segmentation datasets. Common photographer biases are present in all datasets. We observe that SA-1B has greater coverage of image corners compared to LVIS v1 [44] and ADE20K [117], the two most similarly distributed datasets, while COCO [66] and Open Images V5 [60] have a more prominent center bias. In Fig. 6 (legend) we compare these datasets by size. SA-1B has 11× more images and 400× more masks than the second largest, Open Images. On average, it has 36× more masks per image than Open Images. The closest dataset in this respect, ADE20K, still has 3.5× fewer masks per image. Fig. 6 (left) plots the masks-per-image distribution. Next, we look at image-relative mask size (square root of the mask area divided by image area) in Fig. 6 (middle). As expected, since our dataset has more masks per image, it also tends to include a greater percentage of small and medium relative-size masks. Finally, to analyze shape complexity, we look at mask concavity (1 minus mask area divided by area of mask's convex hull) in Fig. 6 (right). Since shape complexity is correlated with mask size, we control for the datasets' mask size distributions by first performing stratified sampling from binned mask sizes. We observe that the concavity distribution of our masks is broadly similar to that of other datasets.

6. Segment Anything RAI Analysis

We next perform a Responsible AI (RAI) analysis of our work by investigating potential fairness concerns and biases when using SA-1B and SAM. We focus on the geographic and income distribution of SA-1B and fairness of SAM across protected attributes of people. We also provide dataset, data annotation, and model cards in §F.

	# 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%

Table 1: Comparison of geographic and income representation. SA-1B has higher representation in Europe and Asia & Oceania as well as middle income countries. Images from Africa, Latin America & Caribbean, as well as low income countries, are underrepresented in all datasets.

Geographic and income representation. We infer the country images were photographed in using standard methods (see §C). In Fig. 7 we visualize the per-country image counts in SA-1B (left) and the 50 countries with the most images (right). We note that the top-three countries are from different parts of the world. Next, in Table 1 we compare the geographic and income representation of SA-1B, COCO [66], and Open Images [60]. SA-1B has a substantially higher percentage of images in Europe and Asia & Oceania as well as in middle income countries. All datasets underrepresent Africa as well as low income countries. We note that in SA-1B, all regions, including Africa, have at least 28 million masks, 10× more than the *total* number of masks of any previous dataset. Finally, we observe that the average number of masks per image (not shown) is fairly consistent across region and income (94-108 per image).

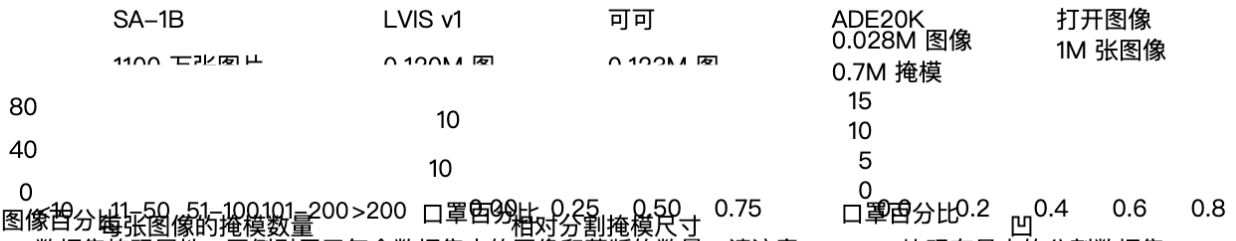


图 6：数据集掩码属性。图例引用了每个数据集中的图像和蒙版的数量。请注意，SA- 1B 比现有最大的分割数据集 Open Images [60] 多了 11 倍的图像和 400 倍的掩模。

图 7：SA- 1B 图像的估计地理分布。世界上大多数国家在SA- 1B中都有超过1000 张图像，图像最多的三个国家来自世界不同地区。

掩模特性。在图 **5** 中，我们绘制了 **SA- 1B** 中对象中心的空间分布与最大的现有分割数据集的比较。所有数据集中都存在常见的摄影师偏见。我们观察到 **SA- 1B** 具有更大的

与 LVIS v1 [44] 和 ADE20K [117]（两个分布最相似的数据集）相比，图像角点的覆盖范围，

而 COCO [66] 和 Open Images V5 [60] 有更多

中心偏向突出。在图 6（图例）中，我们按大小比较这些数据集。 SA- 1B 的图像数量比第二大 Open Images 多 11 倍，掩模数量多 400 倍。平均而言，每个图像的遮罩比 Open Images 多 36 倍。在这方面最接近的数据集 ADE20K 仍然具有 3.5 ×

每张图像的遮罩更少。图 6（左）绘制了掩模- 每像分布。接下来，我们看看图像相对掩模尺寸（掩模面积除以图像面积的平方根）

如图 6（中）所示。正如预期的那样，由于我们的数据集每个图像有更多的掩模，因此它也往往包含更大比例的中小型相对尺寸掩模。最后，为了分析形状复杂性，我们查看掩模凹度（1减去掩模面积除以掩模凸包面积）

图 6（右）。由于形状复杂性与掩模尺寸相关，因此我们通过首先对分箱掩模尺寸执行分层采样来控制数据集的掩模尺寸分布。我们观察到，我们的掩模的凹度分布与其他数据集的凹度分布大致相似。

6. 细分任何 RAI分析

接下来，我们通过调查使用 SA- 1B 和 SAM 时潜在的公平问题和偏见，对我们的工作进行负责任的 AI (RAI) 分析。我们关注 SA- 1B 的地理和收入分配以及 SAM 在受保护的人群属性中的公平性。我们还在 §F 中提供数据集、数据注释和模型卡。

	# 个国家	SA- 1B #imgs	SA- 1B #masks	% 图片 SA- 1B	% 图片 COCO	% 图片 O.I.
非洲	54	30 万	28M	2.8 %	3.0 %	1.7%
亚洲及大洋洲	70	3.9M	4.23 亿	36.2 %	11.4%	14.3 %
欧洲	47	5.4M	5.4 亿	49.8 %	34.2 %	36.2 %
拉丁美洲和加勒比地区	42	38 万	36M	3.5 %	3.1%	5.0 %
北美	4	83 万	80M	7.7 %	48.3 %	42.8 %
高收入国家	81	5.8M	5.98 亿	54.0 %	89.1 %	87.5 %
中等收入国家	108	4.9M	499M	45.0 %	10.5 %	12.0 %
低收入国家	28	10万	9.4M	0.9 %	0.4 %	0.5 %

表 1：地理和收入代表性的比较。 SA- 1B 在欧洲、亚洲和大洋洲以及中等收入国家具有较高的代表性。来自非洲、拉丁美洲和加勒比地区以及低收入国家的图像在所有数据集中都不足。

地理和收入代表性。我们推断国家/ 地区图像是使用标准方法拍摄的（参见§C）。在图 **7** 中，我们可视化了每个国家的图像

SA- 1B（左）和图像最多的 50 个国家/ 地区（右）的计数。我们注意到，排名前三的国家是

来自世界不同地区。接下来，在表 1 中，我们比较了 SA- 1B、COCO [66] 和 Open Images [60] 的地理和收入表示。 SA- 1B 在欧洲、亚洲和大洋洲以及中等收入国家的图像比例要高得多。所有数据集都不足以代表非洲和低收入国家。我们注意到，在 SA- 1B 中，包括非洲在内的所有地区都已

至少 2800 万个掩模，比任何先前数据集的掩模总数多 10 倍。最后，我们观察到每个图像（未显示）的平均掩模数量相当多

	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

Table 2: SAM's performance segmenting people across perceived gender presentation, age group, and skin tone. 95% confidence intervals are shown. Within each grouping, all confidence intervals overlap except older *vs*. middle.

Fairness in segmenting people. We investigate potential fairness concerns across perceived gender presentation, perceived age group, and perceived skin tone by measuring the performance discrepancy of SAM between groups. We use the More Inclusive Annotations for People (MIAP) [87] dataset for gender presentation and age and a proprietary dataset for skin tone (see §C). Our evaluation uses simulated interactive segmentation with random sampling of 1 and 3 points (see §D). Table 2 (top left) shows results for perceived gender presentation. We note that females have been shown to be underrepresented in detection and segmentation datasets [115], but observe that SAM performs similarly across groups. We repeat the analysis for perceived age in Table 2 (bottom left), noting that those who are perceived to be younger and older have been shown to be underrepresented in large-scale datasets [110]. SAM performs best on those who are perceived older (although the confidence interval is large). Finally, we repeat the analysis for perceived skin tone in Table 2 (right), noting that those with lighter apparent skin tones have been shown to be overrepresented and those with darker skin tones underrepresented in large-scale datasets [110]. As MIAP does not contain perceived skin tone annotations, we use a proprietary dataset that contains annotations for the perceived Fitzpatrick skin type [36], which ranges from 1 (lightest skin tone) to 6 (darkest skin tone). While the means vary somewhat, we do not find a significant difference across groups. We believe our findings stem from the nature of the task, and acknowledge biases may arise when SAM is used as a component in larger systems. Finally, in §C we extend the analysis to segmenting clothing where we find an indication of bias across perceived gender presentation.

7. Zero-Shot Transfer Experiments

In this section, we present *zero-shot transfer* experiments with SAM, the Segment Anything Model. We consider five tasks, four of which differ significantly from the promptable segmentation task used to train SAM. These experiments evaluate SAM on datasets and tasks that were not seen during training (our usage of "zero-shot transfer" follows its usage in CLIP [82]). The datasets may include novel image distributions, such as underwater or ego-centric images (*e.g*. Fig. 8) that, to our knowledge, do not appear in SA-1B.

Our experiments begin by testing the core goal of promptable segmentation: producing a valid mask from any prompt. We emphasize the challenging scenario of a *single* foreground point prompt, since it is more likely to be ambiguous than other more specific prompts. Next, we present a sequence of experiments that traverse low, mid, and high-level image understanding and roughly parallel the historical development of the field. Specifically, we prompt SAM to (1) perform edge detection, (2) segment everything, *i.e*. object proposal generation, (3) segment detected objects, *i.e*. instance segmentation, and (4), as a proof-of-concept, to segment objects from free-form text. These four tasks differ significantly from the promptable segmentation task that SAM was trained on and are implemented via prompt engineering. Our experiments conclude with an ablation study.

Implementation. Unless otherwise specified: (1) SAM uses an MAE [47] pre-trained ViT-H [33] image encoder and (2) SAM was trained on SA-1B, noting that this dataset includes only automatically generated masks from the final stage of our data engine. For all other model and training details, such as hyperparameters, refer to §A.

7.1. Zero-Shot Single Point Valid Mask Evaluation

Task. We evaluate segmenting an object from a *single* foreground point. This task is ill-posed as one point can refer to multiple objects. Ground truth masks in most datasets do not enumerate *all* possible masks, which can make automatic metrics unreliable. Therefore, we supplement the standard mIoU metric (*i.e*., the mean of all IoUs between predicted and ground truth masks) with a human study in which annotators rate mask quality from 1 (nonsense) to 10 (pixel-perfect). See §D.1, §E, and §G for additional details.

By default, we sample points from the "center" of ground truth masks (at a maximal value of the mask's interior distance transform), following the standard evaluation protocol in interactive segmentation [92]. Since SAM is capable of predicting multiple masks, we evaluate only the model's most confident mask by default. The baselines are all single-mask methods. We compare mainly to RITM [92], a strong interactive segmenter that performs best on our benchmark compared to other strong baselines [67, 18].

Datasets. We use a newly compiled suite of 23 datasets with diverse image distributions. Fig. 8 lists the datasets and shows a sample from each one (see appendix Table 7 for more details). We use all 23 datasets for mIoU evaluation. For the human study, we use the subset listed in Fig. 9b (due to the resource requirements of such studies). This subset includes both datasets for which SAM outperforms and underperforms RITM according to automatic metrics.

	米卢在 1 分	3分		米卢在 1 分	3分
感知的性别呈现女性化	54.4±1.7	90.4±0.6	感知肤色		
男性	55.7 ±1.7	90.1 ±0.6	1	52.9 ±2.2	91.0 ±0.9
感知年龄组较大	62.9±6.7	92.6±1.3	2	51.5 ±1.4	91.1 ±0.5
中间	54.5±1.3	90.2±0.5			
年轻的	54.2±2.2	91.2±0.7	3	52.2 ±1.9	91.4

表 2：SAM 根据感知的性别表现、年龄组和肤色对人群进行细分的表现。显示 95% 置信区间。在每个分组中，除了老年人和中年人之外，所有置信区间都重叠。

人员划分的公平性。我们通过测量各组之间 SAM 的表现差异来调查感知性别表现、感知年龄组和感知肤色的潜在公平问题。我们使用 More Inclusive Annotations for People (MIAP) [87] 数据集来表示性别和年龄，并使用专有数据集来表示肤色（参见§C）。我们的评估使用模拟交互式分割和 1 点和 3 点随机采样（参见 §D）。表 2（左上）显示了感知性别表现的结果。我们注意到，女性在检测和分割数据集中的代表性不足 [115]，但观察到 SAM 在各组中的表现相似。我们重复分析每个

表 2（左下）中的接收年龄，指出那些被认为更年轻和更年长的人已被证明在大规模数据集中代表性不足 [110]。SAM 对于那些被认为年龄较大的人表现最好（尽管置信区间很大）。最后，我们重复一下肛门——

对表 2（右）中感知肤色的分析，注意到在大规模数据集中，肤色较浅的人所占比例过高，而肤色较深的人所占比例不足 [110]。由于 MIAP 不包含感知肤色注释，因此我们使用专有数据集，其中包含感知肤色注释

Fitzpatrick 皮肤类型 [36]，范围从 1（最浅

肤色）到 6（最深肤色）。虽然平均值有所不同，但我们没有发现各组之间存在显着差异。我们相信我们的发现源于任务的性质，并承认当 SAM 用作大型系统的组件时可能会出现偏差。最后，在§C中，我们将分析扩展到服装细分，我们发现感知性别表现存在偏见。

7. 零射击转移实验

在本节中，我们将介绍 SAM（分段任意模型）的零样本传输实验。我们考虑了五个任务，其中四个与用于训练 SAM 的提示分割任务显着不同。这些实验在数据集和任务上评估了 SAM，而这些数据集和任务是在

训练（我们对“零样本迁移”的使用遵循其在 CLIP [82] 中的用法）。数据集可能包括新颖的图像分布，例如水下或以自我为中心的图像（例如图 8），据我们所知，这些图像未出现在 SA–1B 中。

我们的实验首先测试可提示分割的核心目标：根据任何提示生成有效的掩码。我们强调单个前景点提示的挑战性场景，因为它比其他更具体的提示更有可能不明确。接下来，我们提出了一系列横跨低、中、高级图像理解的实验，并且与该领域的历史发展大致平行。具体来说，我们提示 SAM (1) 执行边缘检测，(2) 分割所有内容，即对象提议生成，(3) 分割检测到的对象，即实例分割，以及 (4) 作为概念验证，分割来自自由格式文本的对象。这四个任务与 SAM 接受训练并通过即时工程实现的即时分割任务显着不同。我们的实验以消融研究结束。

执行。除非另有说明：(1) SAM 使用 MAE [47] 预训练的 ViT–H [33] 图像编码器，(2) SAM 在 SA–1B 上进行训练，注意该数据集仅包含最后阶段自动生成的掩模我们的数据引擎。有关所有其他模型和训练详细信息（例如超参数），请参阅§A。

7.1.零样本单点有效掩模评估任务。我们评估从单个前景点分割对象。这项任务是不适定的，因为一个点可以引用多个对象。大多数数据集中的地面实况掩码不会枚举所有可能的掩码，这可能会使自动指标不可靠。因此，我们通过人类研究补充了标准 mIoU 指标（即预测掩模和真实掩模之间所有 IoU 的平均值），其中注释者对掩模质量进行评分，从 1（无意义）到 10（像素完美）。有关更多详细信息，请参阅§D.1、§E 和§G。默认情况下，我们遵循交互式分割中的标准评估协议，从地面真值掩模的“中心”（在掩模内部距离变换的最大值）采样点。由于 SAM 能够预测多个掩模，因此默认情况下我们仅评估模型最置信度的掩模。基线都是单掩模方法。我们主要与 RITM [92] 进行比较，RITM 是一种强大的交互式分段器，与其他强大的基线相比，它在我们的基准上表现最好 [67, 18]。

数据集。我们使用新编译的 23 个数据集套件

具有不同的图像分布。图8列出了数据集

并显示了每一项的样本（参见附录表 7

更多细节）。我们使用所有 23 个数据集进行 mIoU 评估。对于人类研究，我们使用图 9b 中列出的子集（由于此类研究的资源要求）。该子集包括根据自动指标，SAM 优于和低于 RITM 的数据集。

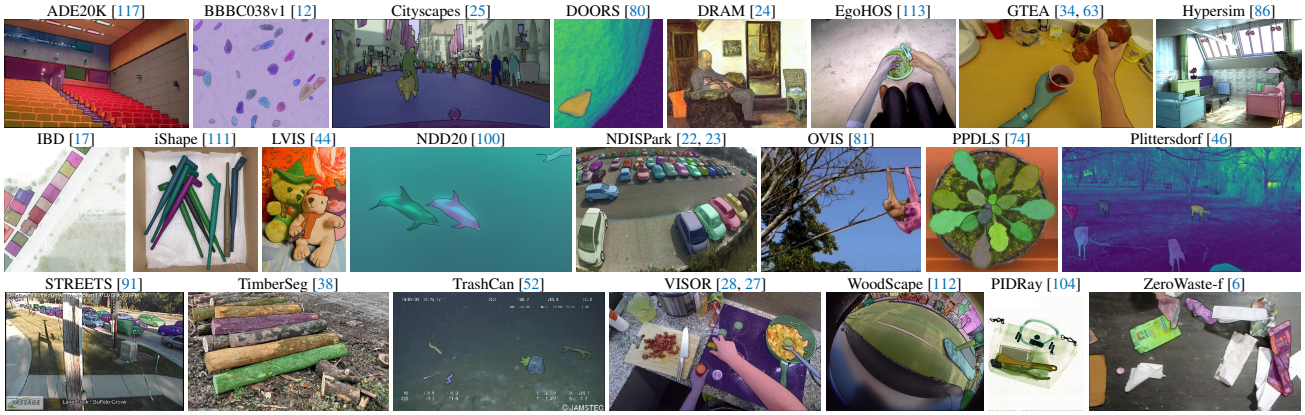
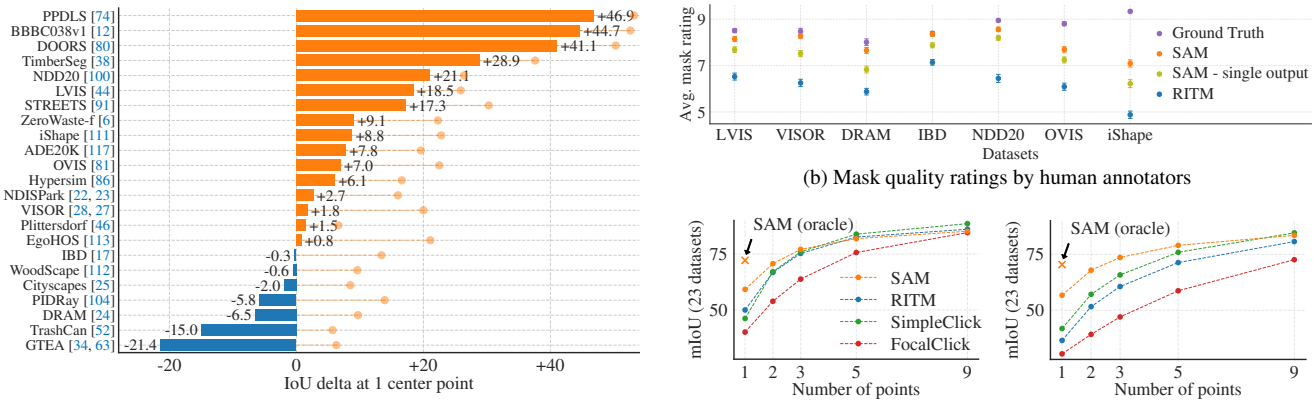


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.



(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.

Results. First, we look at automatic evaluation on the full suite of 23 datasets using mIoU. We compare per-dataset results in Fig. 9a against RITM. SAM yields higher results on 16 of the 23 datasets, by as much as ~47 IoU. We also present an "oracle" result, in which the most relevant of SAM's 3 masks is selected by comparing them to the ground truth, rather than selecting the most confident mask. This reveals the impact of ambiguity on automatic evaluation. In particular, with the oracle to perform ambiguity resolution, SAM outperforms RITM on *all* datasets.

Results of the human study are presented in Fig. 9b. Error bars are 95% confidence intervals for mean mask ratings (all differences are significant; see §E for details). We observe that the annotators consistently rate the quality of SAM's masks substantially higher than the strongest baseline, RITM. An ablated, "ambiguity-unaware" version of SAM with a single output mask has consistently lower ratings, though still higher than RITM. SAM's mean ratings fall between 7 and 9, which corresponds to the qualitative rating guideline: "*A high score (7-9): The object is identifiable and errors are small and rare* (e.g., *missing a small, heavily obscured disconnected component, ...).*" These results indicate that SAM has learned to segment valid masks from a single point. Note that for datasets like DRAM and IBD, where SAM is worse on automatic metrics, *it receives consistently higher ratings in the human study*.

Fig. 9c shows additional baselines, SimpleClick [67] and FocalClick [18], which obtain lower single point performance than RITM and SAM. As the number of points increases from 1 to 9, we observe that the gap between methods decreases. This is expected as the task becomes easier; also, SAM is not optimized for the very high IoU regime. Finally, in Fig. 9d we replace the default center point sampling with random point sampling. We observe that the gap between SAM and the baselines grows and SAM is able to achieve comparable results under either sampling method.

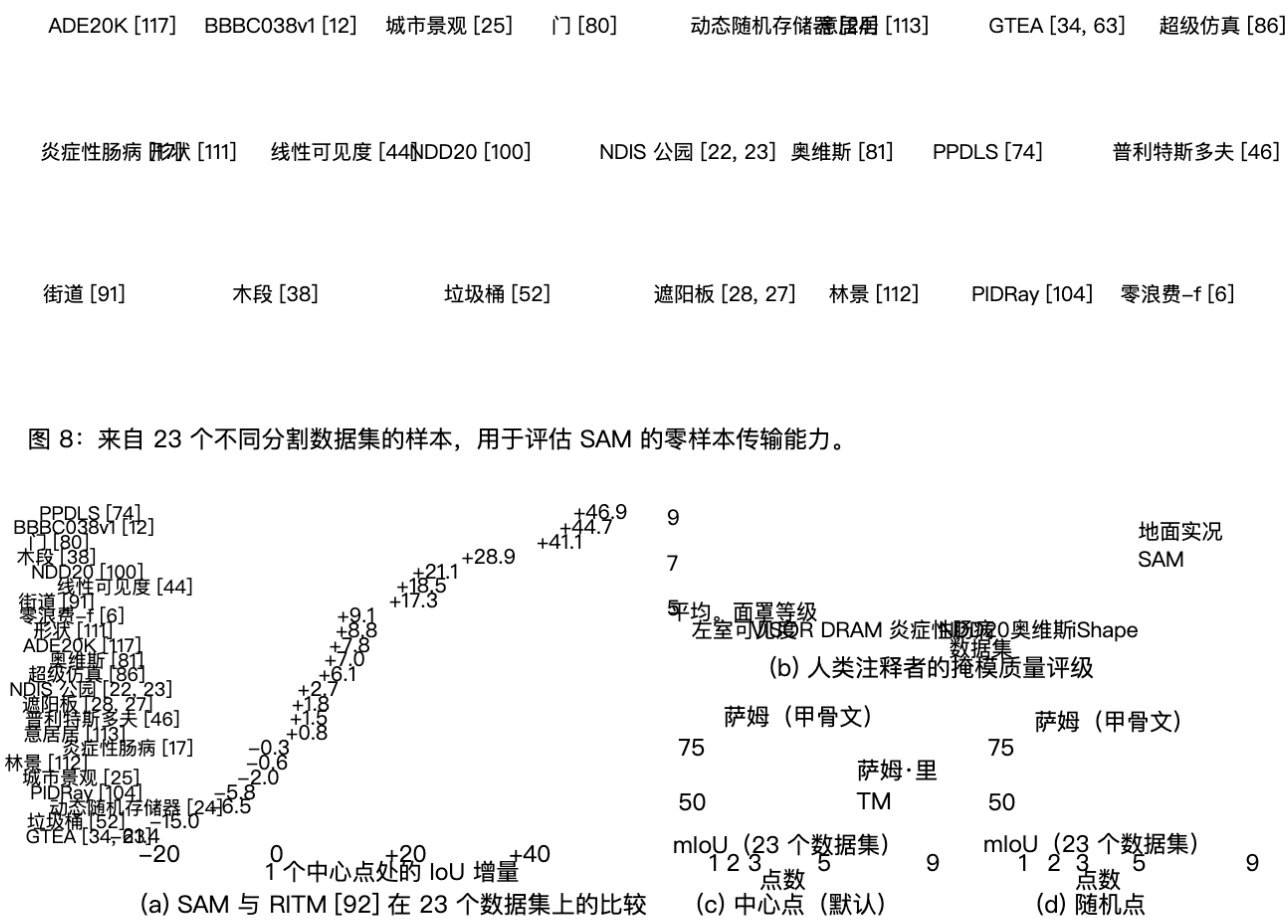


图 9：对 23 个数据集进行指向掩模评估。(a) SAM 和最强单点分段器 RITM 的平均 IoU [92]。由于模糊性，单个掩码可能与真实情况不匹配；圆圈显示 SAM 3 中最相关的“oracle”结果

预测。(b) 注释者对每个数据集的掩模质量评级进行比较，从 1（最差）到 10（最好）。所有方法都使用地面实况掩模中心作为提示。(c, d) 具有不同点数的 MIoU。 SAM 显着优于之前的

结果。首先，我们使用 mIoU 对全套 23 个数据集进行自动评估。我们将图 9a 中每个数据集的结果与 RITM 进行比较。 SAM 产生更高的重新

在 23 个数据集中的 16 个上得到的结果高达 ~47 IoU。我们还提出了一个“预言”结果，其中最相关的

SAM 的 3 个掩模是通过将它们与真实情况进行比较来选择的，而不是选择最有信心的掩模。这揭示了模糊性对自动评估的影响。特别是，通过预言机执行歧义解析，SAM 在所有数据集上都优于 RITM。

人体研究的结果如图 9b 所示。误差线是平均掩模评级的 95% 置信区间（所有差异都很显着；有关详细信息，请参阅§E）。我们观察到，注释者对 SAM 掩模质量的评价始终高于最强基线 RITM。带有单个输出掩码的 SAM 的“模糊性无意识”版本的评级始终较低，但仍高于 RITM。 SAM 的平均评分

落在 7 到 9 之间，对应于定性评级指南：“高分 (7–9)：该对象是可识别的，错误很小且很少见（例如，遗漏了一个小的、严重模糊的断开组件，...）”。这些结果表明 SAM 已经学会从单个点分割有效掩码。请注意，对于 DRAM 和 IBD 等数据集，SAM 在自动指标方面表现较差，但它在人类研究中始终获得较高的评级。

图 9c 显示了额外的基线 SimpleClick [67] 和 FocalClick [18]，它们的单点性能低于 RITM 和 SAM。随着点数从 1 增加到 9，我们观察到方法之间的差距缩小。随着任务变得更容易，这是预料之中的；此外，SAM 并未针对非常高的 IoU 状态进行优化。最后，在图 9d 中，我们用随机点采样替换默认的中心点采样。我们观察到 SAM 与基线之间的差距越来越大，并且 SAM 在任一采样方法下都能够获得可比较的结果。

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.

7.2. Zero-Shot Edge Detection

Approach. We evaluate SAM on the classic low-level task of edge detection using BSDS500 [72, 3]. We use a simplified version of our automatic mask generation pipeline. Specifically, we prompt SAM with a 16×16 regular grid of foreground points resulting in 768 predicted masks (3 per point). Redundant masks are removed by NMS. Then, edge maps are computed using Sobel filtering of unthresholded mask probability maps and standard lightweight postprocessing, including edge NMS (see §D.2 for details).

Results. We visualize representative edge maps in Fig. 10 (see Fig. 15 for more). Qualitatively, we observe that even though SAM was not trained for edge detection, it produces reasonable edge maps. Compared to the ground truth, SAM predicts more edges, including sensible ones that are not annotated in BSDS500. This bias is reflected quantitatively in Table 3: recall at 50% precision (R50) is high, at the cost of precision. SAM naturally lags behind state-of-the-art methods that learn the biases of BSDS500, *i.e.*, which edges to suppress. Nevertheless, SAM performs well compared to pioneering deep learning methods such as HED [108] (also trained on BSDS500) and significantly better than prior, though admittedly outdated, zero-shot transfer methods.

7.3. Zero-Shot Object Proposals

Approach. Next, we evaluate SAM on the mid-level task of object proposal generation [2, 102]. This task has played an important role in object detection research, serving as an intermediate step in pioneering systems (*e.g.*, [102, 41, 84]). To generate object proposals, we run a slightly modified version of our automatic mask generation pipeline and output the masks as proposals (see §D.3 for details).

We compute the standard average recall (AR) metric on LVIS v1 [44]. We focus on LVIS because its large number of categories presents a challenging test. We compare to a *strong* baseline implemented as a ViTDet [62] detector (with cascade Mask R-CNN [48, 11] ViT-H). We note that this "baseline" corresponds to the "Detector Masquerading as Proposal generator" (DMP) method [16] that was shown to game AR, making it a truly demanding comparison.

	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.

Results. In Table 4 we see unsurprisingly that using the detections from ViTDet-H as object proposals (*i.e.*, the DMP method [16] that games AR) performs the best overall. However, SAM does remarkably well on several metrics. Notably, it outperforms ViTDet-H on medium and large objects, as well as rare and common objects. In fact, SAM only underperforms ViTDet-H on small objects and frequent objects, where ViTDet-H can easily learn LVIS-specific annotation biases since it was trained on LVIS, unlike SAM. We also compare against an ablated ambiguity-unaware version of SAM ("single out."), which performs significantly worse than SAM on all AR metrics.

7.4. Zero-Shot Instance Segmentation

Approach. Moving to higher-level vision, we use SAM as the segmentation module of an instance segmenter. The implementation is simple: we run a object detector (the ViTDet used before) and prompt SAM with its output boxes. This illustrates *composing* SAM in a larger system.

Results. We compare the masks predicted by SAM and ViTDet on COCO and LVIS in Table 5. Looking at the mask AP metric we observe gaps on both datasets, where SAM is reasonably close, though certainly behind ViTDet. By visualizing outputs, we observed that SAM masks are often qualitatively better than those of ViTDet, with crisper boundaries (see §D.4 and Fig. 16). To investigate this observation, we conducted an additional human study asking annotators to rate the ViTDet masks and SAM masks on the 1 to 10 quality scale used before. In Fig. 11 we observe that SAM consistently outperforms ViTDet in the human study.

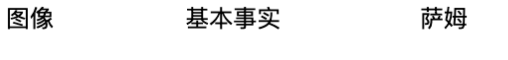


图 10：BSDS500 上的零样本边缘预测。 SAM 没有接受过预测边缘图的训练，也无法在训练期间访问 BSDS 图像或注释。

方法年份	ODS	OIS	AP	R50
海电 [108] 2015	.788	.808	.840	.923
零样本传递方法：Sobel 滤波器 1968				
539	–	–	–	
精明 [13] 1986	600	640	580	–
费尔兹赫特 [35] 2004	610	640	560	–
萨姆 2023	768	786	794	928

表 3：BSDS500 上的零样本传输到边缘检测。

7.2.零射击边缘检测

方法。我们使用 BSDS500 [72, 3] 在边缘检测的经典低级任务上评估 SAM。我们使用自动掩码生成管道的简化版本。

具体来说，我们用 16×16 的规则网格提示 SAM

前景点产生 768 个预测掩模（每点 3 个）。冗余掩码由 NMS 删除。然后，使用未阈值掩模概率图的 Sobel 滤波和标准轻量级后处理来计算边缘图。

处理，包括边缘 NMS（详细信息请参阅§D.2）。

结果。我们在图 10 中可视化了代表性边缘图（更多信息请参见图 15）。定性地，我们观察到即使 SAM 没有经过边缘检测训练，它也会产生合理的边缘图。与真实值相比，SAM 预测了更多边缘，包括 BSDS500 中未注释的合理边缘。这种偏差在表 3 中得到了定量反映：50% 精确度 (R50) 下的召回率很高，但代价是精确度。 SAM 自然落后于学习 BSDS500 偏差（即要抑制哪些边缘）的最先进方法。尽管如此，与 HED [108]（也在 BSDS500 上训练）等开创性的深度学习方法相比，SAM 表现良好，并且明显优于之前的零样本迁移方法（尽管不可否认这是过时的）。

7.3.零样本对象提案

方法。接下来，我们在对象提议生成的中级任务上评估 SAM [2, 102]。该任务在目标检测研究中发挥了重要作用，作为

	掩模AR@1000						
方法	全部	小的	医学。	大的	频率。	com.	稀有的
ViTDet–H [62]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
零次传输方法：SAM							
——单出。	54.9	42.8	76.7	74.4	54.7	59.8	62.0
萨姆	59.3	45.5	81.6	86.9	59.1	63.9	65.8

表 4：LVIS v1 上的对象提案生成。 SAM 是零样本应用，即它没有经过目标提议生成的训练，也没有访问 LVIS 图像或注释。

开创性系统的中间步骤（例如，[102,41,84]）。为了生成对象建议，我们运行自动掩码生成管道的稍微修改版本，并将掩码作为建议输出（有关详细信息，请参阅§D.3）。

我们计算 LVIS v1 上的标准平均召回率 (AR) 指标 [44]。我们专注于 LVIS，因为它的大量类别带来了挑战性的测试。我们与作为 ViTDet [62] 检测器实现的强基线（带有级联 Mask R–CNN [48, 11] ViT–H）进行比较。我们注意到，这个“基线”对应于游戏 AR 中展示的“检测器伪装成提案生成器”(DMP) 方法 [16]，这使其成为真正要求严格的比较。

结果。在表 4 中，我们毫不奇怪地看到，使用 ViTDet–H 的检测作为对象建议（即游戏 AR 的 DMP 方法 [16]）总体表现最佳。然而，SAM 在几个指标上都表现得非常好。值得注意的是，它在中型和大型物体以及稀有和常见物体上的性能优于 ViTDet–H。事实上，SAM 仅在小对象和频繁对象上表现不佳 ViTDet–H，其中 ViTDet–H 可以轻松学习 LVIS 特定注释偏差，因为它是在 LVIS 上训练的，与 SAM 不同。我们还与 SAM 的消除歧义无意识版本（“单出”）进行比较，该版本在所有 AR 指标上的表现均明显差于 SAM。

7.4.零样本实例分割

方法。转向更高层次的视觉，我们使用 SAM 作为实例分割器的分割模块。实现很简单：我们运行一个对象检测器（之前使用的 ViTDet）并通过其输出框提示 SAM。这说明了如何在更大的系统中组合 SAM。

结果。我们在表 5 中比较了 SAM 和 ViTDet 在 COCO 和 LVIS 上预测的掩模。查看掩模 AP 指标，我们观察到两个数据集上的差距，其中 SAM 相当接近，但肯定落后于 ViTDet。通过可视化输出，我们观察到 SAM 掩模在质量上通常优于 ViTDet，并且边界更清晰（参见§D.4 和图 16）。为了调查这一观察结果，我们进行了一项额外的人类研究，要求注释者对 ViTDet 掩模和 SAM 掩模进行评分

之前使用过的 1 到 10 质量等级。在图 11 中，我们观察到 SAM 在人体研究中始终优于 ViTDet。

	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).

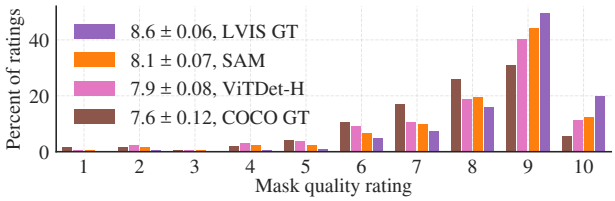


Figure 11: Mask quality rating distribution from our human study for ViTDet and SAM, both applied to LVIS ground truth boxes. We also report LVIS and COCO ground truth quality. The legend shows rating means and 95% confidence intervals. Despite its lower AP (Table 5), SAM has higher ratings than ViTDet,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

We hypothesize that on COCO, where the mask AP gap is larger and the ground truth quality is relatively low (as borne out by the human study), ViTDet learns the specific biases of COCO masks. SAM, being a zero-shot method, is unable to exploit these (generally undesirable) biases. The LVIS dataset has higher quality ground truth, but there are still specific idiosyncrasies (*e.g.*, masks do not contain holes, they are simple polygons by construction) and biases for modal *vs*. amodal masks. Again, SAM is not trained to learn these biases, while ViTDet can exploit them.

7.5. Zero-Shot Text-to-Mask

Approach. Finally, we consider an even higher-level task: segmenting objects from free-form text. This experiment is a proof-of-concept of SAM's ability to process text prompts. While we used the exact same SAM in all prior experiments, for this one SAM's training procedure is modified to make it text-aware, but in a way that does not require new text annotations. Specifically, for each manually collected mask with area larger than 100^2 we extract the CLIP *image* embedding. Then, during training, we prompt SAM with the extracted CLIP image embeddings as its first interaction. The key observation here is that because CLIP's *image* embeddings are trained to align with its *text* embeddings, we can train with image embeddings, but use text embeddings for inference. That is, at inference time we run text through CLIP's text encoder and then give the resulting text embedding as a prompt to SAM (see §D.5 for details).

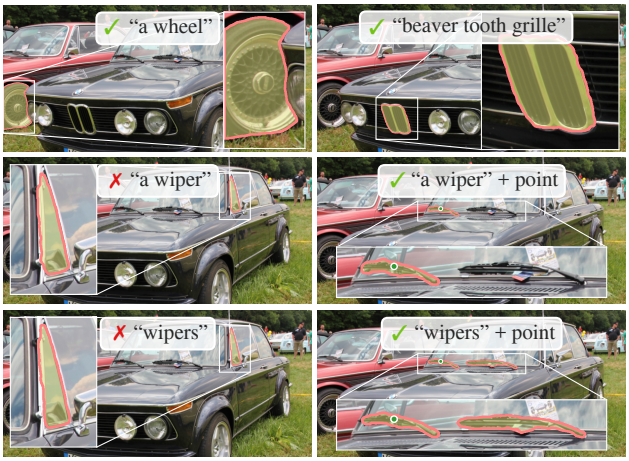


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Results. We show qualitative results in Fig. 12. SAM can segment objects based on simple text prompts like "a wheel" as well as phrases like "beaver tooth grille". When SAM fails to pick the right object from a text prompt only, an additional point often fixes the prediction, similar to [31].

7.6. Ablations

We perform several ablations on our 23 dataset suite with the single center point prompt protocol. Recall that a single point may be ambiguous and that ambiguity may not be represented in the ground truth, which contains only a single mask per point. Since SAM is operating in a zero-shot transfer setting there can be systematic biases between SAM's top-ranked mask *vs*. the masks resulting from data annotation guidelines. We therefore additionally report the best mask with respect to the ground truth ("oracle").

Fig. 13 (left) plots SAM's performance when trained on cumulative data from the data engine stages. We observe that each stage increases mIoU. When training with all three stages, the automatic masks vastly outnumber the manual and semi-automatic masks. To address this, we found that oversampling the manual and semi-automatic masks during training by 10× gave best results. This setup complicates training. We therefore tested a fourth setup that uses only the automatically generated masks. With this data, SAM performs only marginally lower than using all data (~0.5 mIoU). Therefore, by default we use only the automatically generated masks to simplify the training setup.

In Fig. 13 (middle) we look at the impact of data volume. The full SA-1B contains 11M images, which we uniformly subsample to 1M and 0.1M for this ablation. At 0.1M images, we observe a large mIoU decline under all settings. However, with 1M images, about 10% of the full dataset, we observe results comparable to using the full dataset. This data regime, which still includes approximately 100M masks, may be a practical setting for many use cases.

方法	可可 [66]			LVIS v1 [44]		
	美联社	美联社$_S$	美联社$_M$	美联社	美联社$_S$	美联社$_M$
ViTDet-H [62]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
零样本传输方法（仅限分割模块）：SAM								
	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5

表 5：实例分割结果。 ViTDet 框提示 SAM 进行零样本分割。完全监督的 ViTDet 优于 SAM，但在更高质量的 LVIS 掩模上差距缩小了。有趣的是，根据人类评分，SAM 的表现优于 ViTDet（见图 11）。

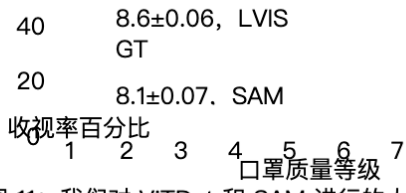


图 11：我们对 ViTDet 和 SAM 进行的人类研究中的掩模质量评级分布，均应用于 LVIS 地面实况盒。我们还报告 LVIS 和 COCO 地面真实质量。图例显示评级平均值和 95% 置信区间。尽管 AP 较低（表 5），但 SAM 的评级高于 ViTDet，这表明 ViTDet 利用了 COCO 和 LVIS 训练数据中的偏差。

我们假设在 COCO 上，掩模 AP 间隙较大且地面实况质量相对较低（正如人类研究所证实的那样），ViTDet 学习 COCO 掩模的特定偏差。 SAM 作为一种零样本方法，无法利用这些（通常是不需要的）偏差。 LVIS 数据集具有更高质量的地面实况，但仍然存在特定的特性（例如，掩模不包含孔，它们在结构上是简单的多边形）以及模态与非模态掩模的偏差。同样，SAM 没有接受过学习这些偏差的训练，而 ViTDet 可以利用它们。

7.5。零样本文本到掩模

方法。最后，我们考虑一个更高级别的任务：从自由格式文本中分割对象。该实验是 SAM 处理文本提示能力的概念验证。虽然我们在之前的所有实验中都使用了完全相同的 SAM，但在本次实验中，我们修改了 SAM 的训练程序以使其具有文本感知能力，但不需要新的文本注释。具体来说，对于每个手动收集的面积大于 100 的掩模，我们提取 CLIP 图像嵌入。然后，在训练过程中，我们提示 SAM 使用提取的 CLIP 图像嵌入作为其第一次交互。这里的关键观察是，由于 CLIP 的图像嵌入经过训练以与其文本嵌入对齐，因此我们可以使用图像嵌入进行训练，但使用文本嵌入进行推理。也就是说，在推理时，我们通过 CLIP 的文本编码器运行文本，然后给出结果

文本嵌入作为 SAM 的提示（有关详细信息，请参阅 §D.5）。

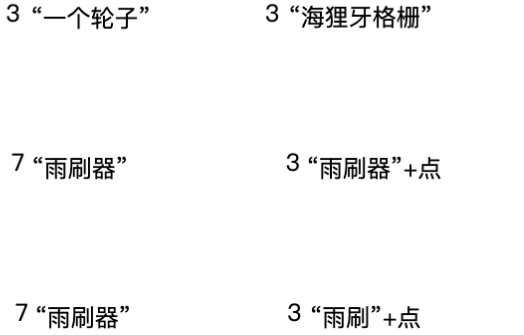


图 12：零样本文本到蒙版。 SAM 可以使用简单而细致的文本提示。当 SAM 无法做出正确预测时，附加点提示可以提供帮助。

结果。我们在图 12 中显示了定性结果。SAM 可以根据简单的文本提示（如“轮子”）以及短语（如“海狸齿格栅”）来分割对象。当 SAM 无法仅从文本提示中选择正确的对象时，附加点通常会修复预测，类似于 [31]。

7.6。消融

我们使用单中心点提示协议对 23 个数据集套件执行多次消融。回想一下，单个点可能是不明确的，并且这种模糊性可能无法在基本事实中表示，每个点仅包含一个掩码。由于 SAM 在零样本传输设置中运行，因此 SAM 的顶级掩码与数据注释指南产生的掩码之间可能存在系统偏差。因此，我们还报告了关于真实情况（“oracle”）的最佳掩码。

图 13（左）绘制了 SAM 在使用数据引擎阶段的累积数据进行训练时的性能。我们观察到每个阶段都会增加 mIoU。在所有三个阶段的训练中，自动面罩的数量远远超过手动和半自动面罩。为了解决这个问题，我们发现在训练期间对手动和半自动掩模进行 10 倍的过采样可以得到最佳结果。这种设置使训练变得复杂。因此，我们测试了仅使用自动生成的掩码的第四种设置。使用此数据，SAM 的性能仅略低于使用所有数据的性能 (~0.5 mIoU)。因此，默认情况下我们仅使用自动生成的掩模来简化训练设置。

在图 13（中）中，我们研究了数据量的影响。完整的 SA–1B 包含 11M 图像，我们将其统一子采样为 1M 和 0.1M 以进行此消融。在 0.1M 图像中，我们观察到所有设置下的 mIoU 都有较大下降。然而，对于 100 万张图像（约占完整数据集的 10%），我们观察到的结果与使用完整数据集相当。该数据机制仍包含大约 100M 个掩码，可能是许多用例的实用设置。

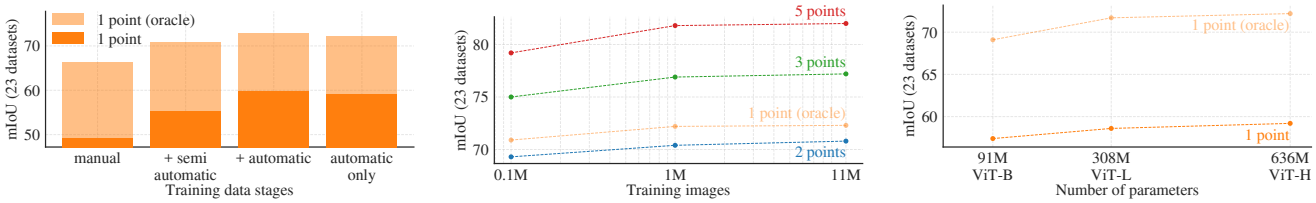


Figure 13: Ablation studies of our data engine stages, image encoder scaling, and training data scaling. (Left) Each data engine stage leads to improvements on our 23 dataset suite, and training with only the automatic data (our default) yields similar results to using data from all three stages. (Middle)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

Finally, Fig. 13 (right) shows results with ViT-B, ViT-L, and ViT-H image encoders. ViT-H improves substantially over ViT-B, but has only marginal gains over ViT-L. Further image encoder scaling does not appear fruitful at this time.

8. Discussion

Foundation models. Pre-trained models have been adapted to downstream tasks since the early days of machine learning [99]. This paradigm has become increasingly important in recent years with a growing emphasis on scale, and such models have recently been (re-)branded as "foundation models": *i.e.* models that are "trained on broad data at scale and are adaptable to a wide range of downstream tasks" [8]. Our work correlates well with this definition, though we note that a foundation model for image segmentation is an inherently limited scope, since it represents an important, yet fractional, subset of computer vision. We also contrast one aspect of our approach with [8], which emphasizes the role of *self-supervised* learning in foundation models. While our model is initialized with a self-supervised technique (MAE [47]), the vast majority of its capabilities come from large-scale *supervised* training. In cases where data engines can scale available annotations, like ours, supervised training provides an effective solution.

Compositionality. Pre-trained models can power new capabilities even beyond ones imagined at the moment of training. One prominent example is how CLIP [82] is used as a *component* in larger systems, such as DALL·E [83]. Our goal is to make this kind of composition straightforward with SAM. We aim to achieve this by requiring SAM to predict a valid mask for a wide range of segmentation prompts. The effect is to create a reliable interface between SAM and other components. For example, MCC [106] can easily use SAM to segment an object of interest and achieve strong generalization to unseen objects for 3D reconstruction from a single RGB-D image. In another example, SAM can be prompted with gaze points detected by a wearable device, enabling new applications. Thanks to SAM's ability to generalize to new domains like ego-centric images, such systems work without need for additional training.

Limitations. While SAM performs well in general, it is not perfect. It can miss fine structures, hallucinates small disconnected components at times, and does not produce boundaries as crisply as more computationally intensive methods that "zoom-in", *e.g.* [18]. In general, we expect dedicated interactive segmentation methods to outperform SAM when many points are provided, *e.g.* [67]. Unlike these methods, SAM is designed for generality and breadth of use rather than high IoU interactive segmentation. Moreover, SAM can process prompts in real-time, but nevertheless SAM's overall performance is not real-time when using a heavy image encoder. Our foray into the text-to-mask task is exploratory and not entirely robust, although we believe it can be improved with more effort. While SAM can perform many tasks, it is unclear how to design simple prompts that implement semantic and panoptic segmentation. Finally, there are domain-specific tools, such as [7], that we expect to outperform SAM in their respective domains.

Conclusion. The Segment Anything project is an attempt to lift image segmentation into the era of foundation models. Our principal contributions are a new task (promptable segmentation), model (SAM), and dataset (SA-1B) that make this leap possible. Whether SAM achieves the status of a foundation model remains to be seen by how it is used in the community, but regardless we expect the perspective of this work, the release of over 1B masks, and our promptable segmentation model will help pave the path ahead.

Acknowledgments. We would like to thank Aaron Adcock and Jitendra Malik for helpful discussion. We thank Vaibhav Aggarwal and Yanghao Li for help with scaling the model. We thank Cheng-Yang Fu, Jiabo Hu, and Robert Kuo for help with data annotation platform. We thank Allen Goodman and Bram Wasti for help in optimizing web-version of our model. Finally, we thank Morteza Behrooz, Ashley Gabriel, Ahuva Goldstand, Sumanth Gurram, Somya Jain, Devansh Kukreja, Joshua Lane, Lilian Luong, Mallika Malhotra, William Ngan, Omkar Parkhi, Nikhil Raina, Dirk Rowe, Neil Sejoor, Vanessa Stark, Bala Varadarajan, and Zachary Winstrom for their help in making the demo, dataset viewer, and other assets and tooling.

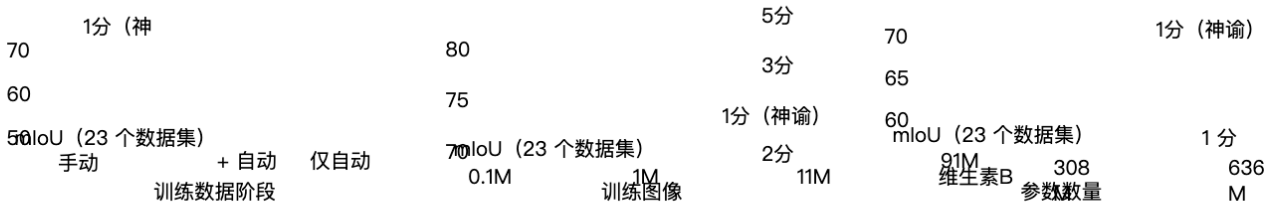


图 13：我们的数据引擎阶段、图像编码器缩放和训练数据缩放的消融研究。 （左）每个数据引擎阶段都会改进我们的 23 个数据集套件，并且仅使用自动数据（我们的默认值）进行训练会产生与使用所有三个阶段的数据相似的结果。 （中）用约 10% 的 SA–1B 训练的 SAM 与完整的 SA–1B 相当。我们默认使用所有 11M 图像进行训练，但使用 1M 图像是一个合理的实际设置。 （右）缩放 SAM 的图像编码器显示出有意义但饱和的增益。然而，在某些设置中，较小的图像编码器可能是首选。

最后，图 13（右）显示了 ViT–B、ViT–L 和 ViT–H 图像编码器的结果。 ViT–H 比 ViT–B 有了显着改善，但仅比 ViT–L 有边际收益。目前，进一步的图像编码器缩放似乎没有成果。

八、讨论

基础模型。自机器学习早期以来，预训练模型就已适应下游任务[99]。近年来，随着对规模的日益重视，这种范式变得越来越重要，并且此类模型最近被（重新）称为“基础模型”：即“在广泛的数据上进行大规模训练并适应广泛的模型”。一系列下游任务”[8]。我们的工作与这个定义很好地相关，尽管我们注意到图像分割的基础模型本质上是有限的范围，因为它代表了计算机视觉的一个重要但又不完整的子集。我们还将我们的方法的一个方面与[8]进行了对比，后者强调了自我监督学习在基础模型中的作用。虽然我们的模型是使用自监督技术（MAE [47]）初始化的，但其绝大多数功能来自大规模监督训练。在数据引擎可以扩展可用注释的情况下（例如我们的），监督训练提供了有效的解决方案。

组合性。预先训练的模型可以提供新的功能，甚至超出训练时的想象。一个突出的例子是如何将 CLIP [82] 用作较大系统（例如 DALL·E [83]）中的组件。我们的目标是通过 SAM 使这种组合变得简单。我们的目标是通过要求 SAM 预测各种分割提示的有效掩码来实现这一目标。其效果是在 SAM 和其他组件之间创建可靠的接口。例如，MCC [106] 可以轻松地使用 SAM 来分割感兴趣的对象，并实现对看不见的对象的强泛化，以便从单个 RGB–D 图像进行 3D 重建。在另一个例子中，SAM 可以通过可穿戴设备检测到的注视点进行提示，从而启用新的应用程序。由于 SAM 能够推广到以自我为中心的图像等新领域，因此此类系统无需额外培训即可运行。

局限性。虽然 SAM 总体表现良好，但它并不完美。它可能会错过精细结构，有时会产生小的不连贯组件的幻觉，并且不会像“放大”的计算密集型方法那样清晰地产生边界，例如[18]。一般来说，当提供许多点时，我们期望专用的交互式分割方法能够优于 SAM，例如[67]。与这些方法不同，SAM 的设计目的是为了通用性和广泛使用，而不是高 IoU 交互式分割。此外，SAM 可以实时处理提示，但是当使用重型图像编码器时，SAM 的整体性能不是实时的。我们对文本到掩码任务的尝试是探索性的，并不完全稳健，尽管我们相信它可以通过更多的努力来改进。虽然 SAM 可以执行许多任务，但尚不清楚如何设计实现语义和全景分割的简单提示。最后，还有一些特定领域的工具，例如 [7]，我们希望它们在各自的领域中优于 SAM。

结论。 Segment Anything 项目试图将图像分割提升到基础模型时代。我们的主要贡献是使这一飞跃成为可能的新任务（即时分割）、模型 (SAM) 和数据集 (SA–1B)。 SAM 是否达到基础模型的地位还有待观察它在社区中的使用情况，但无论我们期望这项工作的角度如何，超过 1B 个掩模的发布以及我们及时的分割模型将有助于铺平前进的道路。

致谢。我们要感谢 Aaron Adcock 和 Jitendra Malik 的有益讨论。我们感谢 Vaibhav Aggarwal 和 Yanghao Li 在扩展模型方面提供的帮助。我们感谢 Cheng–Yang Fu、Jiabo Hu 和 Robert Kuo 在数据注释平台方面提供的帮助。我们感谢 Allen Goodman 和 Bram Wasti 在优化我们模型的网络版本方面提供的帮助。最后，我们感谢 Morteza Behrooz、Ashley Gabriel、Ahuva Goldstand、Sumanth Gurram、Somya Jain、Devansh Kukreja、Joshua Lane、Lilian Luong、Mallika Malhotra、William Ngan、Omkar Parkhi、Nikhil Raina、Dirk Rowe、Neil Sejoor、Vanessa Stark、Bala Varadarajan 和 Zachary Winstrom 在制作演示、数据集查看器以及其他资产和工具方面提供的帮助。

References

[1] Edward H Adelson. On seeing stuff: the perception of materials by humans and machines. *Human vision and electronic imaging VI*, 2001. 5

[2] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? *CVPR*, 2010. 4, 10

[3] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2010. 4, 10, 21, 28

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv:1607.06450*, 2016. 16

[5] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. *arXiv:2106.08254*, 2021. 17

[6] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes. *CVPR*, 2022. 9, 20

[7] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N. Straehle, Bernhard X. Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, Kemal Eren, Jaime I. Cervantes, Buote Xu, Fynn Beuttenmueller, Adrian Wolny, Chong Zhang, Ullrich Koethe, Fred A. Hamprecht, and Anna Kreshuk. ilastik: interactive machine learning for (bio)image analysis. *Nature Methods*, 2019. 12

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021. 1, 12

[9] Gustav Bredell, Christine Tanner, and Ender Konukoglu. Iterative interaction training for segmentation editing networks. *MICCAI*, 2018. 17

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NeurIPS*, 2020. 1, 4

[11] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. *CVPR*, 2018. 10

[12] Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature Methods*, 2019. 9, 19, 20

[13] John Canny. A computational approach to edge detection. *TPAMI*, 1986. 10, 21

[14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with Transformers. *ECCV*, 2020. 5, 16, 17

[15] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. *ECCV*, 2008. 5, 17

[16] Neelima Chavali, Harsh Agrawal, Aroma Mahendru, and Dhruv Batra. Object-proposal evaluation protocol is' gameable'. *CVPR*, 2016. 10, 21

[17] Jiazhou Chen, Yanghui Xu, Shufang Lu, Ronghua Liang, and Liangliang Nan. 3D instance segmentation of MVS buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 9, 19, 20, 23, 24

[18] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. FocalClick: towards practical interactive image segmentation. *CVPR*, 2022. 8, 9, 12, 19

[19] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *CVPR*, 2022. 4

[20] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 2021. 5, 16, 17

[21] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv:2204.02311*, 2022. 1

[22] Luca Ciampi, Carlos Santiago, Joao Costeira, Claudio Gennaro, and Giuseppe Amato. Domain adaptation for traffic density estimation. *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2021. 9, 20

[23] Luca Ciampi, Carlos Santiago, Joao Costeira, Claudio Gennaro, and Giuseppe Amato. Night and day instance segmented park (NDISPark) dataset: a collection of images taken by day and by night for vehicle detection, segmentation and counting in parking areas. *Zenodo*, 2022. 9, 20

[24] Nadav Cohen, Yael Newman, and Ariel Shamir. Semantic segmentation in art paintings. *Computer Graphics Forum*, 2022. 9, 19, 20, 23, 24

[25] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. *CVPR*, 2016. 9, 19, 20

[26] Bruno da Silva, George Konidaris, and Andrew Barto. Learning parameterized skills. *ICML*, 2012. 4

[27] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *IJCV*, 2022. 9, 20, 23, 24

[28] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations. *NeurIPS*, 2022. 9, 19, 20, 23, 24

[29] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? *CVPR workshops*, 2019. 18

[30] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. CrowdWorkSheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. *ACM Conference on Fairness, Accountability, and Transparency*, 2022. 25

[31] Henghui Ding, Scott Cohen, Brian Price, and Xudong Jiang. PhraseClick: toward achieving flexible interactive segmentation by phrase and click. *ECCV*, 2020. 11

[32] Piotr Dollár and C Lawrence Zitnick. Fast edge detection using structured forests. *TPAMI*, 2014. 21

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 5, 8, 16

[34] Alireza Fathi, Xiaofeng Ren, and James M. Rehg. Learning to recognize objects in egocentric activities. *CVPR*, 2011. 9, 19, 20

[35] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004. 10

[36] Thomas B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of Dermatology*, 1988. 8

[37] Marco Forte, Brian Price, Scott Cohen, Ning Xu, and François Pitié. Getting to 99% accuracy in interactive segmentation. *arXiv:2003.07932*, 2020. 5, 17

[38] Jean-Michel Fortin, Olivier Gamache, Vincent Grondin, François Pomerleau, and Philippe Giguère. Instance segmentation for autonomous log grasping in forestry operations. *IROS*, 2022. 9, 20

参考

[1] 爱德华·H·阿德尔森。关于观察事物：人类和机器对材料的感知。人类视觉与电子成像VI，2001. 5

[2] 博格丹·亚历克斯、托马斯·德塞拉尔斯和维托里奥·法拉利。什么是对象？ CVPR, 2010. 4, 10

[3] Pablo Arbel´ aez、Michael Maire、Charles Fowlkes 和 Jitendra Malik。轮廓检测和分层图像分割。 TPAMI, 2010. 4, 10, 21, 28

[4] 吉米·雷巴、杰米·瑞安·基罗斯和杰弗里·E·辛顿。层标准化。 arXiv:1607.06450, 2016. 16

[5]包航波，李东，魏福如。 BEiT：图像转换器的 BERT 预训练。 arXiv:2106.08254, 2021. 17

[6] Dina Bashkirova、Mohamed Abdelfattah、Ziliang Zhu、James Akl、Fadi Alladkani、Ping Hu、Vitaly Ablavsky、Berk Calli、Sarah Adel Bargal 和 Kate Saenko。 ZeroWaste 数据集：在杂乱场景中实现可变形对象分割。 CVPR，2022 年 9 月 20 日

[7] Stuart Berg、Dominik Kutra、Thorben Kroeger、Christoph N. Straehle、Bernhard X. Kausler、Carsten Haubold、Martin Schiegg、Janez Ales、Thorsten Beier、Markus Rudy、Kemal Eren、Jaime I. Cervantes、Buote Xu、Fynn Beuttenmueller 、Adrian Wolny、Chong Chang、Ullrich Koethe、Fred A. Hamprecht 和 Anna Kreshuk。 ilastik：用于（生物）图像分析的交互式机器学习。自然方法，2019. 12

[8] Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill 等。论基础模型的机遇与风险。 arXiv:2108.07258, 2021. 1, 12

[9] 古斯塔夫·布雷德尔、克里斯汀·坦纳和恩德·科努科格鲁。分段编辑网络的迭代交互训练。米凯，2018. 17

[10] 汤姆·布朗、本杰明·曼、尼克·莱德、梅兰妮·苏比亚、贾里德·D·卡普兰、普拉富拉·达里瓦尔、阿尔温德·尼拉坎坦、普拉纳夫·希亚姆、吉里什·萨斯特里、阿曼达·阿斯克尔、桑迪尼·阿加瓦尔、阿里尔·赫伯特-沃斯、格雷琴·克鲁格、汤姆·赫尼汉、Rewon Child 、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。语言模型是小样本学习者。 NeurIPS, 2020. 1, 4

[11]蔡兆伟，努诺·瓦斯康塞洛斯。 Cascade R-CNN：深 、梁楠. MVS 建筑物的 3D 实例分割。 IEEE 地球科学与遥感学报，2022 年。9、19、20、23、24

[18] 陈希，赵志彦，张一蕾，段漫妮，齐栋联，赵恒爽。 FocalClick：走向实用的交互式图像分割。 CVPR，2022

19 [19] Bowen Cheng、Ishan Misra、Alexander G Schwing、Alexander Kirillov 和 Rohit Girdhar。用于通用图像分割的掩模注意力掩模变换器。 CVPR，2022

像素分类并不是语义分割所需的全部。 NeurIPS, 2021. 5, 16, 17

[21] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。 PaLM：通过路径扩展语言建模。 arXiv:2204.02311, 2022. 1

[22] 卢卡·钱皮、卡洛斯·圣地亚哥、若昂·科斯特拉、克劳迪奥·根纳罗和朱塞佩·阿马托。交通密度估计的域适应。计算机视觉、成像与计算机图形学理论与应用国际联合会议，2021年9月20日

[23] 卢卡·钱皮、卡洛斯·圣地亚哥、若昂·科斯特拉、克劳迪奥·根纳罗和朱塞佩·阿马托。夜间和白天实例分割公园（NDISPark）数据集：白天和夜间拍摄的图像集合，用于停车区域的车辆检测、分割和计数。泽诺多, 2022. 9, 20

[24] 纳达夫·科恩、雅埃尔·纽曼和阿里尔·沙米尔。艺术绘画中的语义分割。计算机图形学论坛，2022. 9, 19, 20, 23, 24

[25] Marius Cordts、Mohamed Omran、Sebastian Ramos、Timo Rehfeld、Markus Enzweiler、Rodrigo Benenson、Uwe Franke、Stefan Roth 和 Bernt Schiele。用于语义城市场景理解的 Cityscapes 数据集。 CVPR, 2016. 9, 19, 20

[26] 布鲁诺·达席尔瓦、乔治·科尼达里斯和安德鲁·巴范德马滕。物体识别适合所有人吗？ CVPR 研讨会，2019. 18

[30] 马克·德·亚兹、伊恩·基夫利坎、雷切尔·罗森、迪伦·贝克、拉兹万·阿米罗内塞、维诺德库马尔·普拉巴卡兰和艾米丽·丹顿。 CrowdWorkSheets：说明众包数据集注释背后的个人和集体身份。 ACM 公平、问责和透明度会议，2022 年。 25

[31] 丁恒辉，斯科特·科恩，布莱恩·普莱斯，蒋旭东。 PhraseClick：通过短语和点击实现灵活的交互式细分。 ECCV，2020. 11

[32] Piotr Doll´ ar 和 C Lawrence Zitnick。使用结构化森林进行快速边缘检测。 TPAMI，2014年。21

[33] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像相当于 16x16 个单词：用于大规模图像识别的 Transformer。 ICLR，2021 年第 5、8、16 期

arXiv:2003.07932, 2020. 5, 17

[38] Jean-Michel Fortin、Olivier Gamache、Vincent Grondin、Fran ¸ois Pomerleau 和 Philippe Gigu`ere。林

[39] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021. 25

[40] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *CVPR*, 2021. 16, 18, 22

[41] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2014. 10

[42] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv:1706.02677*, 2017. 17

[43] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. *CVPR*, 2022. 20

[44] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. *CVPR*, 2019. 2, 6, 7, 9, 10, 11, 19, 20, 21, 24

[45] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *NeurIPS*, 2012. 5, 17

[46] Timm Haucke, Hjalmar S. Kühl, and Volker Steinhage. SOCRATES: Introducing depth in visual wildlife monitoring using stereo vision. *Sensors*, 2022. 9, 20

[47] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022. 5, 8, 12, 16, 17

[48] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *ICCV*, 2017. 10

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 16

[50] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv:1606.08415*, 2016. 16

[51] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv:2203.15556*, 2022. 1

[52] Jungseok Hong, Michael Fulton, and Junaed Sattar. TrashCan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv:2007.08097*, 2020. 9, 19, 20

[53] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. *ECCV*, 2016. 17

[54] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. *arXiv:2211.06220*, 2022. 4

[55] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*, 2021. 1

[56] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020. 1

[57] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *IJCV*, 1988. 4

[58] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 2022. 21

[59] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *CVPR*, 2019. 4

[60] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2, 6, 7, 18, 19

[61] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv:1910.09700*, 2019. 28

[62] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ECCV*, 2022. 5, 10, 11, 16, 21, 23, 24

[63] Yin Li, Zhefan Ye, and James M. Rehg. Delving into egocentric actions. *CVPR*, 2015. 9, 20

[64] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. *CVPR*, 2018. 5, 17, 19

[65] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *ICCV*, 2017. 5, 17

[66] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. *ECCV*, 2014. 2, 4, 6, 7, 11, 18, 19, 20

[67] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. SimpleClick: Interactive image segmentation with simple vision transformers. *arXiv:2210.11006*, 2022. 8, 9, 12, 19

[68] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 17

[69] Cathy H Lucas, Daniel OB Jones, Catherine J Hollyhead, Robert H Condon, Carlos M Duarte, William M Graham, Kelly L Robinson, Kylie A Pitt, Mark Schildhauer, and Jim Regetz. Gelatinous zooplankton biomass in the global oceans: geographic variation and environmental drivers. *Global Ecology and Biogeography*, 2014. 20

[70] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. *BMVC*, 2018. 4, 17

[71] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. *CVPR*, 2018. 6

[72] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV*, 2001. 10, 21, 28

[73] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *3DV*, 2016. 5, 17

[74] Massimo Minervini, Andreas Fischbach, Hanno Scharr, and Sotirios A. Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognition Letters*, 2016. 9, 20

[75] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the conference on fairness, accountability, and transparency*, 2019. 25, 28

[39] Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna Wallach、Hal Daum´ e Iii 和 Kate Crawford。数据集的数据表。 ACM 通讯，2021 年。 25

[40] Golnaz Ghiasi、Yin Cui、Aravind Srinivas、Rui Qi、Tsung–Yi Lin、Ekin D Cubuk、Quoc V Le 和 Barret Zoph。简单的复制粘贴是用于实例分割的强大数据增强方法。 CVPR，2021 年 16、18、22

[41]罗斯·吉尔希克、杰夫·多纳休、特雷弗·达雷尔和吉腾德拉·马利克。丰富的特征层次结构，用于准确的对象检测和语义分割。 CVPR，2014 年。10

[42] Priya Goyal、Piotr Doll´ ar、Ross Girshick、Pieter Noordhuis、Lukasz Wesolowski、Aapo Kyrola、Andrew Tulloch、Yangqing Jia 和 Kaiming He。准确的大型小批量 SGD：1 小时内训练 ImageNet。 arXiv:1706.02677, 2017. 17

[43] Kristen Grauman、Andrew Westbury、Eugene Byrne、Zachary Chavis、Antonino Furnari、Rohit Girdhar、Jackson Hamilton、Hao Jiang、Miao Liu、Xingyu Liu、Miguel Martin、Tushar Nagarajan、Ilija Radosavovic、Santhosh Kumar Ramakrishnan、Fiona Ryan、Jayant夏尔马、迈克尔·雷、徐萌萌、徐忠聪、赵晨、Siddhant Bansal、Dhruv Batra、Vincent Cartilier、Sean Crane、Tien Do、Morrie Doulaty、Akshay Erapalli、Christoph Feichtenhofer、Adriano Fragomeni、Qichen Fu、Christian Fuegen、Abrham Gebreselasie , Cristina Gonzalez, James Hillis, 黄旭华, 黄一飞, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, 大词汇量实例分割。 CVPR，2019. 2, 6, 7, 9, 10, 11, 19, 20, 21, 24

[45] Abner Guzman–Rivera、Dhruv Batra 和 Pushmeet Kohli。多项选择学习：学习产生多种结构化输出。 NeurIPS, 2012. 5, 17 [46] Timm Haucke、Hjalmar S. 苏格拉底：引入立体视觉深度视觉野生动物监测。传感器，2022 年 9 月 20 日

[47] 何凯明，陈新雷，谢赛宁，李阳浩，Piotr Doll´ ar，Ross Girshick。屏蔽自动编码器是可扩展的视觉学习器。 CVPR，2022 年。5、8、12、16、17

[48] Kaiming He、Georgia Gkioxari、Piotr Doll´ ar 和 Ross Girshick。掩模 R–CNN。国际商会，2017. 10

[49] 何凯明，张翔宇，任少清，孙健．用于图像识别的深度残差学习。 CVPR，2016. 16

[50] 丹·亨德里克斯和凯文·金佩尔。高斯误差线性单位 (gelus)。 arXiv:1606.08415, 2016. 16

1

[52] Jungseok Hong、Michael Fulton 和 Junaed Sattar。 TrashCan：用于视觉检测海洋垃圾的语义分段数据集。 arXiv:2007.08097, 2020. 9, 19, 20

[53] 高晃，孙宇，刘庄，Daniel Sedra，Kilian Q Weinberger。具有随机深度的深层网络。欧洲CV，2016. 17

4 [55] Chao Jia，Yinfei Yang，Ye Xia，Yi–Ting Chen，Zarana Parekh，Hieu Pham，Quoc Le，Yun–Hsuan Sung，Zhen Li 和 Tom Duerig。通过嘈杂的文本监督扩大视觉和视觉语言表示学习。 ICML，2021. 1

[56] 贾里德·卡普兰、萨姆·麦坎德利什、汤姆·赫尼汉、汤姆·B·布朗、本杰明·切斯、雷旺·柴尔德、斯科特·格雷、亚历克·雷德福德、杰弗里·吴和达里奥·阿莫迪。神经语言模型的缩放定律。 arXiv:2001.08361, 2020.1

学习分类。 IEEE 机器人和自动化快报，2022 年。 21

[59] 亚历山大·基里洛夫、何凯明、罗斯·吉尔希克、卡斯滕·罗瑟和彼得·多尔。全景分割。 CVPR，2019. 4

[60] Alina Kuznetsova、Hassan Rom、Neil Alldrin、Jasper Uijlings、Ivan Krasin、Jordi Pont–Tuset、Shahab Kamali、Stefan Popov、Matteo Malloci、Alexander Kolesnikov、Tom Duerig 和 Vittorio Ferrari。开放图像数据集 v4：大规模统一图像分类、对象检测和视觉关系检测。国际JCV，2020.2,6,7,18,19

[61] 亚历山大·拉科斯特、亚历山德拉·卢乔尼、维克多·施密特和托马斯·丹德雷斯。量化机器学习的碳排放。 arXiv:1910.09700, 2019. 28

[62] 李阳浩，毛汉子，罗斯·吉尔希克，何凯明。探索用于物体检测的普通视觉变压器主干。 ECCV, 2022. 5, 10, 11, 16, 21, 23, 24

佩罗纳、Deva Ramanan、Piotr Doll´ ar 和 C Lawrence Zitnick。 Microsoft COCO：上下文中的常

[67]刘勤，徐振林，Gedas Bertasius，Marc Niethammer。 SimpleClick：使用简单的视觉转换器进行交互式图像分割。 arXiv:2210.11006, 2022.8,9,12,19

[68] 伊利亚·洛什奇洛夫和弗兰克·哈特。解耦权重衰减正则化。 ICLR，2019. 17

[69] 凯西·H·卢卡斯、丹尼尔·OB·琼斯、凯瑟琳·J·霍利黑德、罗伯特·H·康登、卡洛斯·M·杜阿尔特、威廉·M·格雷厄姆、凯利·L·罗宾逊、凯莉·A·皮特、马克·希尔豪尔和吉姆·雷杰茨。全球海洋中的胶状浮游动物生物量：地理变化和环境驱动因素。全球生态与生物地理学，2014 年。20

[70] Sabarinath Mahadevan、Paul Voigtlaender 和 Bastian Leibe。迭代训练的交互式分割。 BMVC, 2018. 4, 17

[71]Kevis–Kokitsi Maninis、Sergi Caelles、Jordi Pont–Tuset 和 Luc Van Gool。深度极值切割：从极值点到对象分割。 CVPR，2018. 6

[72] 大卫·马丁、查尔斯·福克斯、多伦·塔尔和吉腾德拉·马利克。人类分割自然图像的数据库及其在评估分割算法和测量生态统计方面的应用。 ICCV，2001年。10、21、28

[76] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. *ICCV*, 2017. 6

[77] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv:2104.10350*, 2021. 28

[78] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017. 18

[79] Mengyang Pu, Yaping Huang, Yuming Liu, Qingji Guan, and Haibin Ling. EDTER: Edge detection with transformer. *CVPR*, 2022. 10

[80] Mattia Pugliatti and Francesco Topputo. DOORS: Dataset fOr bOuldeRs Segmentation. *Zenodo*, 2022. 9, 20

[81] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *ICCV*, 2022. 9, 20, 23, 24

[82] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021. 1, 2, 4, 5, 8, 12, 16, 22

[83] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ICML*, 2021. 1, 4, 12

[84] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 6, 10

[85] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. *ICCV*, 2003. 4

[86] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *ICCV*, 2021. 9, 19, 20

[87] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. A step toward more inclusive people annotations for fairness. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021. 8, 19

[88] Sefik Ilkin Serengil and Alper Ozpinar. LightFace: A hybrid deep face recognition framework. *ASYU*, 2020. 26

[89] Sefik Ilkin Serengil and Alper Ozpinar. HyperExtended LightFace: A facial attribute analysis framework. *ICEET*, 2021. 26

[90] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. TextonBoost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation. *ECCV*, 2006. 4

[91] Corey Snyder and Minh Do. STREETS: A novel camera network dataset for traffic flow. *NeurIPS*, 2019. 9, 20

[92] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. *ICIP*, 2022. 5, 8, 9, 17, 19, 23, 24, 28

[93] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 2014. 16

[94] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. *CVPR*, 1999. 4

[95] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020. 5, 16

[96] Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Action recognition in RGB-D egocentric videos. *ICIP*, 2017. 20

[97] Yansong Tang, Zian Wang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Multi-stream deep neural networks for RGB-D egocentric action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 20

[98] The World Bank. The world by income and regions, 2022. https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html. 18

[99] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? *NeurIPS*, 1995. 12

[100] Cameron Trotter, Georgia Atkinson, Matt Sharpe, Kirsten Richardson, A. Stephen McGough, Nick Wright, Ben Burville, and Per Berggren. NDD20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. *arXiv:2005.13359*, 2020. 9, 19, 20, 23, 24

[101] United States Environmental Protection Agency. Greenhouse Gas Equivalencies Calculator. https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator, 2022. 28

[102] Koen EA van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. Segmentation as selective search for object recognition. *ICCV*, 2011. 10

[103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 5, 16

[104] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. *CVPR*, 2021. 9, 19, 20

[105] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. *CVPR*, 2022. 21

[106] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3D reconstruction. *CVPR*, 2023. 12

[107] Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. *CVPR*, 2010. 20

[108] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *ICCV*, 2015. 10

[109] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. *CVPR*, 2016. 4, 19

[110] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020. 8

[111] Lei Yang, Yan Zi Wei, Yisheng HE, Wei Sun, Zhenhang Huang, Haibin Huang, and Haoqiang Fan. iShape: A first step towards irregular shape instance segmentation. *arXiv:2109.15068*, 2021. 9, 20, 23, 24

[112] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. WoodScape: A multi-task, multi-camera fisheye dataset for autonomous driving. *ICCV*, 2019. 9, 20

[113] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. *ECCV*, 2022. 9, 19, 20

[114] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards unified image segmentation. *NeurIPS*, 2021. 4

[115] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv:1707.09457*, 2017. 8

[116] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 20

[117] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *IJCV*, 2019. 2, 7, 9, 20

[82] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。从自然语言监督中学习可迁移的视觉模型。 ICML，2021 年。1、2、4、5、8、12、16、22

[83] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零样本文本到图像生成。 ICML，2021.1,4,12

[84] 任少清，何凯明，Ross Girshick，孙健。 Faster R-CNN：通过区域提议网络实现实时目标检测。 NeurIPS, 2015. 6, 10

[85] 任晓峰，Jitendra Malik。学习用于分割的分类模型。国际商业联合会，2003年。4

[86] Mike Roberts、Jason Ramapuram、Anurag Ranjan、Atulit Kumar、Miguel Angel Bautista、Nathan Paczan、Russ Webb 和 Joshua M. Susskind。 Hypersim：用于整体室内场景理解的逼真合成数据集。 ICCV，2021. 9, 19, 20

[87] 坎迪斯·舒曼、苏珊娜·里科、乌察夫·普拉布、维托里奥·法拉利和卡罗琳·潘托法鲁。朝着更具包容性的人员注释迈出一步，以实现公平。 2021 年 AAAI/ACM 人工智能、伦理与社会会议论文集，2021. 8, 19

[88] 塞菲克·伊尔金·塞伦吉尔和阿尔珀·奥兹皮纳尔。 LightFace：混合深度人脸识别框架。纽约大学，2020. 26

[89] 塞菲克·伊尔金·塞伦吉尔和阿尔珀·奥兹皮纳尔。 HyperExtended LightFace：面部属性分析框架。 ICEET，2021. 26

[90] 杰米·肖顿、约翰·温、卡斯滕·罗瑟和安东尼奥·克里米尼西。 TextonBoost：用于多类对象识别和分割的联合外观、形状和上下文建模。欧洲CV，2006年。4

[91] 科里·斯奈德和明·杜。 STREETS：一种新颖的交通流量摄像机网络数据集。 NeurIPS，2019 年 9 月 20 日

[92] 康斯坦丁·索菲乌克、伊利亚·A·彼得罗夫和安东·科努申。通过掩模指导恢复交互式分割的迭代训练。 ICIP，2022 年。5、8、9、17、19、23、24、28

[93] Nitish Srivastava、Geoffrey Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。 Dropout：防止神经网络过度拟合的简单方法。机器学习研究杂志，2014 年。 16

[94]克里斯·斯托弗和W·埃里克·格里姆森。用于实时跟踪的自适应背景混合模型。 CVPR，1999。4

[95]Matthew Tancik、Pratul Srinivasan、Ben Mildenhall、Sara Fridovich-Keil、Nithin Raghavan、Utkarsh Singhal、Ravi Ramamoorthi、Jonathan Barron 和 Ren Ng。傅立叶特征使网络能够学习低维域中的高频函数。 NeurIPS，2020 年 5 月 16 日

20 [97] 唐岩松，王子安，路继文，冯建江，周杰。用于 RGB-D 自我中心动作识别的多流深度神经网络。 IEEE 视频技术电路和系统汇刊，2019 年。 20

[98] 世界银行。按收入和地区划分的世界

2022.

[99] 塞巴斯蒂安·特龙。学习第 n 件事比学习第一件事更容易吗？神经IPS，1995。12

[100] Cameron Trotter、Georgia Atkinson、Matt Sharpe、Kirsten Richardson、A. Stephen McGough、Nick Wright、Ben Burville 和 Per Berggren。 NDD20：用于粗粒度和细粒度分类的大规模少镜头海豚数据集。 arXiv:2005.13359, 2020.9,19,20,23,24

[101]美国环境保护署。温室气体气体当量计算器，2022 年。 28

[102]Koen EA van de Sande、Jasper RR Uijlings、Theo Gevers 和 Arnold WM Smeulders。分割作为对象识别的选择性搜索。国际商业联合会，2011. 10

[103] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin。您所需要的就是关注。 NeurIPS，2017 年 5 月 16 日

[104]王博英，张立波，温龙银，刘翔龙，吴彦军。实现现实世界违禁物品检测：大规模 X 射线基准。 CVPR，2021 年 9、19、20

[105] 王伟耀、Matt Feiszli、Heng Wang、Jitendra Malik 和 Du Tran。开放世界实例分割：从学习的成对亲和力中利用伪地面事实。 CVPR，2022. 21

[106]吴朝元、Justin Johnson、Jitendra Malik、Christoph Feichtenhofer 和 Georgia Gkioxari。用于 3D 重建的多视图压缩编码。 CVPR，2023. 12

[107] 肖建雄、詹姆斯·海斯、克里斯塔·埃因格、奥德·奥利瓦和安东尼奥·托拉尔巴。 SUN数据库：从修道院到动物园的大规模场景识别。 CVPR，2010 年。20

[108]谢赛宁，屠卓文。整体嵌套边缘检测。国际商会，2015. 10

Padraig Varley、Derek O'Dea、Michal Uric´ar、Stefan Milz、Martin Simon、Karl Amende 等。 WoodScape：用于自动驾驶的多任务、多摄像头鱼眼

[113] 张凌志，周胜浩，西蒙·斯坦特，石建波。细粒度的以自我为中心的手部物体分割：数据集、模型和应用程序。 ECCV, 2022. 9, 19, 20

[114] 张文伟，庞江淼，陈凯，陈变洛。 K-Net：迈向统一图像分割。 NeurIPS，2021. 4

[115] 赵杰宇、王天禄、Mark Yatskar、Vicente Ordonez

托尼奥·托拉尔巴Places：1000万张图像数据库，用于场景识别。 TPAMI，2017. 20

[117]周博雷、赵航、泽维尔·普伊格、肖特特、桑贾·费德勒、阿德拉·巴留索和安东尼奥·托拉尔巴。通过 ADE20K

Appendix

Table of contents:

A. Segment Anything Model and Task Details

Image encoder. In general, the image encoder can be any network that outputs a $C \times H \times W$ image embedding. Motivated by scalability and access to strong pre-training, we use an MAE [47] pre-trained Vision Transformer (ViT) [33] with minimal adaptations to process high resolution inputs, specifically a ViT-H/16 with 14×14 windowed attention and four equally-spaced global attention blocks, following [62]. The image encoder's output is a $16\times$ downscaled embedding of the input image. Since our runtime goal is to process each prompt in real-time, we can afford a high number of image encoder FLOPs because they are computed only once per image, *not* per prompt.

Following standard practices (*e.g.*, [40]), we use an input resolution of 1024×1024 obtained by rescaling the image and padding the shorter side. The image embedding is therefore 64×64. To reduce the channel dimension, following [62], we use a 1×1 convolution to get to 256 channels, followed by a 3×3 convolution also with 256 channels. Each convolution is followed by a layer normalization [4].

Prompt encoder. Sparse prompts are mapped to 256-dimensional vectorial embeddings as follows. A point is represented as the sum of a positional encoding [95] of the point's location and one of two learned embeddings that indicate if the point is either in the foreground or background. A box is represented by an embedding pair: (1) the positional encoding of its top-left corner summed with a learned embedding representing "top-left corner" and (2) the same structure but using a learned embedding indicating "bottom-right corner". Finally, to represent free-form text we use the text encoder from CLIP [82] (any text encoder is possible in general). We focus on geometric prompts for the remainder of this section and discuss text prompts in depth in §D.5.

Dense prompts (*i.e.*, masks) have a spatial correspondence with the image. We input masks at a $4\times$ lower resolution than the input image, then downscale an additional $4\times$ using two 2×2, stride-2 convolutions with output channels 4 and 16, respectively. A final 1×1 convolution maps the channel dimension to 256. Each layer is separated by GELU activations [50] and layer normalization. The mask and image embedding are then added element-wise. If there is no mask prompt, a learned embedding representing "no mask" is added to each image embedding location.

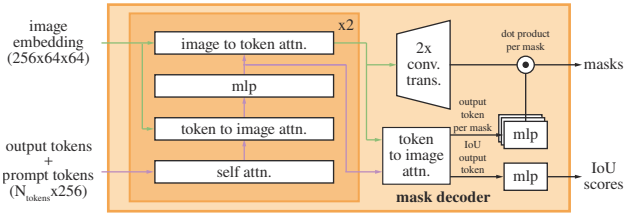


Figure 14: Details of the lightweight mask decoder. A two-layer decoder updates both the image embedding and prompt tokens via cross-attention. Then the image embedding is upscaled, from which the updated output tokens are used to dynamically predict masks. (Not illustrated for figure clarity: At every 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.)

Lightweight mask decoder. This module efficiently maps the image embedding and a set of prompt embeddings to an output mask. To combine these inputs, we take inspiration from Transformer segmentation models [14, 20] and modify a standard Transformer decoder [103]. Before applying our decoder, we first insert into the set of prompt embeddings a learned output token embedding that will be used at the decoder's output, analogous to the `[class]` token in [33]. For simplicity, we refer to these embeddings (*not* including the image embedding) collectively as "tokens".

Our decoder design is shown in Fig. 14. Each decoder layer performs 4 steps: (1) self-attention on the tokens, (2) cross-attention from tokens (as queries) to the image embedding, (3) a point-wise MLP updates each token, and (4) cross-attention from the image embedding (as queries) to tokens. This last step updates the image embedding with prompt information. During cross-attention, the image embedding is treated as a set of 64^2 256-dimensional vectors. Each self/cross-attention and MLP has a residual connection [49], layer normalization, and a dropout [93] of 0.1 at training. The next decoder layer takes the updated tokens and the updated image embedding from the previous layer. We use a two-layer decoder.

To ensure the decoder has access to critical geometric information the positional encodings are added to the image embedding whenever they participate in an attention layer. Additionally, the *entire* original prompt tokens (including their positional encodings) are re-added to the updated tokens whenever they participate in an attention layer. This allows for a strong dependence on both the prompt token's geometric location and type.

After running the decoder, we upsample the updated image embedding by $4\times$ with two transposed convolutional

附录

目录：

A. 分割任何模型和任务细节

图像编码器。一般来说，图像编码器可以是任何输出 C×H ×W 图像嵌入的网络。出于可扩展性和强大的预训练的动机，我们使用 MAE [47] 预训练的 Vision Transformer (ViT) [33] 进行最小的调整来处理高分辨率输入，特别是具有 14×14 窗口的 ViT–H/16注意力和四个等距的全局注意力块，如下[62]。图像编码器的输出是输入图像的 16 倍缩小嵌入。由于我们的运行时目标是实时处理每个提示，因此我们可以承受大量的图像编码器 FLOP，因为它们仅针对每个图像而不是每个提示计算一次。

遵循标准实践（例如，[40]），我们使用通过重新缩放图像并填充较短边获得的 1024×1024 输入分辨率。因此图像嵌入是 64×64。要减少通道尺寸，请遵循

根据[62]，我们使用 1×1 卷积来达到 256 个通道

nels，然后是同样具有 256 个通道的 3×3 卷积。每个卷积后面都有一个层归一化[4]。

提示编码器。稀疏提示映射到 256 维向量嵌入，如下所示。一个点被表示为该点位置的位置编码[95]和两个学习嵌入之一的总和，这两个嵌入表示该点是在前景还是背景。盒子由嵌入对表示：（1）其左上角的位置编码与表示“左上角”的学习嵌入相加；（2）相同的结构，但使用表示“右下角”的学习嵌入。最后，为了表示自由格式的文本，我们使用 CLIP [82] 中的文本编码器（通常任何文本编码器都是可能的）。我们将在本节的其余部分重点讨论几何提示，并在 §D.5 中深入讨论文本提示。

密集提示（即掩模）与图像具有空间对应关系。我们以比输入图像低 4 倍的分辨率输入掩码，然后使用两个 2×2、stride–2 卷积和输出通道将其再缩小 4 倍。

分别为第 4 条和第 16 条。最终的 1×1 卷积将通道维度映射到 256。每一层都由 GELU 激活 [50] 和层归一化分隔。面具

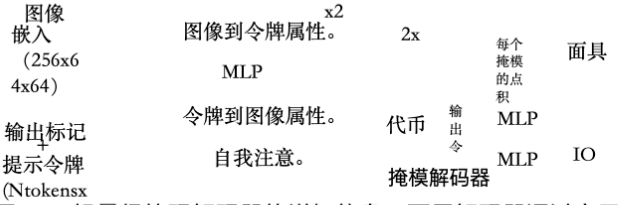


图 14：轻量级掩码解码器的详细信息。两层解码器通过交叉注意力更新图像嵌入和提示标记。然后，图像嵌入被放大，更新的输出标记用于动态预测掩模。（为了图形清晰，未示出：在每个注意层，位置编码被添加到图像嵌入中，并且整个原始提示标记（包括位置编码）被重新添加到标记查询和键中。）

然后按元素添加图像嵌入。如果没有掩模提示，则将表示“无掩模”的学习嵌入添加到每个图像嵌入位置。

轻量级掩模解码器。该模块有效地将图像嵌入和一组提示嵌入映射到输出掩码。为了结合这些输入，我们从 Transformer 分割模型 [14, 20] 中汲取灵感，并修改标准 Transformer 解码器 [103]。在应用解码器之前，我们首先将学习的输出标记嵌入插入到提示嵌入集中，该输出标记嵌入将在解码器的输出中使用，类似于[33]中的[class]标记。为了简单起见，我们将这些嵌入（不包括图像嵌入）统称为“令牌”。

我们的解码器设计如图 14 所示。每个解码器层执行 4 个步骤：(1) 对标记进行自我关注，(2) 从标记（作为查询）到图像嵌入的交叉关注，(3) 点–明智的 MLP 更新每个标记，以及 (4) 从图像嵌入（作为查询）到标记的交叉注意力。最后一步用提示信息更新图像嵌入。在交叉注意期间，图像em–

床上用品被视为一组 64 256 维向量。每个自/交叉注意力和 MLP 都有一个剩余连接

化[49]、层归一化和训练时的dropout[93]为0.1。下一个解码器层从上一层获取更新的令牌和更新的图像嵌入。我们使用两层解码器。

为了确保解码器能够访问关键的几何信息，只要参与关注层，位置编码就会添加到图像嵌入中。此外，每当它们参与注意力层时，整个原始提示标记（包括它们的位置编码）都会被重新添加到更新的标记中。这允许对提示标记的几何位置和类型有很强的依赖性。

运行解码器后，我们使用两个转置卷积对更新后的图像嵌入进行 4 倍上采样

layers (now it's downscaled 4× relative to the input image). Then, the tokens attend once more to the image embedding and we pass the updated output token embedding to a small 3-layer MLP that outputs a vector matching the channel dimension of the upscaled image embedding. Finally, we predict a mask with a spatially point-wise product between the upscaled image embedding and the MLP's output.

The transformer uses an embedding dimension of 256. The transformer MLP blocks have a large internal dimension of 2048, but the MLP is applied only to the prompt tokens for which there are relatively few (rarely greater than 20). However, in cross-attention layers where we have a 64×64 image embedding, we reduce the channel dimension of the queries, keys, and values by 2× to 128 for computational efficiency. All attention layers use 8 heads.

The transposed convolutions used to upscale the output image embedding are 2×2, stride 2 with output channel dimensions of 64 and 32 and have GELU activations. They are separated by layer normalization.

Making the model ambiguity-aware. As described, a single input prompt may be ambiguous in the sense that it corresponds to multiple valid masks, and the model will learn to average over these masks. We eliminate this problem with a simple modification: instead of predicting a single mask, we use a small number of output tokens and predict multiple masks simultaneously. By default we predict three masks, since we observe that three layers (whole, part, and subpart) are often enough to describe nested masks. During training, we compute the loss (described shortly) between the ground truth and each of the predicted masks, but only backpropagate from the lowest loss. This is a common technique used for models with multiple outputs [15, 45, 64]. For use in applications, we'd like to rank predicted masks, so we add a small head (operating on an additional output token) that estimates the IoU between each predicted mask and the object it covers.

Ambiguity is much rarer with multiple prompts and the three output masks will usually become similar. To minimize computation of degenerate losses at training and ensure the single unambiguous mask receives a regular gradient signal, we only predict a single mask when more than one prompt is given. This is accomplished by adding a fourth output token for an additional mask prediction. This fourth mask is never returned for a single prompt and is the only mask returned for multiple prompts.

Losses. We supervise mask prediction with a linear combination of focal loss [65] and dice loss [73] in a 20:1 ratio of focal loss to dice loss, following [20, 14]. Unlike [20, 14], we observe that auxiliary deep supervision after each decoder layer is unhelpful. The IoU prediction head is trained with mean-square-error loss between the IoU prediction and the predicted mask's IoU with the ground truth mask. It is added to the mask loss with a constant scaling factor of 1.0.

Training algorithm. Following recent approaches [92, 37], we simulate an interactive segmentation setup during training. First, with equal probability either a foreground point or bounding box is selected randomly for the target mask. Points are sampled uniformly from the ground truth mask. Boxes are taken as the ground truth mask's bounding box, with random noise added in each coordinate with standard deviation equal to 10% of the box sidelength, to a maximum of 20 pixels. This noise profile is a reasonable compromise between applications like instance segmentation, which produce a tight box around the target object, and interactive segmentation, where a user may draw a loose box.

After making a prediction from this first prompt, subsequent points are selected uniformly from the error region between the previous mask prediction and the ground truth mask. Each new point is foreground or background if the error region is a false negative or false positive, respectively. We also supply the mask prediction from the previous iteration as an additional prompt to our model. To provide the next iteration with maximal information, we supply the unthresholded mask logits instead of the binarized mask. When multiple masks are returned, the mask passed to the next iteration and used to sample the next point is the one with the highest predicted IoU.

We find diminishing returns after 8 iteratively sampled points (we have tested up to 16). Additionally, to encourage the model to benefit from the supplied mask, we also use two more iterations where no additional points are sampled. One of these iterations is randomly inserted among the 8 iteratively sampled points, and the other is always at the end. This gives 11 total iterations: one sampled initial input prompt, 8 iteratively sampled points, and two iterations where no new external information is supplied to the model so it can learn to refine its own mask predictions. We note that using a relatively large number of iterations is possible because our lightweight mask decoder requires less than 1% of the image encoder's compute and, therefore, each iteration adds only a small overhead. This is unlike previous interactive methods that perform only one or a few interactive steps per optimizer update [70, 9, 37, 92].

Training recipe. We use the AdamW [68] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and a linear learning rate warmup [42] for 250 iterations and a step-wise learning rate decay schedule. The initial learning rate (*lr*), after warmup, is 8e^{-4}. We train for 90k iterations (~2 SA-1B epochs) and decrease the *lr* by a factor of 10 at 60k iterations and again at 86666 iterations. The batch size is 256 images. To regularize SAM, we set weight decay (*wd*) to 0.1 and apply drop path [53] (*dp*) with a rate of 0.4. We use a layer-wise learning rate decay [5] (*ld*) of 0.8. No data augmentation is applied. We initialize SAM from an MAE [47] pre-trained ViT-H. We distribute training across 256 GPUs, due to the large image encoder and 1024×1024 input size. To limit GPU mem-

层（现在相对于输入图像缩小了 4 倍）。然后，令牌再次参与图像嵌入，我们将更新后的输出令牌嵌入传递给小型 3 层 MLP，该 MLP 输出与放大的图像嵌入的通道尺寸相匹配的向量。最后，我们用放大图像嵌入和 MLP 输出之间的空间逐点乘积来预测掩模。

Transformer 使用的嵌入维度为 256。Transformer MLP 块的内部维度较大，为 2048，但 MLP 仅应用于相对较少（很少大于 20）的提示标记。然而，在交叉注意力层中，我们有 64×64 图像嵌入，我们减少了通道维度

查询、键和值的数量增加了 2 倍到 128，用于计算

效率。所有注意力层都使用 8 个头。

用于升级输出图像嵌入的转置卷积为 2×2，步幅为 2，输出通道为 di–

64 和 32 的提及并具有 GELU 激活。它们通过层归一化分开。

使模型具有模糊性意识。如上所述，单个输入提示可能是不明确的，因为它对应于多个有效掩码，并且模型将学习对这些掩码进行平均。我们通过简单的修改消除了这个问题：我们使用少量的输出标记并同时预测多个掩码，而不是预测单个掩码。默认情况下，我们预测三个掩码，因为我们观察到三层（整体、部分和子部分）通常足以描述嵌套掩码。在训练过程中，我们计算真实值和每个预测掩模之间的损失（稍后描述），但仅从最低损失进行反向传播。这是用于具有多个输出的模型的常用技术[15,45,64]。为了在应用程序中使用，我们希望对预测掩码进行排名，因此我们添加一个小头（在附加输出标记上操作）来估计每个预测掩码与其覆盖的对象之间的 IoU。

对于多个提示，歧义性要少得多，并且三个输出掩码通常会变得相似。为了最大限度地减少训练时退化损失的计算并确保单个明确的掩模接收规则的梯度信号，当给出多个提示时，我们仅预测单个掩模。这是通过添加第四个输出标记以进行额外的掩模预测来实现的。这第四个掩码永远不会为单个提示返回，并且是为多个提示返回的唯一掩码。

损失。我们使用焦点损失 [65] 和骰子损失 [73] 的线性组合来监督掩模预测，焦点损失与骰子损失的比率为 20:1，遵循 [20, 14]。与[20, 14]不同，我们观察到每个解码器层之后的辅助深度监督是没有帮助的。 IoU 预测头使用 IoU 预测和预测掩模与真实掩模的 IoU 之间的均方误差损失进行训练。它以恒定比例因子 1.0 添加到掩模损失中。

训练算法。遵循最近的方法 [92, 37]，我们在训练期间模拟交互式分割设置。首先，以相同的概率为目标掩模随机选择前景点或边界框。点是从地面实况掩模中均匀采样的。框被视为真实掩模的边界框，在每个坐标中添加随机噪声，标准差等于框边长的 10%，最大为 20 个像素。此噪声分布是实例分割（在目标对象周围产生紧密的框）和交互式分割（用户可以在其中绘制松散的框）等应用程序之间的合理折衷。根据第一个提示进行预测后，从先前掩模预测和地面真实掩模之间的误差区域中统一选择后续点。如果错误区域分别为假阴性或假阳性，则每个新点分别为前景或背景。我们还提供前一次迭代的掩模预测作为模型的附加提示。为了向下一次迭代提供最大信息，我们提供未阈值掩码 logits 而不是二值化掩码。当返回多个掩码时，传递到下一次迭代并用于采样下一个点的掩码是预测 IoU 最高的掩码。

我们发现在 8 个迭代采样点之后收益递减（我们测试了多达 16 个）。此外，为了鼓励模型从提供的掩模中受益，我们还使用了两次迭代，其中没有采样额外的点。这些迭代之一被随机插入到

8个迭代采样点，另一个始终处于

结尾。这给出了总共 11 次迭代：一次采样初始输入

输入提示、8 个迭代采样点和两次迭代，其中没有向模型提供新的外部信息，因此模型可以学习改进自己的掩模预测。我们注意到，使用相对大量的迭代是可能的，因为我们的轻量级掩模解码器需要图像编码器的不到 1% 的计算，因此，每次迭代仅增加很小的开销。这与以前的交互式方法不同，以前的交互式方法每次优化器更新仅执行一个或几个交互式步骤 [70,9,37,92]。

训练食谱。我们使用 AdamW [68] 优化器 (β1 =

0.9，β2 = 0.999）和线性学习率预热[42]

250 次迭代和逐步学习率衰减时间表。预热后的初始学习率 (lr) 为 8e。我们

训练 90k 次迭代（~2 SA–1B epoch）并减少

lr 在 60k 次迭代时增加了 10 倍，在 86666 次迭代时再次增加了 10 倍 –

饮食。批量大小为 256 张图像。为了规范 SAM，

我们将权重衰减 (wd) 设置为 0.1，并以 0.4 的速率应用下降路径 [53] (dp)。我们使用 0.8 的分层学习率衰减 [5] (ld)。不应用数据增强。我们从 MAE [47] 预训练的 ViT–H 初始化 SAM。我们

由于图像较大，将训练分布在 256 个 GPU 上

编码器和 1024×1024 输入尺寸。限制 GPU 内存

ory usage, we train with up to 64 randomly sampled masks per GPU. Additionally, we find that lightly filtering SA-1B masks to discard any that cover more than 90% of the image qualitatively improves results.

For ablations and others variations on training (*e.g.*, text-to-mask §D.5), we deviate from the default recipe above as follows. When training with data from the first and second data engine stages only, we augment the input with large-scale jitter [40] with a scale range of [0.1, 2.0]. Intuitively, data augmentation may be helpful when training data is more limited. To train ViT-B and ViT-L, we use 180k iterations with batch size 128 distributed across 128 GPUs. We set *lr* = $8e^{-4}/4e^{-4}$, *ld* = 0.6/0.8, *wd* = 0.1, and *dp* = 0.6/0.4 for ViT-B/L, respectively.

B. Automatic Mask Generation Details

Here we discuss details of the data engine's fully automatic stage that was used to generate the released SA-1B.

Cropping. Masks were generated from a regular grid of 32×32 points on the full image and 20 additional zoomed-in image crops arising from 2×2 and 4×4 partially overlapping windows using 16×16 and 8×8 regular point grids, respectively. The original high-resolution images were used for cropping (this was the only time we used them). We removed masks that touch the inner boundaries of the crops. We applied standard greedy box-based NMS (boxes were used for efficiency) in two phases: first within each crop and second across crops. When applying NMS within a crop, we used the model's predicted IoU to rank masks. When applying NMS across crops, we ranked masks from most zoomed-in (*i.e.*, from a 4×4 crop) to least zoomed-in (*i.e.*, the original image), based on their source crop. In both cases, we used an NMS threshold of 0.7.

Filtering. We used three filters to increase mask quality. First, to keep only *confident* masks we filtered by the model's predicted IoU score at a threshold of 88.0. Second, to keep only *stable* masks we compared two binary masks resulting from the same underlying soft mask by thresholding it at different values. We kept the prediction (*i.e.*, the binary mask resulting from thresholding logits at 0) only if the IoU between its pair of -1 and +1 thresholded masks was equal to or greater than 95.0. Third, we noticed that occasionally an automatic mask would cover the entire image. These masks were generally uninteresting, and we filtered them by removing masks that covered 95% or more of an image. All filtering thresholds were selected to achieve both a large number of masks and high mask quality as judged by professional annotators using the method described in §5.

Postprocessing. We observed two error types that are easily mitigated with postprocessing. First, an estimated 4% of masks include small, spurious components. To address these, we removed connected components with area less than 100 pixels (including removing entire masks if the largest component is below this threshold). Second, another estimated 4% of masks include small, spurious holes. To address these, we filled holes with area less than 100 pixels. Holes were identified as components of inverted masks.

Automatic mask generation model. We trained a special version of SAM for fully automatic mask generation that sacrifices some inference speed for improved mask generation properties. We note the differences between our default SAM and the one used for data generation here: it was trained on manual and semi-automatic data only, it was trained for longer (177656 iterations instead of 90k) with large-scale jitter data augmentation [40], simulated interactive training used only point and mask prompts (no boxes) and sampled only 4 points per mask during training (reducing from our default of 9 to 4 sped up training iterations and had no impact on 1-point performance, though it would harm mIoU if evaluating with more points), and finally the mask decoder used 3 layers instead of 2.

SA-1B examples. We show SA-1B samples in Fig. 2. For more examples, please see our dataset explorer.

C. RAI Additional Details

Inferring geographic information for SA-1B. While the images in SA-1B are not geo-tagged, each image has a caption describing its contents and where it was taken. We infer approximate image geo-locations from these captions using an Elmo-based named entity recognition model [78]. Each extracted location entity is mapped to every matching country, province, and city. Captions are mapped to a single country by first considering the matching countries, then provinces, and finally cities. We note that there are ambiguities and potential for biases with this method (*e.g.*, "Georgia" may refer to the country or the US state). As such, we use the extracted locations to analyze the dataset as a whole, but do not release the inferred locations. The captions will not be released publicly as required by the image provider.

Inferring geographic information for COCO and Open Images. The COCO [66] and Open Images [60] datasets do not provide geo-locations. Following [29], we retrieve geographic metadata using the Flickr API. We retrieved locations for 24% of the COCO training set (19,562 images) and for Open Images we retrieved 18% of the training set (493,517 images, after only considering images with masks). We note that the geographic information is approximate, and the sample of images with this information may not fully match the full dataset distribution.

Inferring income information. We use each image's inferred country to look up its income level using the levels defined by The World Bank [98]. We collapse the upper-middle and lower-middle levels into a single middle level.

根据使用情况，我们使用每个 GPU 最多 64 个随机采样的掩模进行训练。此外，我们发现轻微过滤 SA–1B 掩模以丢弃覆盖超过 90% 图像的任何掩模可以在质量上改善结果。

对于训练的消融和其他变化（例如，textto–mask §D.5），我们偏离了上面的默认配方，如下所示。当仅使用来自第一和第二数据引擎阶段的数据进行训练时，我们使用尺度范围为 [0.1, 2.0] 的大规模抖动 [40] 来增强输入。直观上，当训练数据较为有限时，数据增强可能会有所帮助。为了训练 ViT–B 和 ViT–L，我们使用 180k 次迭代，批量大小为 128，分布在 128 个 GPU 上。我们设置 lr = 8e/4e，ld = 0.6/0.8，wd = 0.1，并且

ViT–B/L 的 dp 分别为 0.6/0.4。

B. 自动掩码生成详细信息

在这里，我们讨论用于生成已发布的 SA–1B 的数据引擎全自动阶段的详细信息。

裁剪。掩模是从完整图像上的 32×32 点的规则网格和 20 个附加缩放生成的

在由 2×2 和 4×4 部分过度产生的图像裁剪中

分别使用 16×16 和 8×8 规则点网格的研磨窗口。原始高分辨率图像用于裁剪（这是我们唯一一次使用它们）。我们移除了接触作物内部边界的遮罩。我们分两个阶段应用了标准的基于贪婪框的 NMS（框用于提高效率）：首先是在每种作物内，其次是跨作物。当在作物中应用 NMS 时，我们使用模型的预测 IoU 对掩模进行排名。当在作物上应用 NMS 时，我们对大多数作物的掩模进行了排名

根据源裁剪，放大（即从 4×4 裁剪）到最小放大（即原始图像）。在这两种情况下，我们都使用 0.7 的 NMS 阈值。

过滤。我们使用三个过滤器来提高面罩质量。首先，为了仅保留置信掩码，我们根据模型预测的 IoU 分数（阈值为 88.0）进行过滤。其次，为了仅保持稳定的掩模，我们通过将其阈值设置为不同的值来比较由同一底层软掩模产生的两个二进制掩模。仅当一对 −1 和 +1 阈值掩码之间的 IoU 等于或大于 95.0 时，我们才保留预测（即，阈值 Logits 产生的二进制掩码为 0）。第三，我们注意到有时自动蒙版会覆盖整个图像。这些蒙版通常没什么意思，我们通过删除覆盖图像 95% 或更多区域的蒙版来过滤它们。所有过滤阈值均经过选择，以实现大量掩模和高掩模质量（由专业注释者使用§5中描述的方法判断）。

后期处理。我们观察到两种错误类型可以通过后处理轻松缓解。首先，估计有 4% 的掩模包含小型假元件。为了解决这些

我们删除了面积小于 100 像素的连接组件（包括如果最大组件低于此阈值则删除整个蒙版）。其次，另外估计有 4% 的口罩存在小假孔。到

为了解决这些问题，我们填充了面积小于 100 像素的孔。孔被确定为倒置掩模的组成部分。

自动掩模生成模型。我们训练了一个特殊版本的 SAM 以实现全自动掩模生成，该版本牺牲了一些推理速度以提高掩模生成性能。我们注意到我们的默认 SAM 与此处用于数据生成的 SAM 之间的差异：它仅在手动和半自动数据上进行训练，它通过大规模抖动数据增强进行了更长时间的训练（177656 次迭代而不是 90k）[40]，模拟交互训练仅使用点和掩模提示（无框）

并在训练期间每个掩模仅采样 4 个点（减少

从我们默认的 9 到 4 次加速训练迭代，并且对 1 点性能没有影响，尽管如果使用更多点进行评估会损害 MIoU），最后

掩码解码器使用 3 层而不是 2 层。

SA–1B 示例。我们在图 2 中展示了 SA–1B 样本。有关更多示例，请参阅我们的数据集浏览器。

C. RAI 其他详细信息

推断 SA–1B 的地理信息。虽然 SA–1B 中的图像没有地理标记，但每张图像都有一个说明文字，描述其内容和拍摄地点。我们使用基于 Elmo 的命名实体识别模型 [78] 从这些标题中推断出近似的图像地理位置。每个提取的位置实体都映射到每个匹配的国家、省份和城市。通过首先考虑匹配的国家/地区，然后考虑省份，最后考虑城市，将字幕映射到单个国家/地区。我们注意到这种方法存在歧义和潜在的偏见（例如，“佐治亚州”可能指的是该国或美国的州）。因此，我们使用提取的位置来分析整个数据集，但不发布推断的位置。字幕不会按照图片提供者的要求公开发布。

推断 COCO 和 Open Images 的地理信息。 COCO [66] 和 Open Images [60] 数据集不提供地理位置。遵循 [29]，我们使用 Flickr API 检索地理元数据。我们检索了 24% 的 COCO 训练集（19,562 张图像）的位置，对于开放图像，我们检索了 18% 的训练集 –

ing 集（493,517 张图像，仅考虑带有掩模的图像）。我们注意到，地理信息是近似的，具有此信息的图像样本可能不完全匹配完整的数据集分布。

推断收入信息。我们使用每张图像推断的国家/地区来使用世界银行定义的水平来查找其收入水平[98]。我们将中上层和中下层折叠成一个中间层。

	mIoU at 1 point	mIoU at 3 points
perceived gender presentation		
feminine	76.3 ±1.1	90.7 ±0.5
masculine	81.0 ±1.2	92.3 ±0.4
perceived age group		
older	81.9 ±3.8	92.8 ±1.6
middle	78.2 ±0.8	91.3 ±0.3
young	77.3 ±2.7	91.5 ±0.9

Table 6: SAM's performance segmenting clothing across perceived gender presentation and age group. The intervals for perceived gender are disjoint, with mIoU for masculine being higher. Confidence intervals for age group overlap.

Fairness in segmenting people. To investigate SAM's fairness at segmenting people we use the More Inclusive Annotations for People (MIAP) [87] test set annotations for Open Images [60], which allows us to compare SAM's performance across perceived gender presentation and perceived age group. MIAP provides box annotations, while we need ground truth masks for this analysis. To get ground truth masks, we select each person-category mask from Open Images if its corresponding bounding box is within a 1% margin (based on relative box side lengths) of an annotated bounding box in MIAP, resulting in 3.9k masks.

Fairness in segmenting clothing. We extend our analysis from §6 to clothing segmentation. We look at SAM's performance on clothing relative to the attributes of those wearing the clothes. We use all 6.5k ground truth masks from Open Images that have a category under the clothing superclass and reside within a person box from MIAP. In Table 6 we compare performance across perceived gender presentation and age group. We find that SAM is better at segmenting clothing on those who present predominantly masculine, with disjoint 95% confidence intervals. The gap closes when moving from 1 to 3 point evaluation. Differences for perceived age group are not significant. Our results indicate there is a bias when segmenting clothing across perceived gender presentation with a one point prompt, and we encourage users of SAM to be mindful of this limitation.

D. Experiment Implementation Details

D.1. Zero-Shot Single Point Valid Mask Evaluation

Datasets. We built a new segmentation benchmark to evaluate the zero-shot transfer capabilities of our model using a suite of 23 diverse segmentation datasets from prior work. A description of each dataset is given in Table 7. For examples, see main text Fig. 8. This suite covers a range of domains including egocentric [34, 28, 113], microscopy [12], X-ray [104], underwater [52, 100], aerial [17], simulation [86], driving [25], and painting [24] images. For efficient evaluation we subsampled datasets with more than 15k masks. Specifically, we randomly picked images so that the total number of masks in the sampled images was ~10k. We blurred faces of people in all the datasets.

Point sampling. Our default point sampling follows standard practice in interactive segmentation [109, 64, 92]. The first point is chosen deterministically as the point farthest from the object boundary. Each subsequent point is the farthest from the boundary of the error region between ground truth and the previous prediction. Some experiments (where specified) use a more challenging sampling strategy in which the first point is a *random* point, rather than a deterministically selected "center" point. Each subsequent point is selected as described above. This setting better reflects use cases in which the first point is not reliably near the center of the mask, such as prompting from eye gaze.

Evaluation. We measure IoU between a prediction after N point prompts and a ground truth mask, where $N = \{1, 2, 3, 5, 9\}$ and points are sampled iteratively with either of the strategies described above. The per-dataset mIoU is the per-mask IoU averaged across all objects in the dataset. Finally, we report the top-line metric by averaging the per-dataset mIoUs across all 23 datasets. Our evaluation differs from the standard interactive segmentation evaluation protocol which measures the average number of points needed to achieve X% IoU, with up to 20 points. We focus on predictions after just one, or possibly a few points, since many of our use cases involve a single or very few prompts. Given our application focus, which requires real-time prompt processing, we expect the best interactive segmentation models to outperform SAM when using a large number of points.

Baselines. We use three recent strong interactive baselines: RITM [92], FocalClick [18], and SimpleClick [67]. For each, we use the largest models trained on the broadest datasets publicly released by the authors. For RITM, we use `HRNet32 IT-M` trained on the combination of COCO [66] and LVIS [44] introduced by the authors. For FocalClick, we use `SegFormerB3-S2` trained on a "combined dataset" that includes 8 different segmentation datasets [18]. For SimpleClick, we use `ViT-H448` trained on a combination of COCO and LVIS. We follow the suggested default strategies for data pre-processing (*i.e.*, data augmentations or image resizing) and do not change or adapt any parameters for our evaluation. In our experiments, we observe that RITM outperforms other baselines on our 23 dataset suite with 1 point evaluation. Therefore, we use RITM as the default baseline. When evaluating with more points we report results for all baselines.

Single point ambiguity and oracle evaluation. In addition to IoU after N points prompts, we report SAM's "oracle" performance at 1 point by evaluating the predicted mask that best matches ground truth from amongst SAM's three predictions (rather than using the one that SAM itself ranks first, as we do by default). This protocol addresses possible single point prompt ambiguity by relaxing the requirement to guess the one right mask among several valid objects.

	米卢在 1 分	3分		米卢在 1 分	3分
感知的性别呈现女性化	76.3±1.1	90.7±0.5	感知年龄组较年长	81.9±3.8	92.8±1.6
男性	81.0 ±1.2	92.3 ±0.4	中年	78.2 ±0.8	91.3 ±0.3
			青年	77.3±2.7	91.5±0.9

表 6：SAM 根据感知的性别表现和年龄组对服装进行细分的表现。感知性别的区间是不相交的，男性的 mIoU 较高。年龄组重叠的置信区间。

人员划分的公平性。为了调查 SAM 在分割人群方面的公平性，我们使用了针对开放图像的更包容性注释 (MIAP) [87] 测试集注释 [60]，这使我们能够比较 SAM 在感知性别表现和感知年龄组中的表现。 MIAP 提供框注释，而我们需要地面实况掩码来进行此分析。为了获得真实掩模，我们从开放图像中选择每个人类别掩模，如果其相应的边界框位于 MIAP 中带注释的边界框的 1% 边距（基于相对框边长）内，从而产生 3.9k 个掩模。

服装细分的公平性。我们将分析从§6 扩展到服装细分。我们根据穿着者的属性来查看 SAM 在服装上的表现。我们使用 Open Images 中的所有 6.5k 地面真实掩模，这些掩模在服装超类下有一个类别，并且驻留在 MIAP 的人员框中。在表 6 中，我们比较了感知性别表现和年龄组的表现。我们发现 SAM 更擅长对那些以男性为主的人进行服装分类，具有不相交的 95% 置信区间。差距缩小

当从 1 点评价转向 3 点评价时。感知年龄组的差异并不显着。我们的结果表明，在通过一点提示对感知的性别表现进行服装分类时存在偏差，我们鼓励 SAM 用户注意这一限制。

D. 实验实施细节

D.1.零样本单点有效掩模评估

数据集。我们建立了一个新的分割基准，使用先前工作中的 23 个不同分割数据集来评估我们模型的零样本传输能力。表 7 给出了每个数据集的描述。例如，请参见正文图 8。该套件涵盖了一系列领域，包括自我中心 [34,28,113]、显微镜 [12]、X 射线 [104]、水下 [52, 100]、空中 [17]、模拟 [86]、驾驶 [25] 和绘画 [24] 图像。为了有效评估，我们对具有超过 15k 个掩模的数据集进行了二次采样。具体来说，我们随机挑选图像，以便采样图像中的掩模总数为

～10k。我们对所有数据集中的人脸进行了模糊处理。

点采样。我们的默认点采样遵循交互式分割的标准实践 [109,64,92]。第一个点被确定性地选择为距对象边界最远的点。每个后续点距离真实值和先前预测之间的误差区域边界最远。一些实验（如果指定）使用更具挑战性的采样策略，其中第一个点是随机点，而不是确定性选择的“中心”点。如上所述选择每个后续点。此设置更好地反映了第一个点不能可靠地靠近掩模中心的用例，例如眼睛注视的提示。

评估。我们测量 N 点提示后的预测与真实掩码之间的 IoU，其中 N = {1, 2, 3, 5, 9}，并且使用上述任一策略对点进行迭代采样。每个数据集的 mIoU 是数据集中所有对象的每个掩模 IoU 的平均值。最后，我们通过对所有 23 个数据集的每个数据集 mIoU 进行平均来报告最重要的指标。我们的评估不同于标准交互式分割评估协议，后者测量所需的平均点数

达到 X % IoU，最高 20 分。我们只专注于一个或可能几个点之后的预测，因为我们的许多用例都涉及单个或很少的提示。鉴于我们的应用重点需要实时提示处理，我们期望在使用大量点时最好的交互式分割模型能够胜过 SAM。

基线。我们使用三个最近的强交互基线：RITM [92]、FocalClick [18] 和 SimpleClick [67]。对于每一个，我们都使用在作者公开发布的最广泛数据集上训练的最大模型。对于 RITM，我们使用作者引入的 COCO [66] 和 LVIS [44] 组合训练的 HRNet32 IT–M。

对于 FocalClick，我们使用 SegFormerB3–S2 在

“组合数据集”，包括 8 种不同的分割

数据集[18]。对于 SimpleClick，我们使用在 COCO 和 LVIS 组合上训练的 ViT–H448。我们遵循建议的默认数据预处理策略（即数据增强或图像调整大小），并且不会更改或调整任何评估参数。在我们的实验中，我们观察到 RITM 优于其他基线

在我们的 23 个数据集套件上进行 1 分评估。因此，我们使用 RITM 作为默认基线。当使用更多点进行评估时，我们会报告所有基线的结果。

单点歧义和预言评估。除了 N 点提示后的 IoU 之外，我们还通过评估 SAM 的三个预测中与真实情况最匹配的预测掩码来报告 SAM 在 1 点时的“oracle”性能（而不是使用 SAM 本身排名第一的那个，就像我们所做的那样）默认）。该协议通过放宽在多个有效对象中猜测一个正确掩码的要求来解决可能的单点提示歧义。

dataset	abbreviation & link	image type	description	mask type	source split	# images sampled	# masks sampled
Plant Phenotyping Datasets Leaf Segmentation [74]	PPDLS	Plants	Leaf segmentation for images of tobacco and ara plants.	Instance	N/A	182	2347
BBBC038v1 from Broad Bioimage Benchmark Collection [12]	BBBC038v1	Microscopy	Biological images of cells in a variety of settings testing robustness in nuclei segmentation.	Instance	Train	227	10506
Dataset fOr bOuldeRs Segmentation [80]	DOORS	Boulders	Segmentation masks of single boulders positioned on the surface of a spherical mesh.	Instance	DS1	10000	10000
TimberSeg 1.0 [38]	TimberSeg	Logs	Segmentation masks of individual logs in piles of timber in various environments and conditions. Images are taken from an operator's point-of-view.	Instance	N/A	220	2487
Northumberland Dolphin Dataset 2020 [100]	NDD20	Underwater	Segmentation masks of two different dolphin species in images taken above and under water.	Instance	N/A	4402	6100
Large Vocabulary Instance Segmentation [44]	LVIS	Scenes	Additional annotations for the COCO [66] dataset to enable the study of long-tailed object detection and segmentation.	Instance	Validation (v0.5)	945	9642
STREETS [91]	STREETS	Traffic camera	Segmentation masks of cars in traffic camera footage.	Instance	N/A	819	9854
ZeroWaste-f [6]	ZeroWaste-f	Recycling	Segmentation masks in cluttered scenes of deformed recycling waste.	Instance	Train	2947	6155
iShape [111]	iShape	Irregular shapes	Segmentation masks of irregular shapes like antennas, logs, fences, and hangers.	Instance	Validation	754	9742
ADE20K [117]	ADE20K	Scenes	Object and part segmentation masks for images from SUN [107] and Places [116] datasets.	Instance	Validation	302	10128
Occluded Video Instance Segmentation [81]	OVIS	Occlusions	Instance segmentation masks in videos, focusing on objects that are occluded.	Instance	Train	2044	10011
Hypersim [86]	Hypersim	Simulation	Photorealistic synthetic dataset of indoor scenes with instance masks.	Instance	Evermotion archinteriors volumes 1-55 excluding 20,25,40,49	338	9445
Night and Day Instance Segmented Park [22, 23]	NDISPark	Parking lots	Images of parking lots from video footage taken at day and night during different weather conditions and camera angles for vehicle segmentation.	Instance	Train	111	2577
EPIC-KITCHENS VISOR [28, 27]	VISOR	Egocentric	Segmentation masks for hands and active objects in ego-centric video from the cooking dataset EPIC-KITCHENS [27].	Instance	Validation	1864	10141
Plittersdorf dataset [46]	Plittersdorf	Stereo images	Segmentation masks of wildlife in images taken with the SOCRATES stereo camera trap.	Instance	Train, validation, test	187	546
Egocentric Hand-Object Segmentation [113]	EgoHOS	Egocentric	Fine-grained egocentric hand-object segmentation dataset. Dataset contains mask annotations for existing datasets.	Instance	Train (including only Ego4D [43] and THU-READ [97, 96])	2940	9961
InstanceBuilding 2D [17]	IBD	Drones	High-resolution drone UAV images annotated with roof instance segmentation masks.	Instance	Train (2D annotations)	467	11953
WoodScape [112]	WoodScape	Fisheye driving	Fisheye driving dataset with segmentation masks. Images are taken from four surround-view cameras.	Instance	Set 1	107	10266
Cityscapes [25]	Cityscapes	Driving	Stereo video of street scenes with segmentation masks.	Panoptic	Validation	293	9973
PIDray [104]	PIDRay	X-ray	Segmentation masks of prohibited items in X-ray images of baggage.	Instance	Test (hard)	3733	8892
Diverse Realism in Art Movements [24]	DRAM	Paintings	Domain adaptation dataset for semantic segmentation of art paintings.	Semantic	Test	718	1179
TrashCan [52]	TrashCan	Underwater	Segmentation masks of trash in images taken by underwater ROVs. Images are sourced from the J-EDI [69] dataset.	Instance	Train (instance task)	5936	9540
Georgia Tech Egocentric Activity Datasets [34, 63]	GTEA	Egocentric	Videos are composed of four different subjects performing seven types of daily activities with segmentation masks of hands.	Instance	Train (segmenting hands task)	652	1208

Table 7: Segmentation datasets used to evaluate zero-shot segmentation with point prompts. The 23 datasets cover a broad range of domains; see column "image type". To make our evaluation efficient, we subsample datasets that have more than 15k masks. Specifically, we randomly sampled images so that the total number of masks in the images is ~10k.

数据集	缩写和链接	图像类型	描述	面罩类型	来源分割	# 采样图像	# 采样的掩模
植物表型数据集叶子分割 [74]	PPDLS	植物	烟草和阿拉植物图像的叶子分割。	实例	不适用	182	2347
来自 Broad Bioimage Benchmark Collection 的 BBBC038v1 [12]	BBBC038v1	显微镜检查	各种环境下的细胞生物图像，测试细胞核分割的稳健性。	实例	火车	227	10506
巨石分割数据集 [80]	门	巨石	位于球形网格表面上的单个巨石的分割掩模。	实例	DS1	10000	10000
TimberSeg 1.0 [38]	木段	日志	在各种环境和条件下对木材堆中的单个原木进行分割掩模。图像是从操作员的角度拍摄的。	实例	不适用	220	2487
诺森伯兰海豚数据集 2020 [100]	NDD20	水下	水上和水下拍摄的图像中两种不同海豚物种的分割掩模。	实例	不适用	4402	6100
大词汇量实例分割 [44]	左室可视度	可视度	COCO [66] 数据集的附加注释可用于长尾对象检测和分割的研究。	实例	验证（v0.5）	945	9642
街道 [91]	街道	交通摄像	交通摄像机镜头中汽车的分割掩模。	实例	不适用	819	9854
零浪费–f [6]	零浪费–f	回收	变形回收废物杂乱场景中的分割掩模。	实例	火车	2947	6155
形状 [111]	iShape	不规则形状	不规则形状的分割蒙版，如天线、原木、栅栏和衣架。	实例	验证	754	9742
ADE20K [117]	ADE20K	场景	来自 SUN [107] 和 Places [116] 数据集的图像的对象和部分分割掩模。	实例	验证	302	10128
遮挡视频实例分割[81]	奥维斯	遮挡	视频中的实例分割掩模，重点关注被遮挡的对象。	实例	火车	2044	10011
超级仿真 [86]	超级仿真	模拟	带有实例蒙版的室内场景的真实感合成数据集。	实例	Evermotion archinteriors 第 1–55 卷，不包括 20、25、40、49	338	9445
昼夜实例分段公园 [22, 23]	NDIS公园	停车场	根据不同天气条件下白天和晚上拍摄的视频片段以及用于车辆分割的摄像机角度拍摄的停车场图像。	实例	火车	111	2577
EPIC–厨房遮阳板 [28, 27]	遮阳板	以自我为中心	来自烹饪数据集 EPIC–KITCHENS [27] 的以自我为中心的视频中的手和活动物体的分割掩模。	实例	验证	1864	10141
Plittersdorf 数据集 [46]	普利特斯多夫	立体图像	使用 SOCRATES 立体相机陷阱拍摄的图像中野生动物的分割掩模。	实例	训练、验证、测试	187	546
以自我为中心的手部物体分割 [113]	自我居屋	以自我为中心	细粒度的以自我为中心的手部物体分割数据集。数据集包含现有数据集的掩模注释。	实例		2940	9961
实例建筑 2D [17]	炎症性肠病	航空	使用屋顶实例分割掩模注释的高分辨率无人机图像。	实例	火车（2D注释）	467	11953
林景 [112]	木景	鱼眼驾驶	带有分割掩模的鱼眼驾驶数据集。图像由四个环视摄像头拍摄。	实例	套装1	107	10266
城市景观 [25]	城市景观	驾驶	带有分割蒙版的街景立体视频。	全景式	验证	293	9973
PIDray [104]	PIDRay	X射线	行李X射线图像中违禁物品的分割掩模。	实例	测试（困难）	3733	8892
艺术运动中的多元化现实主义 [94]	动态随机存取存储器	绘画	用于艺术绘画语义分割的领域适应数据集。	语义学	测试	718	1179
垃圾桶 [52]	垃圾箱	水下	水下 ROV 拍摄的图像中垃圾的分割掩模。图像来源于 J-EDI [69] 数据集。	实例	火车（实例任务）	5936	9540
佐治亚理工学院以自我为中心的活动数据集 [34, 63]	GTEA	以自我为中心	视频由四个不同的主题组成，他们执行七种类型的日常活动，并带有手部分割掩模。	实例	训练（分段手任务）	652	1208

表 7：用于评估带有点提示的零样本分割的分割数据集。这 23 个数据集涵盖了广泛的领域；参见“图像类型”栏。为了提高评估效率，我们对具有超过 15k 个掩模的数据集进行了二次采样。具体来说，我们随机采样图像，使得图像中的掩模总数为 ~10k。

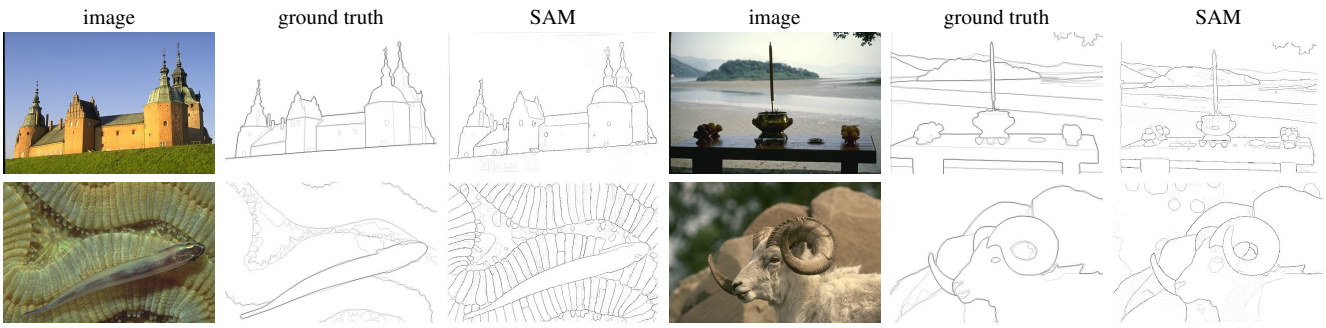


Figure 15: Additional visualizations of zero-shot edge predictions on BSDS500. Recall that SAM was not trained to predict edge maps and did not have access to BSDS images and annotations during training.

D.2. Zero-Shot Edge Detection

Dataset and metrics. We perform zero-shot edge detection experiments on BSDS500 [72, 3]. The ground truth for each image comes from the manual annotations of five different subjects. We report results on the 200 image test subset using the four standard metrics for edge detection [3, 32]: optimal dataset scale (ODS), optimal image scale (OIS), average precision (AP), and recall at 50% precision (R50).

Method. For zero-shot transfer, we use a simplified version of our automatic mask generation pipeline. We prompt SAM with a 16×16 regular grid of foreground points, which yields 768 predicted masks (three per point). We do not filter by predicted IoU or stability. Redundant masks are removed by NMS. Then we apply a Sobel filter to the remaining masks' unthresholded probability maps and set values to zero if they do not intersect with the outer boundary pixels of a mask. Finally, we take a pixel-wise max over all the predictions, linearly normalize the result to [0,1], and apply edge NMS [13] to thin the edges.

Visualizations. In Fig. 15, we show additional examples of zero-shot edge predictions from SAM. These qualitative examples further illustrate how SAM tends to output sensible edge maps, despite not being trained for edge detection. We see that the edges can align well with the human annotations. Although, as previously mentioned, since SAM is not trained for edge detection it does not learn the biases of the BSDS500 dataset and often outputs more edges than are present in the ground truth annotations.

D.3. Zero-Shot Object Proposals

Dataset and metrics. We report the standard average recall (AR) metric for masks at 1000 proposals on the LVIS v1 validation set [44]. Since LVIS has high-quality masks for 1203 object classes, it provides a challenging test for object proposal generation. We focus on AR@1000 due to the open-world nature of our model, which will likely produce many valid masks outside even the 1203 classes in LVIS. To measure performance on frequent, common, and rare categories, we use AR@1000 but measured against a ground truth set containing just the corresponding LVIS categories.

Baseline. We use cascade ViTDet-H as a baseline, the strongest model from [62] by AP on LVIS. As noted in the main text, an object detector trained in-domain can "game" AR [16] and is expected to be a stronger baseline than other models that focus on open-world proposals or segmentation [58, 105]. To produce 1000 proposals, we disable score thresholding in the three cascade stages and as raise the maximum number of predictions per stage to 1000.

Method. We use a modified version of SAM's automatic mask generation pipeline for zero-shot transfer. First, to make inference time comparable to that of ViTDet we do not process image crops. Second, we remove filtering by predicted IoU and stability. This leaves two tunable parameters to get ∼1000 masks per image: the input point grid and the NMS threshold duplicate mask suppression. We choose a 64×64 point grid and an NMS threshold of 0.9, which produces ∼900 masks per image on average. At evaluation, if greater than 1000 masks have been proposed in an image, they are ranked by the average of their confidence and stability scores, then truncated to the top 1000 proposals.

We hypothesize that SAM's ability to output multiple masks is especially valuable for this task, since recall should benefit from proposals generated at multiple scales from a single input point. To test this, we compare to an ablated version SAM that only outputs a single mask instead of three (SAM - single-output). Since this model produces fewer masks, we further increase the number of points sampled and NMS threshold to 128×128 and 0.95, respectively, obtaining ∼950 masks per image on average. Additionally, single-output SAM does not produce the IoU score used to rank masks for NMS in the automatic mask generation pipeline, so instead masks are ranked randomly. Testing suggests this has similar performance to more sophisticated methods of ranking masks, such as using the max logit value of the mask as a proxy for model confidence.

图 15：BSDS500 上零样本边缘预测的附加可视化。回想一下，SAM 没有接受过预测边缘图的训练，并且在训练期间无法访问 BSDS 图像和注释。

D.2。零射击边缘检测

数据集和指标。我们在 BSDS500 上进行了零样本边缘检测实验 [72, 3]。每个图像的基本事实来自五个不同的手动注释

科目。我们使用边缘检测的四个标准指标 [3, 32] 报告 200 个图像测试子集的结果：最佳数据集规模 (ODS)、最佳图像规模 (OIS)、平均精度 (AP) 和 50% 精度的召回率 (R50）。

方法。对于零样本传输，我们使用自动掩模生成管道的简化版本。我们用 16×16 的前景点规则网格提示 SAM，

生成 768 个预测掩模（每点 3 个）。我们不通过预测的 IoU 或稳定性进行过滤。冗余掩码由 NMS 删除。然后，我们对剩余蒙版的未阈值概率图应用索贝尔滤波器，如果它们不与蒙版的外边界像素相交，则将值设置为零。最后，我们对所有预测取像素级最大值，将结果线性归一化为 [0,1]，并应用边缘 NMS [13] 来细化边缘。

可视化。在图 15 中，我们展示了 SAM 零样本边缘预测的其他示例。这些定性示例进一步说明了 SAM 如何倾向于输出合理的边缘图，尽管没有接受过边缘检测训练。我们看到边缘可以与人类注释很好地对齐。尽管如前所述，由于 SAM 未接受边缘检测训练，因此它不会学习 BSDS500 数据集的偏差，并且通常会输出比真实注释中存在的边缘更多的边缘。

D.3。零样本对象提案

数据集和指标。我们报告了 LVIS v1 验证集上 1000 个提案中掩码的标准平均召回率 (AR) 指标 [44]。由于 LVIS 拥有高质量的口罩

1203个对象类别，它为对象提供了具有挑战性的测试

项目提案生成。由于我们模型的开放世界性质，我们专注于 AR@1000，这可能会产生

其存在 LVIS 的 1000 类之外还有许多有效的口罩。衡量频

对于稀有类别，我们使用 AR@1000，但针对仅包含相应 LVIS 类别的地面实况集进行测量。

基线。我们使用级联 ViTDet–H 作为基线，这是 AP 在 LVIS 上的[62]中最强的模型。正如正文中所指出的，在域内训练的对象检测器可以“游戏” AR [16]，并且预计将成为比其他专注于开放世界提案或分割的模型更强的基线 [58, 105]。为了产生 1000 个提案，我们在三个级联阶段禁用分数阈值，并将每个阶段的最大预测数量提高到 1000。

方法。我们使用 SAM 自动掩模生成管道的修改版本来实现零样本传输。首先，为了使推理时间与 ViTDet 相当，我们不处理图像裁剪。其次，我们通过预测的 IoU 和稳定性去除过滤。这留下了两个可调参数来为每个图像获取 ~1000个掩模：输入点网格和NMS阈值重复掩模抑制。我们选择

64×64点网格和NMS阈值0.9，

平均每张图像生成约 900 个掩模。在评估时，

如果在图像中提出了超过 1000 个掩模，则按照置信度的平均值对它们进行排名

我们假设 SAM 输出多个掩码的能力对于这项任务特别有价值，因为召回应该受益于从单个输入点以多个尺度生成的建议。为了测试这一点，我们与仅输出单个掩码而不是三个掩码的消融版本 SAM 进行比较（SAM – 单输出）。由于该模型产生的掩模较少，我们进一步将采样点数和 NMS 阈值分别增加到 128×128 和 0.95，

平均每张图像获得约 950 个掩模。此外，单输出 SAM 不会生成用于在自动掩码生成管道中对 NMS 掩码进行排序的 IoU 分数，因此掩码会随机排序。测试表明，这与更复杂的掩码排名方法具有相似的性能，例如使用掩码的最大 Logit 值作为模型置信度的代理。

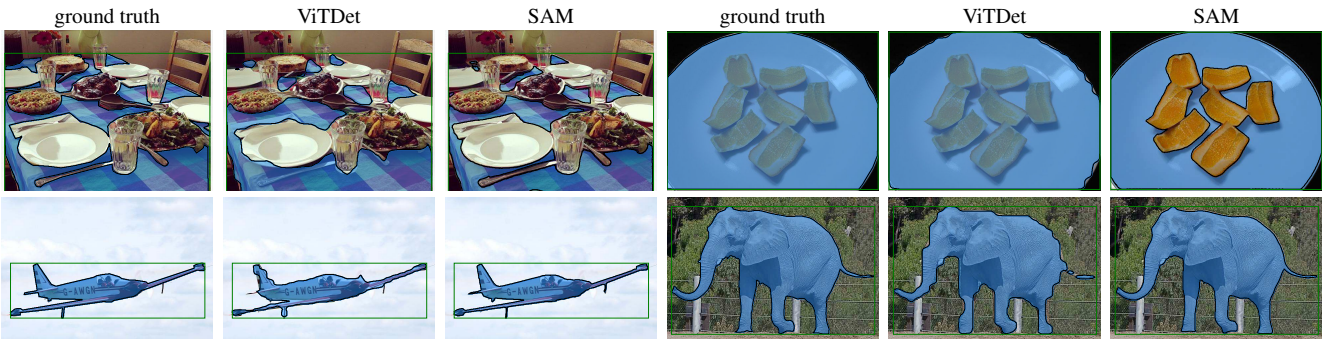


Figure 16: Zero-shot instance segmentation on LVIS v1. SAM produces higher quality masks than ViTDet. As a zero-shot model, SAM does not have the opportunity to learn specific training data biases; see top-right as an example where SAM makes a modal prediction, whereas the ground truth in LVIS is amodal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

D.4. Zero-Shot Instance Segmentation

Method. For zero-shot instance segmentation, we prompt SAM with the boxes output by a fully-supervised ViTDet-H on COCO and LVIS v1 validation splits. We apply an additional mask refinement iteration by feeding the most confident predicted mask, together with the box prompt, back to the mask decoder to produce the final prediction. We show zero-shot instance segmentations predicted on LVIS in Fig. 16. Compared to ViTDet, SAM tends to produce higher quality masks with cleaner boundaries. We confirm this observation with human studies in §7.4. Note that as a zero-shot model, SAM is not able to learn annotation biases in a dataset. For instance, we see that SAM makes a valid modal prediction for the plate, whereas LVIS masks cannot contain holes by design so the plate is annotated amodally.

D.5. Zero-Shot Text-to-Mask

Model and training. We use the largest publicly available CLIP model [82] (`ViT-L/14@336px`) to compute text and image embeddings, which we ℓ^2 normalize prior to use. To train SAM, we use masks from the first two stages of our data engine. Moreover, we discard all masks with an area smaller than 100^2 pixels. We train this model with large-scale jitter [40] for 120k iterations with batch size 128. All other training parameters follow our default settings.

Generating training prompts. To extract an input prompt we first expand the bounding box around each mask by a random factor from $1\times$ to $2\times$, square-crop the expanded box to maintain its aspect ratio, and resize it to 336×336 pixels. Before feeding the crop to the CLIP image encoder, with 50% probability we zero-out pixels outside the mask. To ensure the embedding focuses on the object, we use masked attention in the last layer to restrict attention from the output token to the image positions inside the mask. Finally, our prompt is the output token embedding. For training we supply the CLIP-based prompt first, followed by additional iterative point prompts to refine the prediction.

Figure 17: Visualization of thresholding the similarities of mask embeddings from SAM's latent space. A query is indicated by the magenta box; top row shows matches at a low threshold, bottom row at a high threshold. The most similar mask embeddings in the same image can often be semantically similar to the query mask embedding, even though SAM is not trained with explicit semantic supervision.

Inference. During inference we use the CLIP text encoder without any modifications to create a prompt for SAM. We rely on the fact that text and image embeddings are aligned by CLIP, which allows us to train without any explicit text supervision while using text-based prompts for inference.

D.6. Probing the Latent Space of SAM

Finally, we perform an initial investigation to qualitatively probe the latent space learned by SAM. In particular, we are interested in whether SAM is able to capture any semantics in its representation even though is not trained with explicit semantic supervision. To do so, we compute *mask embeddings* by extracting an image embedding from SAM from an image crop around a mask and its horizontally flipped version, multiplying the image embedding by the binary mask, and averaging over spatial locations. In Fig. 17, we show 3 examples of a query mask and similar masks (in the latent space) in the same image. We observe

基本事实 维生素D 萨姆 基本事实 维生素D 萨姆

图 16：LVIS v1 上的零样本实例分段。 SAM 生产的掩模质量比 ViTDet 更高。作为一个零样本模型，SAM 没有机会学习特定的训练数据偏差；请参见右上角的示例，其中 SAM 进行模态预测，而 LVIS 中的基本事实是非模态的，因为 LVIS 中的掩模注释没有漏洞。

D.4。零样本实例分割

方法。对于零样本实例分割，我们通过完全监督的 ViTDet–H 在 COCO 和 LVIS v1 验证分割上输出的框来提示 SAM。我们通过将最置信的预测掩模与框提示一起反馈给掩模解码器以产生最终预测来应用额外的掩模细化迭代。我们在图 16 中展示了在 LVIS 上预测的零样本实例分割。与 ViTDet 相比，SAM 倾向于生成具有更清晰边界的更高质量的掩模。我们通过第 7.4 节中的人体研究证实了这一观察结果。请注意，作为零样本模型，SAM 无法学习数据集中的注释偏差。例如，我们看到 SAM 对板进行了有效的模态预测，而 LVIS 掩模在设计上不能包含孔，因此板被非模态注释。

D.5。零样本文本到掩模

模型和训练。我们使用最大的公开可用的 CLIP 模型 [82] (ViT–L/14@336px) 来计算文本和图像嵌入，我们在使用之前对其进行标准化。为了训练 SAM，我们使用数据引擎前两个阶段的掩码。此外，我们丢弃所有面积小于 100 像素的掩模。我们使用大规模抖动 [40] 训练该模型，进行 120k 次迭代，批量大小为 128。所有其他训练参数均遵循我们的默认设置。

生成培训提示。为了提取输入提示，我们首先将每个蒙版周围的边界框扩展为 1× 到 2× 的随机因子，对扩展框进行方形裁剪以保持其纵横比，并将其大小调整为 336×336 像素。在将裁剪图像输入 CLIP 图像编码器之前，我们有 50% 的概率将掩模外部的像素清零。为了确保嵌入集中在对象上，我们在最后一层使用掩模注意力来将注意力从输出标记限制到掩模内的图像位置。最后，我们的提示是输出标记嵌入。对于训练，我们首先提供基于 CLIP 的提示，然后提供额外的迭代点提示来完善预测。

图 17：对 SAM 潜在空间中掩模嵌入的相似性进行阈值化的可视化。查询由洋红色框表示；顶行显示低阈值的匹配，底行显示高阈值的匹配。即使 SAM 没有经过显式语义监督的训练，同一图像中最相似的掩码嵌入通常在语义上与查询掩码嵌入相似。

推理。在推理过程中，我们使用 CLIP 文本编码器而不进行任何修改来创建 SAM 提示。我们依赖于通过 CLIP 对齐文本和图像嵌入的事实，这使我们能够在没有任何显式文本监督的情况下进行训练，同时使用基于文本的提示进行推理。

D.6。探索 SAM 的潜在空间

最后，我们进行初步调查，定性地探究 SAM 学到的潜在空间。特别是，我们感兴趣的是 SAM 是否能够捕获其表示中的任何语义，即使没有经过显式语义监督的训练。为此，我们通过从掩模及其水平翻转版本周围的图像裁剪中的 SAM 中提取图像嵌入来计算掩模嵌入，将图像嵌入乘以二进制掩模，并对空间位置求平均值。在图 17 中，我们展示了同一图像中查询掩码和相似掩码（在潜在空间中）的 3 个示例。我们观察到

that the nearest neighbors for each query show some, albeit imperfect, shape and semantic similarity. Although these results are preliminary, they indicate that the representations from SAM may be useful for a variety of purposes, such as further data labeling, understanding the contents of datasets, or as features for downstream tasks.

E. Human Study Experimental Design

Here we describe details of the human study used to evaluate mask quality in §7.1 and §7.4. The purpose of the human study is to address two limitations of using IoU to ground truth as a measure of predicted mask quality. The first limitation is that, for ambiguous inputs such as a single point, the model may be strongly penalized for returning a valid mask of a different object than the ground truth. The second limitation is that ground truth masks may include various biases, such as systematic errors in the edge quality or decisions to modally or amodally segment occluding objects. A model trained in-domain can learn these biases and obtain a higher IoU without necessarily producing better masks. Human review can obtain a measure of mask quality independent of an underlying ground truth mask in order to alleviate these issues.

Models. For single-point evaluation, we use RITM [92], single-output SAM, and SAM to test two hypotheses. First, we hypothesize that SAM produces visually higher quality masks than baseline interactive segmentation models when given a single point, even when metrics such as IoU with ground truth do not reveal this. Second, we hypothesize that SAM's ability to disambiguate masks improves mask quality for single point inputs, since single output SAM may return masks that average over ambiguous masks.

For instance segmentation experiments, we evaluate cascade ViTDet-H [62] and SAM in order to test the hypothesis that SAM produces visually higher quality masks, even if it obtains a lower AP due to the inability to learn specific annotation biases of the validation dataset.

Datasets. For single-point experiments, we select 7 datasets from our set of 23 datasets, since the full suite is too large for human review. We choose LVIS v0.5 [17], VISOR [28, 27], DRAM [24], IBD [17], NDD20 [100], OVIS [81], and iShape [111], which provide a diverse collection of images, including scene-level, ego-centric, drawn, overhead, underwater, and synthetic imagery. Additionally, this set includes datasets both where SAM outperforms RITM with IoU metrics and vice-versa. For instance segmentation experiments, we use the LVIS v1 validation set, allowing for direct comparison to ViTDet, which was trained on LVIS.

Methodology. We presented masks generated by the models to professional annotators and asked them to rate each mask using provided guidelines (see §G for the complete guidelines). Annotators were sourced from the same company that collected manually annotated masks for the data engine. An annotator was provided access to an image, the predicted mask of a single model, and the input to the model (either a single point or single box) and asked to judge the mask on three criterion: Does the mask correspond to a valid object? Does the mask have a clean boundary? and Does the mask correspond to the input? They then submitted a rating from 1-10 indicating the overall mask quality.

A score of 1 indicates a mask that corresponds to no object at all; a low score (2-4) indicates that the mask has huge errors, such including huge regions of other objects or having large areas of nonsensical boundaries; a middle score (5-6) indicates masks that are mostly sensible but still have significant semantic or boundary errors; a high score (7-9) indicates masks with only minor boundary errors; and a score of 10 is for masks with no visible errors. Annotators were provided with five different views, each designed to help identify different error types.

For single point experiments, 1000 masks per dataset were selected randomly from the same subsets used for benchmarking zero-shot interactive segmentation (see §D.1 for details on these subsets). The model input was the centermost point, calculated as the largest value of the distance transform from the edge of the mask. For instance segmentation experiments, 1000 masks were selected from the LVIS v1 validation set, and the model input was the LVIS ground truth box. In all experiments, masks with a size smaller than 24^2 pixels were excluded from sampling, to prevent showing raters a mask that was too small to judge accurately. For both memory and display reasons, large images were rescaled to have a max side-length of 2000 before predicting a mask. In all experiments, the same inputs were fed to each model to produce a predicted mask.

For comparison, the ground truth masks from each dataset were also submitted for rating. For single-point experiments, this gave 4000 total rating jobs per dataset (1000 masks each for RITM, SAM single-output, SAM, and ground truth); for instance segmentation experiments, it gave 3000 total jobs (ViTDet, SAM, and ground truth).

For each dataset, these jobs were inserted with random ordering into a queue from which 30 annotators drew jobs. In initial testing of the review study, we provided each job to five different annotators and found reasonable consistency in scores: the average standard deviation in score over the five annotators was 0.83. Additionally, the annotation company deployed quality assurance testers who spot checked a fraction of results for extreme departures from the guidelines. Thus for our experiments each job (*i.e.*, rating one mask in one image) was completed by only a single annotator. Average time spent per annotator per job was 90 seconds, longer than our initial target of 30 seconds, but still sufficiently fast to collect a large number of ratings on each of the 7 selected datasets.

每个查询的最近邻居显示出一些（尽管不完美）形状和语义相似性。尽管这些结果是初步的，但它们表明 SAM 的表示可能可用于多种目的，例如进一步的数据标记、理解数据集的内容或作为下游任务的特征。

E. 人体研究实验设计

在这里，我们在第 7.1 节和第 7.4 节中描述了用于评估口罩质量的人体研究的详细信息。人类研究的目的是解决使用 IoU 与真实值作为预测掩模质量的衡量标准的两个局限性。第一个限制是，对于诸如单点之类的模糊输入，模型可能会因为返回与地面实况不同的对象的有效掩模而受到严厉惩罚。第二个限制是真实掩模可能包含各种偏差，例如边缘质量的系统误差或模态或非模态分割遮挡对象的决策。域内训练的模型可以学习这些偏差并获得更高的 IoU，而不必产生更好的掩模。人工审查可以获得独立于底层真实掩模的掩模质量测量，以缓解这些问题。

楷模。对于单点评估，我们使用 RITM [92]、单输出 SAM 和 SAM 来检验两个假设。首先，我们假设当给定一个点时，SAM 会比基线交互式分割模型产生视觉上更高质量的掩模，即使 IoU 等指标没有揭示这一点。其次，我们假设 SAM 消除掩码歧义的能力提高了单点输入的掩码质量，因为单输出 SAM 可能会返回比模糊掩码平均的掩码。

例如分割实验，我们评估级联 ViTDet-H [62] 和 SAM，以测试 SAM 产生视觉上更高质量的掩模的假设，即使由于无法学习验证数据集的特定注释偏差而获得较低的 AP 。

数据集。对于单点实验，我们选择7个数据集

来自我们的 23 个数据集，因为全套数据太大

供人工审核。我们选择LVIS v0.5 [17]，VISOR [28，

27]、DRAM [24]、IBD [17]、NDD20 [100]、OVIS [81] 和 iShape [111]，它们提供了各种图像集合，包括场景级、以自我为中心的、绘制的、开销的、水下和合成图像。此外，该集还包括 SAM 在 IoU 指标方面优于 RITM 的数据集，反之亦然。例如分割实验，

我们使用 LVIS v1 验证集，可以与在 LVIS 上训练的 ViTDet 进行直接比较。

方法。我们向专业注释者展示了模型生成的蒙版，并要求他们使用提供的指南对每个蒙版进行评分（完整指南请参阅§G）

注释器来自为数据引擎收集手动注释掩码的同一家公司。向注释者提供对图像、单个模型的预测掩模以及模型的输入（单个点或单个框）的访问权限，并要求根据三个标准判断掩模：掩模是否对应于有效对象？掩模的边界是否清晰？掩码是否与输入相对应？然后他们提交了 1 到 10 的评分，表明面罩的整体质量。

得分 1 表示掩模根本不对应任何对象；低分 (2–4) 表示掩模存在巨大错误，例如包含其他对象的大区域或具有大面积无意义的边界；中间分数 (5–6) 表示掩码大部分是合理的，但仍然存在显着的语义或边界错误；高分 (79) 表示掩模仅具有较小的边界错误；和一个

10 分表示掩模没有明显错误。为注释者提供了五种不同的视图，每种视图都旨在帮助识别不同的错误类型。

对于单点实验，每个数据集的 1000 个掩模是从用于基准零样本交互式分割的相同子集中随机选择的（有关这些子集的详细信息，请参阅§D.1）。模型输入是最中心的点，计算为距掩模边缘的距离变换的最大值。例如分段

心理实验中，选取了1000个口罩

LVIS v1 验证集，模型输入是 LVIS 地面真值框。在所有实验中，掩模的尺寸

小于 24 像素的像素被排除在采样之外，以防止向评估者显示的掩模太小而无法准确判断。出于内存和显示的原因，大图像

在预测面具之前，年龄被重新调整为最大边长 2000。在所有实验中，相同的输入被输入到每个模型以产生预测的掩模。

为了进行比较，还提交了每个数据集的真实掩模进行评级。对于单点实验，每个数据集总共有 4000 个评级作业

（RITM、SAM 单输出、SAM 和地面实况各 1000 个掩码）；例如分割实验，

它总共提供了 3000 个工作岗位（ViTDet、SAM 和 Ground Truth）。

对于每个数据集，这些作业以随机顺序插入到一个队列中，30 个注释者从中提取作业。在审查研究的初始测试中，我们将每项工作提供给五位不同的注释者，并发现分数具有合理的一致性：五位注释者分数的平均标准差为 0.83。此外，注释公司还部署了质量保证测试人员，抽查一小部分结果是否与指南存在严重偏差。因此，对于我们的实验，每项工作（即，对一张图像中的一个掩模进行评级）仅由一个anno-完成

塔托尔。每个注释者每个作业花费的平均时间为 90 秒

等等，比我们最初的 30 秒目标长，但仍然足够快，可以收集每个内容的大量评分

7 个选定的数据集。

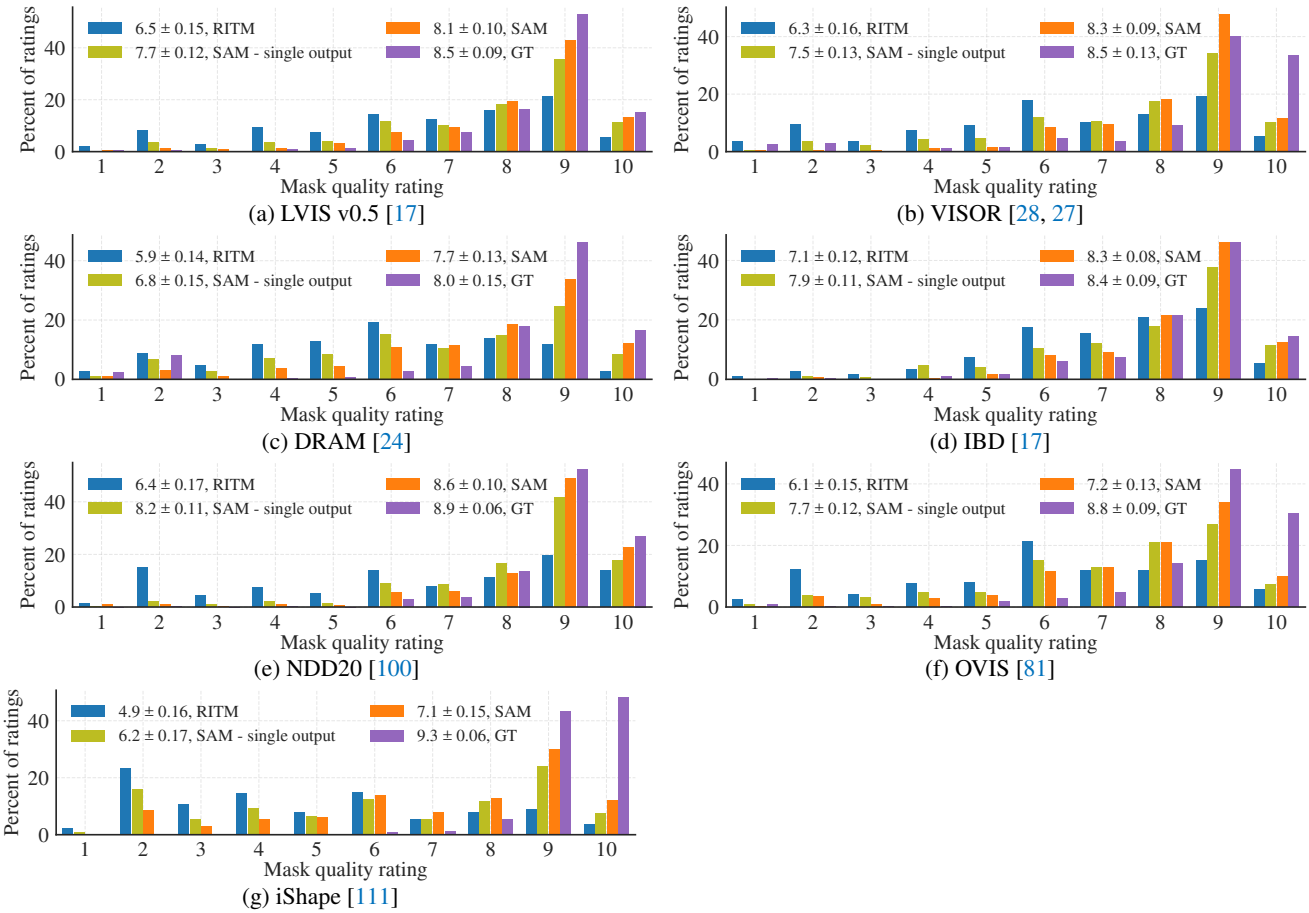


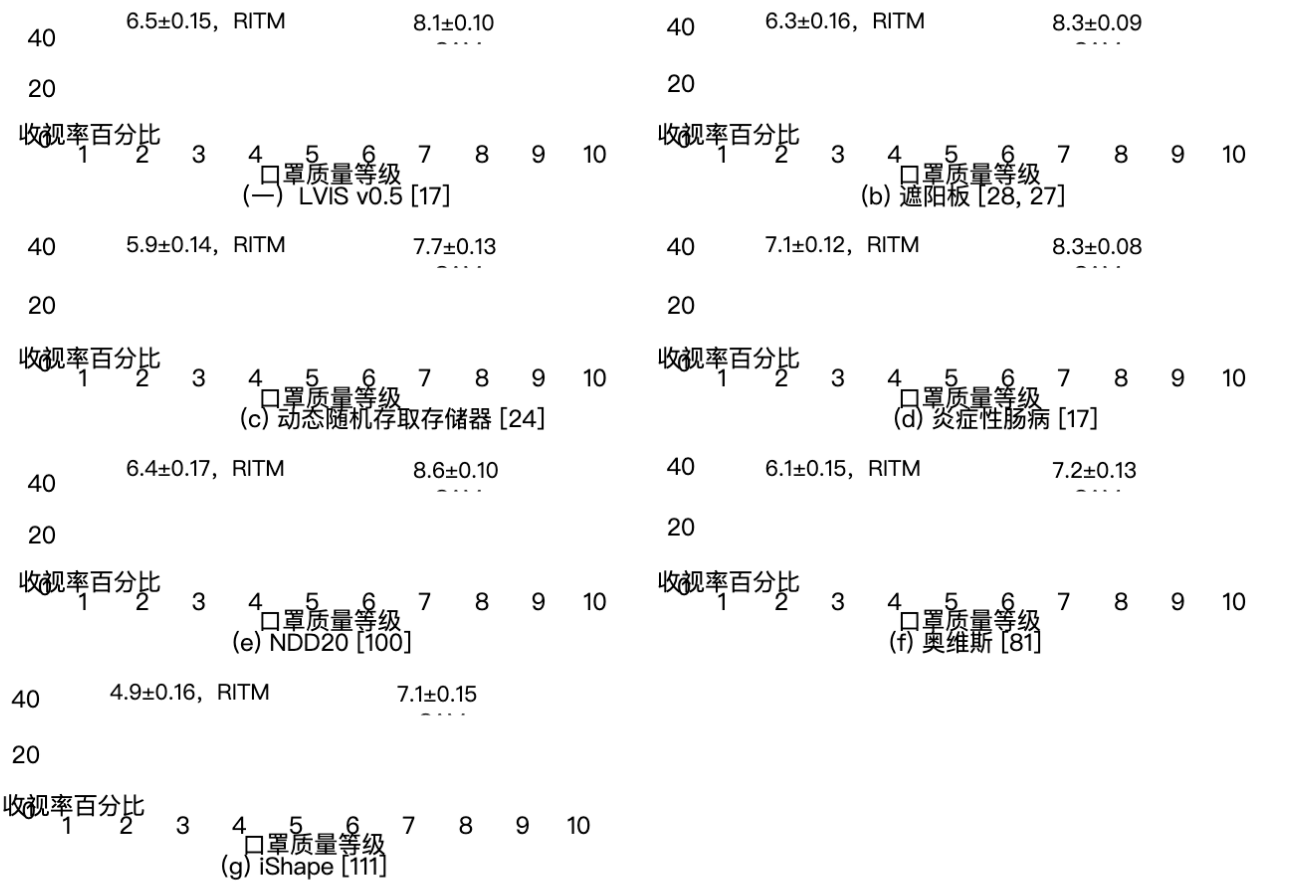
Figure 18: Mask quality rating distributions by dataset from our human evaluation study.

	SAM > baseline		SAM > SAM single out.	
dataset	p-value	$CI_{99}(\Delta\mu)$	p-value	$CI_{99}(\Delta\mu)$
point input (RITM [92] baseline):				
LVIS v0.5 [44]	4e-69	(1.40, 1.84)	2e-11	(0.29, 0.64)
VISOR [28, 27]	7e-98	(1.81, 2.24)	7e-26	(0.58, 0.94)
DRAM [24]	1e-76	(1.54, 2.00)	2e-24	(0.62, 1.03)
IBD [17]	2e-57	(1.03, 1.39)	1e-15	(0.32, 0.62)
NDD20 [100]	2e-86	(1.88, 2.37)	5e-08	(0.19, 0.55)
OVIS [81]	2e-64	(1.38, 1.84)	3e-10	(0.27, 0.63)
iShape [111]	2e-88	(1.97, 2.47)	7e-23	(0.65, 1.10)
box input (ViTDet-H [62] baseline):				
LVIS v1 [44]	2e-05	(0.11, 0.42)	N/A	N/A

Table 8: Statistical tests showing significance that SAM has higher mask quality ratings than baseline and single-output SAM. P-values are calculated by paired t-test, while confidence intervals for the difference in mean scores are calculated by paired bootstrap on 10k samples. All p-values are significant, and all confidence intervals exclude zero.

Results. Fig. 18 shows histograms over ratings for each dataset in the single-point experiments. We run statistical tests for two hypotheses: (1) that SAM gets higher scores than the baseline model (RITM or ViTDet) and (2) that SAM gets higher scores than single-output SAM. P-values are calculated via a paired t-test on the means of the model scores, which we supplement with a paired bootstrap test on 10k samples to find the 99% confidence interval for the difference of means. Table 8 shows p-values and confidence intervals for these tests. All statistical tests are strongly significant, and all confidence intervals exclude zero.

For instance segmentation, Fig. 11 of the main text shows the histogram for ratings. To compare to COCO ground truth, we additionally include 794 ratings of COCO ground truth masks that were collected during our testing of the human review process. These masks were presented to raters using an identical setup as the LVIS results. For fair comparison, results for LVIS in Fig. 11 were subsampled to the same 794 inputs for each model and ground truth. For Table 8, the full 1000 ratings are used to run statistical tests, which show that SAM's mask quality improvement over ViTDet is statistically significant.



(一) LVIS v0.5 [17]
(b) 遮阳板 [28, 27]
(c) 动态随机存取存储器 [24]
(d) 炎症性肠病 [17]
(e) NDD20 [100]
(f) 奥维斯 [81]
(g) iShape [111]

图 18：根据我们的人类评估研究数据集掩盖质量评级分布。

数据集	SAM > 基线 p 值	SAM > 基线 CI99(Δμ)	SAM > SAM 单出。 p 值	SAM > SAM 单出。 CI99(Δμ)
点输入（RITM [92]基线）：LVIS v0.5 [44]	1e−99	(1.40, 1.84)	2e−11	(0.29, 0.64)
遮阳板 [28, 27]	7e−98	(1.81, 2.24)	7e−26	(0.58, 0.94)
动态随机存储器 [24]	1e−76	(1.54, 2.00)	2e−24	(0.62, 1.03)
炎症性肠病 [17]	2e−57	(1.03, 1.39)	1e−15	(0.32, 0.62)
NDD20 [100]	2e−86	(1.88, 2.37)	5e−08	(0.19, 0.55)
奥维斯 [81]	2e−64	(1.38, 1.84)	3e−10	(0.27, 0.63)
形状 [111]	2e−88	(1.97, 2.47)	7e−23	(0.65, 1.10)
框输入（ViTDet-H [62] 基线）：LVIS v1 [44]	2e−05	(0.11, 0.42)	不适用	不适用

表 8：统计测试表明 SAM 具有比基线和单输出 SAM 更高的掩模质量等级。 P 值通过配对 t 检验计算，而平均分数差异的置信区间则通过 10k 样本的配对 bootstrap 计算。所有 p 值均显着，并且所有置信区间均排除零。

结果。图 18 显示了单点实验中每个数据集的评分直方图

我们对两个假设进行统计测试：(1) SAM 比基线模型（RITM 或 ViTDet）获得更高的分数，(2) SAM 比单输出 SAM 获得更高的分数。 P 值是通过对模型分数均值进行配对 t 检验来计算的，我们对 10k 样本进行配对自举检验进行补充，以找到均值差异的 99% 置信区间。表 8 显示了这些检验的 p 值和置信区间。所有统计检验均非常显着，并且所有置信区间均排除零。

例如分段，正文的图 11 显示了评级的直方图。与COCO比较

地面真相，我们还包括在我们测试人类审查过程期间收集的 COCO 地面真相面具的 794 个评级。使用与 LVIS 结果相同的设置将这些掩模呈现给评估者。为了公平起见

比较，图 11 中 LVIS 的结果是二次采样的

每个模型和地面实况都有相同的 794 个输入。

对于表 8，使用全部 1000 个评级来运行统计测试，这表明 SAM 相对于 ViTDet 的掩模质量改进具有统计显着性。

F. Dataset, Annotation, and Model Cards

In §F.1 we provide a Dataset Card for SA-1B, following [39], in a list of questions and answers. Next, we provide a Data Annotation Card in §F.2 for the first two stages of our data engine described in §4, following CrowdWorkSheets [30], again as a list of questions and answers. We provide a Model Card following [75] in Table 9.

F.1. Dataset Card for SA-1B

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.* The contributions of our dataset to the vision community are fourfold: (1) We release a dataset of 11M images and 1.1B masks, by far the largest segmentation dataset to date. (2) The dataset we release is privacy protecting: we have blurred faces and license plates in all images. (3) The dataset is licensed under a broad set of terms of use which can be found at https://ai.facebook.com/datasets/segment-anything. (4) The data is more geographically diverse than its predecessors, and we hope it will bring the community one step closer to creating fairer and more equitable models.
2. *Who created the dataset* (e.g., *which team, research group) and on behalf of which entity* (e.g., *company, institution, organization)?* The dataset was created by the FAIR team of Meta AI. The underlying images were collected and licensed from a third party photo company.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.* Meta AI funded the creation of the dataset.
4. *Any other comments?* No.

Composition

1. *What do the instances that comprise the dataset represent* (e.g., *documents, photos, people, countries)? Are there multiple types of instances* (e.g., *movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.* All of the instances in the dataset are photos. The photos vary in subject matter; common themes of the photo include: locations, objects, scenes. All of the photos are distinct, however there are some sets of photos that were taken of the same subject matter.
2. *How many instances are there in total (of each type, if appropriate)?* There are 11 million images.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set* (e.g., *geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not* (e.g., *to cover a more diverse range of instances, because instances were withheld or unavailable).* The dataset is composed of images licensed from a photo provider. The dataset contains all instances licensed. The images are photos, *i.e.* not artwork, although there are a few exceptions. The dataset includes all generated masks for each image in the dataset. We withheld ∼2k randomly selected images for testing purposes.
4. *What data does each instance consist of? "Raw" data* (e.g., *unprocessed text or images) or features? In either case, please provide a description.* Each instance in the dataset is an image. The images were processed to blur faces and license plates to protect the identities of those in the image.
5. *Is there a label or target associated with each instance? If so, please provide a description.* Each image is annotated with masks. There are no categories or text associated with the masks. The average image has ∼100 masks, and there are ∼1.1B masks in total.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing* (e.g., *because it was unavailable). This does not include intentionally removed information, but might include,* e.g., *redacted text.* Yes. Each image is accompanied by a short caption that describes the content and place of the photo in a free form text. Per our agreement with the photo provider we are not allowed to release these captions. However, we use them in our paper to analyze the geographical distribution of the dataset.
7. *Are relationships between individual instances made explicit* (e.g., *users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.* No, there are no known relationships between instances in the dataset.
8. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Errors:* The masks are generated by a segmentation model, so there may be errors or inconsistencies in the masks. *Redundancies:* While no two images are the same, there are instances of images of the same subject taken close together in time.
9. *Is the dataset self-contained, or does it link to or otherwise rely on external resources* (e.g., *websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset* (i.e., *including the external resources as they existed at the time the dataset was created); c) are there any restrictions* (e.g., *licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.* The dataset is self-contained.
10. *Does the dataset contain data that might be considered confidential* (e.g., *data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.* No.
11. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.* We have two safety measures to prevent objectionable content: (1) Photos are licensed from a photo provider and had to meet the terms of service of the photo provider. We requested that all objectionable content be filtered from the images we licensed. (2) If a user observes objectionable image(s) in the dataset, we invite them to report the image(s) at segment-anything@meta.com for removal. Despite the measures taken, we observe that a small portion of images contains scenes of protests or other gatherings that focus on a diverse spectrum of religious beliefs or political opinions that may be offensive. We were not able to produce a filtering strategy that removes all such images and rely on users to report this type of content.
12. *Does the dataset identify any subpopulations* (e.g., *by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.* The dataset does not identify any subpopulations of the people in the photos.
13. *Is it possible to identify individuals* (i.e., *one or more natural persons), either directly or indirectly* (i.e., *in combination with other data) from the dataset? If so, please describe how.* No. Images were subjected to a face blurring model to remove any personally identifiable information. If a user observes any anonymization issue, we invite them to report the issue and the image id(s) at segment-anything@meta.com.
14. *Does the dataset contain data that might be considered sensitive in any way* (e.g., *data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.* The dataset contains scenes of protests, or other gatherings that may suggest religious beliefs, political opinions or union memberships. However, the faces of all people in the dataset have been anonymized via facial blurring, so it is not possible to identify any person in the dataset.
15. *Any other comments?* No.

Collection Process

1. *How was the data associated with each instance acquired? Was the data directly observable* (e.g., *raw text, movie ratings), reported by subjects* (e.g., *survey responses), or indirectly inferred/derived from other data* (e.g., *part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.* The released masks associated with each image were automatically inferred by our segmentation model, SAM. The masks that were collected using model-assisted manual annotation will not be released. Quality was validated as described in §5.
2. *What mechanisms or procedures were used to collect the data* (e.g., *hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?* The images in the dataset are licensed from an image provider. They are all photos taken by photographers with different cameras.

F. 数据集、注释和模型卡

在§F.1中，我们在问题和答案列表中提供了SA–1B的数据集卡，遵循[39]。接下来，我们亲

在 §F.2 中提供了数据注释卡，用于 §4 中描述的数据引擎的前两个阶段，遵循 CrowdWorkSheets [30]，再次作为问题和答案列表。我们按照表 9 中的 [75] 提供了一个模型卡

F.1。 SA–1B 数据集卡

动机 1. 创建数据集的目的是什么？是否有特定的任务？是否有需要填补的具体空白？请提供说明。我们的数据集对视觉社区的贡献有四个方面： （1）我们发布了包含 1100 万张图像和 1.1B 个掩模的数据集，这是迄今为止最大的分割数据集。 （2）我们发布的数据集是隐私保护的：我们在所有图像中都模糊了人脸和车牌。 (3) 该数据集根据广泛的使用条款获得许可，可在 https://ai.facebook.com/datasets/segment–anything 上找到。 (4) 数据比之前的数据在地域上更加多样化，我

哪个实体（例如公司、机构、组织）？该数据集由 Meta AI 的 FAIR 团队创建。底层图像是从第三方照片公司收集

3. 谁资助了数据集的创建？如果有相关补助金，请提供资助者姓名以及资助名称和编号。 Meta AI 资助了该数据集的创建。

4.还有其他意见吗？不。

作品

照片、人物、国家）？是否存在多种类型的实例（例如，电影、用户和评分；人和他们之间的交互；节点和边）？请提供说明。数据集中的所有实例都是照片。照片的主题各不相同；照片的常见主题包括：地点、物体、场景。所

2. 总共有多少个实例（每种类型，如果适用）？有 1100 万张图像。

3. 数据集是否包含所有可能的实例，还是来自较大集合的实例样本（不一定是随机的）？如果数据集是一个样本，那么更大的集合是多少？样本是否代表更大的集合（例如地理覆盖范围）？如果是，请描述如何验证/核实这种代表性。如果它不能代表更大的集合，请描述为什么不代表（例如，为了覆盖更多样化的实例，因为实例被保留或不可用）。该数据集由照片提供商许可的图像组成。该数据集包含所有许可的实例。这些图像是照片，即不是艺术

文本或图像）或功能？无论哪种情况，请提供说明。数据集中的每个实例都是一个图像。这些图像经过处理以模糊

5. 是否有与每个实例关联的标签或目标？如果是，请提供描述。每个图像都用蒙版进行注释。没有与蒙版关联的类别或文本。平均图像有~100个掩模，总共有~1.1B个掩模。

6. 个别实例是否缺少任何信息？如果是这样，请提供描述，解释为什么缺少此信息（例如，因为它不可用）。这不包括故意删除的信息，但可能包括例如经过编辑的文本。是的。每张图像都附有一个简短的标题，以自由格式的文本描述照片的内容和地点。根据我们与照片提供商的

7. 各个实例之间的关系是否明确（例如，用户的电影收视率、社交网络链接）？如果是这样，请描述如何使这些关系变得明确。不，数据集中的实例之间没有已知

8. 数据集中是否存在任何错误、噪声源或冗余？如果是，请提供描述。错误：掩模是由分割模型生成的，因此掩模中可能存在错误或不一致。冗余：虽然没有两张图像是相同的，但存在同一主题的图像在时间上接近的情况。

资源（例如网站、推文、其他数据集）？如果它链接到或依赖于外部资源，a) 是否能保证它们将存在并随着时间的推移保持不变； b) 是否有完整数据集的官方存档版本（即，包括创建数据集时存在的外部资源）； c) 是否存在与可能适用于数据集使用者的任何外部资源相关的任何限制（例如许可证、费用）？请酌情提供所有外部资源以及与之相关的任何限制的描述，以及链接或其他访问点

10. 数据集是否包含可能被视为机密的数据（例如，受法律特权或医患保密保护的数据（包括个人非公开通信内容的数据）？如果是，请提供描述。不。

11. 数据集是否包含如果直接查看可能具有冒犯性、侮辱性、威胁性或可能引起焦虑的数据？如果是这样，请描述原因。我们有两项安全措施来防止令人反感的内容：(1) 照片已获得照片提供商的许可，并且必须符合照片提供商的服务条款。我们要求从我们许可的图像中过滤掉所有令人反感的内容。 (2) 如果用户在数据集中观察到令人反感的图像，我们邀请他们通过segmentanything@meta.com 报告该图像以进行删除。尽管采取了这些措施，我们还是观察到一小部分图像包含抗议或其他集会的场景，这些场景集中于可能令人反感的各种宗教信仰或政治观点。我们无法制定删除所有

请描述如何识别这些亚群，并提供其在数据集中各自分布的描述。该数据集无法识别照片中人物的任何亚群。

13. 是否有可能识别个人（即一个或多个自然人），即直接或间接（即与其他数据结合）来自数据集？如果是这样，请描述如何。不会。图像经过面部模糊模型处理，以删除任何个人身份信息。如果用户发现任何匿名化问题，我们邀请他们通过segment–anything@meta.com 报告

14. 数据集是否包含可能以任何方式被视为敏感的数据（例如，揭示种族或族裔出身、性取向、宗教信仰、政治观点或工会成员身份或地点的数据；财务或健康数据；生物识别或遗传数据）；政府身份证明的形式，例如社会安全号码；犯罪记录）？如果是，请提供描述。该数据集包含抗议或其他可能暗示宗教信仰、政治观点或工会会员身份的集会场景。然而，数据集中所有人的面孔都已通过面部模糊进行匿名化，因此不可能识别数据集中的任何人。

15.还有其他意见吗？不。

收款流程

1. 每个实例的相关数据是如何获取的？数据是直接可观察的（例如，原始文本、电影评级）、由受试者报告（例如，调查回复），还是从其他数据间接推断/导出（例如，词性标签、基于模型的年龄或语言猜测）？如果数据是由受试者报告的或从其他数据间接推断/得出的，则数据是否经过验

软件设备或传感器、人工管理、软件程序、软件 API）？这些机制或程序是如何验证的？数据集中的图像已获得图像提供商的许可。它们都是摄影师用不同相机拍摄的照

3. *If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*, deterministic, probabilistic with specific sampling probabilities)?* We withheld ∼2k randomly selected images for testing purposes. The rest of the licensed images are included in the dataset.
4. *Who was involved in the data collection process (*e.g.*, students, crowdworkers, contractors) and how were they compensated (*e.g.*, how much were crowdworkers paid)?* The released masks were automatically inferred by SAM. For details on our model-assisted manual annotation process see our Data Annotation Card in §F.2. Note these masks will not be released.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (*e.g.*, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.* The licensed photos vary in their date taken over a wide range of years up to 2022.
6. *Were any ethical review processes conducted (*e.g.*, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. If the dataset does not relate to people, you may skip the remaining questions in this section.* We underwent an internal privacy review to evaluate and determine how to mitigate any potential risks with respect to the privacy of people in the photos. Blurring faces and license plates protects the privacy of the people in the photos.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (*e.g.*, websites)?* We licensed the data from a third party photo provider.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.* The images are licensed from a third party who provided appropriate representations regarding the collection of any notices and consents as required from individuals. In addition, all identifiable information (*e.g.* faces, license plates) was blurred. Under the terms of the dataset license it is prohibited to attempt to identify or associate an image with a particular individual.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.* The images are licensed from a third party who provided appropriate representations regarding the collection of any notices and consents as required from individuals. In addition, all identifiable information (*e.g.* faces, license plates) was blurred from all images. For avoidance of doubt, under the terms of the dataset license it is prohibited to attempt to identify or associate an image with a particular individual.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).* We invite users to report at segment-anything@meta.com for image(s) removal.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (*e.g.*, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.* To eliminate any potential impact on people whose photos are included in the dataset, identifiable information (faces, license plates) has been blurred.
12. *Any other comments?* No.

Preprocessing / Cleaning / Labeling

1. *Was any preprocessing / cleaning / labeling of the data done (*e.g.*, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.* We resized the high-resolution licensed images such that the shorter side is 1500 pixels and only processed the images to remove any identifiable and personal information from the photos (faces, license plates).
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (*e.g.*, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.* No, as we removed the data for safety reasons and to respect privacy, we do not release the unaltered photos.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.* We used the RetinaFace [88, 89] model (https://github.com/serengil/retinaface) to detect faces. The model used to blur license plates has not been made public.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.* The dataset was used to train our segmentation model, SAM.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.* No. However, all users of the dataset must cite it, so its use is trackable via citation explorers.
3. *What (other) tasks could the dataset be used for?* We intend the dataset to be a large-scale segmentation dataset. However, we invite the research community to gather additional annotations for the dataset.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (*e.g.*, stereotyping, quality of service issues) or other risks or harms (*e.g.*, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?* We have an analysis of the approximate geographic and income level coverage of our dataset in §6. While we believe our dataset to be more representative than most of the publicly existing datasets at this time, we acknowledge that we do not have parity across all groups, and we encourage users to be mindful of potential biases their models have learned using this dataset.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.* Full terms of use for the dataset including prohibited use cases can be found at https://ai.facebook.com/datasets/segment-anything.
6. *Any other comments?* No.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (*e.g.*, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.* The dataset will be available for the research community.
2. *How will the dataset will be distributed (*e.g.*, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?* The dataset is available at https://ai.facebook.com/datasets/segment-anything.
3. *When will the dataset be distributed?* The dataset will be released in 2023.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.* Yes. The license agreement and terms of use for the dataset can be found at https://ai.facebook.com/datasets/segment-anything. Users must agree to the terms of use before downloading or using the dataset.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.* Full terms of use and restrictions on use of the SA-1B dataset can be found at https://ai.facebook.com/datasets/segment-anything.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.* The license and restrictions on use of the SA-1B dataset can be found at https://ai.facebook.com/datasets/segment-anything.
7. *Any other comments?* No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?* The dataset will be hosted at https://ai.facebook.com/datasets/segment-anything and maintained by Meta AI.
2. *How can the owner/curator/manager of the dataset be contacted (*e.g.*, email address)?* Please email segment-anything@meta.com.
3. *Is there an erratum? If so, please provide a link or other access point.* No.
4. *Will the dataset be updated (*e.g.*, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (*e.g.*, mailing list,*

3. 如果数据集是来自较大集合的样本，那么采样策略是什么（例如，确定性、具有特定采样概率的概率性）？我们保留了 ~2k 随机选择的图像用于测试目的。其余许可图像包含在数据集中。

员工、承包商）以及他们如何获得报酬（例如，众包工作者的报酬是多少）？发布的掩码由 SAM 自动推断。有关模型辅助手动注释过程的详细信息，请参阅§F.2 中的数

5. 数据是在什么时间范围内收集的？此时间范围是否与实例关联的数据的创建时间范围相匹配（例如，最近对旧新闻文章的爬网）？如果没有，请描述创建与实例关联的数据的时间范围。获得许可的照片的拍摄日期各不相同，截至 2022 年

查看板）？如果是，请提供这些审核流程的描述，包括结果，以及任何支持文档的链接或其他访问点。如果数据集与人无关，您可以跳过本节中的其余问题。我们进行了内部隐私审查，以评估并确定如何减轻照片中人物隐私的任何潜在风险。模糊脸部和车牌可以保护照片中人物的隐

7. 您是直接从相关个人那里收集数据，还是通过第三方或其他来源（例如网站）获取数据？我们从第三方照片提供商处获得了数据许可。

请描述（或用屏幕截图或其他信息显示）如何提供通知，并提供链接或其他访问点，或以其他方式复制通知本身的确切语言。这些图像已获得第三方许可，该第三方根据个人要求提供了有关收集任何通知和同意的适当陈述。此外，所有可识别信息（例如面孔、车牌）都变得模糊。根据数据集许可条款，禁止尝试识别图像或将图像与特定个

9. 相关个人是否同意收集和使用其个人信息？

数据？如果是这样，请描述（或用屏幕截图或其他信息显示）如何请求和提供同意，并提供链接或其他访问点，或以其他方式复制个人同意的确切语言。这些图像已获得第三方许可，该第三方根据个人要求提供了有关收集任何通知和同意的适当陈述。此外，所有图像中的所有可识别信息（例如面部、车牌）都被模糊化。为避免疑义，根据数据集许可条款，禁止尝试识别图像或将图像与特定个人关

10. 如果获得同意，是否向同意的个人提供了

将来或出于某些用途撤销其同意的机制？如果是，请提供描述以及该机制的链接或其他访问点（如果适用）。我们邀请用户通过segmentanything@meta.com 举报图像被

11. 是否对数据集及其使用对数据主体的潜在影响进行了分析（例如数据保护影响分析）？如果是，请提供此分析的描述，包括结果，以及任何支持文档的链接或其他访问点。为了消除对数据集中包含照片的人的任何潜在影响，可识别信息（面部、车牌）已被模糊化。

12.还有其他意见吗？不。

预处理/清洁/标签

1. 是否对数据进行了任何预处理/清理/标记（例如，离散化或分桶、标记化、词性标记、SIFT 特征提取、实例删除、缺失值处理）？如果是，请提供描述。如果没有，您可以跳过本节中的其余问题。我们调整了高分辨率许可图像的大小，使短边为 1500 像素，并且仅处理图像以删除照片中的任何可识别信息和个人信息（面部、车牌）。

2. 除了预处理/清理/标记的数据之外，是否还保存了“原始”数据（例如，以支持未来意外的用途）？如果是这样

我们使用 RetinaFace [88, 89] 模型 (https://github.com/serengil/retinaface) 来检测人脸。用遍于模糊车牌的模型尚未公开。

1. 该数据集是否已用于任何任务？如果是，请提供描述。该数据集用于训练我们的分割模型 SAM。

2. 是否有一个存储库链接到使用该数据集的任何或所有论文或系统？如果是，请提供链接或其他访问点。不可以。

成为一个大规模的分割数据集。但是，我们邀请研究界收集数据集的其他注释。

4. 是否有关于数据集的组成或方式的信息

收集和预处理/清洁/标记可能会影响未来的使用？例如，数据集消费者可能需要了解哪些信息，以避免使用可能导致对个人或群体的不公平待遇（例如，刻板印象、服务质量问题）或其他风险或损害（例如，法律风险、财务损害）？如果是，请提供描述。数据集消费者可以采取什么措施来减轻这些风险或危害？我们在第 6 节中对数据集的大致地理和收入水平覆盖范围进行了分析。虽然我们相信我们的数据集比目前大多数公开存在的数据集更具

5. 是否存在不应该使用该数据集的任务？如果是，请提供描述。数据集的完整使用条款，包括禁止使用

6. 还有其他意见吗？不。

分配

数据集是代表哪个公司、机构、组织）创建的？如果是，请提供描述。该数据集将可供研究界使用。

2. 数据集将如何分发（例如网站上的 tarball、API、GitHub）？数据集是否有数字对象标识符 (DOI)？这

3. 数据集什么时候分发？该数据集将于 2023 年发布。

财产 (IP) 许可和/或适用的使用条款 (ToU)？如果是，请描述本许可和/或 ToU，并提供任何相关许可条款或 ToU 以及与这些限制相关的任何费用的链接或其他访问点，或以其他方式复制。是的。

数据集的许可协议和使用条款可以在

5. 是否有任何第三方对数据施加基于IP或其他的限制

与实例相关？如果是，请描述这些限制，并提供任何相关许可条款以及与这些限制相关的任何费用的链接或其他访问点，或以其他方式复制。 SA-1B 数据集的完整使用条

6. 数据集或单个实例是否适用任何出口管制或其他监管限制？如果是这样，请描述这些限制，并提供任何支持文档的链接或其他访问点，或以其他方式复制。 SA-1B数据集的许

7. 还有其他意见吗？不。

维护

托管于 https://ai.facebook.com/datasets/segment-anything 并由 Meta AI 维护。

2. 如何联系数据集的所有者/策展人/管理者（例如电子邮件地址）？请发送电子邮件至segment-anything@meta.com。

立场，删除实例）？如果是这样，请描述更新的频率、由谁以及如何向数据集使用者传达更新（例如，邮件列表、

GitHub)? To aid reproducibility of research using SA-1B, the only updates will be to remove reported images.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (*e.g.*, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.* There are no limits on data retention. We took measures to remove personally identifiable information from any images of people. Users may report content for potential removal here: segment-anything@meta.com.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.* No, as the only updates will be to remove potentially harmful content, we will not keep older versions with the content.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.* We encourage users to gather further annotations for SA-1B. Any users who generate annotations will be liable for hosting and distributing their annotations.
8. *Any other comments?* No.

F.2. Data Annotation Card

Task Formulation

1. *At a high level, what are the subjective aspects of your task?* Segmenting objects present in an image is inherently a subjective task. For instance, one annotator may segment two boots as one mask, whereas another may segment each boot separately. Depending on annotators's skills, the quality of the mask and the number of masks per image are different between annotators. Despite these subjective aspects of the task, we believed efficient annotation was possible as the data was annotated in a per-mask fashion with the main focus on the diversity of the data rather than completeness.
2. *What assumptions do you make about annotators?* Our annotators worked full time on our annotation task with very small attrition rate. This made it possible to train the annotators providing feedback and answering their questions on a regular basis. Specifically: (1) By giving a clear understanding of the goals of this work and providing clear guidelines, including visuals and video recordings of the tasks, annotators had enough context to understand and perform the tasks reasonably. (2) Sharing objectives and key results and meeting weekly with annotators increased the likelihood that annotators improved annotation quality and quantity over time.
3. *How did you choose the specific wording of your task instructions? What steps, if any, were taken to verify the clarity of task instructions and wording for annotators?* As our task was annotating images, the annotation guidelines included visual examples. Our research team completed 30 annotation tasks to identify any obvious challenges using the annotation tool, collectively decide how to handle complex cases, and refine the guidelines. The research team met with the annotators weekly for feedback sessions. Videos of the research team performing the task were shared live with the annotators, followed by Q&A sessions. Annotators were able to give feedback on unclear aspects, both during the feedback session and asynchronously.
4. *What, if any, risks did your task pose for annotators and were they informed of the risks prior to engagement with the task?* No identified risks. Images were filtered for objectionable content prior to the annotation phase.
5. *What are the precise instructions that were provided to annotators?* We provide only high-level instructions: Given an image, we aim at segmenting every possible object. Annotators generate a mask for every potential object they can identify. An object can be segmented using our interactive segmentation tool either by using corrective foreground/background clicks to add/remove parts of the mask or by drawing a bounding box around the object. Masks can be refined using pixel-precise tools.

Selecting Annotations

1. *Are there certain perspectives that should be privileged? If so, how did you seek these perspectives out?* We chose to work with annotators that have worked on other vision annotation tasks before.
2. *Are there certain perspectives that would be harmful to include? If so, how did you screen these perspectives out?* No.
3. *Were sociodemographic characteristics used to select annotators for your task? If so, please detail the process.* No.
4. *If you have any aggregated socio-demographic statistics about your annotator pool, please describe. Do you have reason to believe that sociodemographic characteristics of annotators may have impacted how they annotated the data? Why or why not?* We worked with 130 annotators. The annotators were all based in Kenya. We do not believe sociodemographic characteristics of annotators meaningfully impacted the annotated data.
5. *Consider the intended context of use of the dataset and the individuals and communities that may be impacted by a model trained on this dataset. Are these communities represented in your annotator pool?* The Segment Anything 1B (SA-1B) dataset is to be used for research purposes only. The SA-1B dataset is one of the most geographically diverse segmentation dataset, as discussed in §6. In addition, we analyze the responsible AI axes of a model trained on the dataset in §6.

Platform and Infrastructure Choices

1. *What annotation platform did you utilize? At a high level, what considerations informed your decision to choose this platform? Did the chosen platform sufficiently meet the requirements you outlined for annotator pools? Are any aspects not covered?* We used a proprietary annotation platform.
2. *What, if any, communication channels did your chosen platform offer to facilitate communication with annotators? How did this channel of communication influence the annotation process and/or resulting annotations?* We manually reviewed annotations and shared feedback with the annotators on a weekly basis. We communicated common mistakes or inconsistencies and the corresponding corrections. In addition, the annotators were given feedback for improvements daily by the annotation QA team. Outside the weekly feedback sessions, annotators had access to a spreadsheet and chat group to facilitate communication with the research team. This process greatly improved the average speed and quality of the annotations.
3. *How much were annotators compensated? Did you consider any particular pay standards, when determining their compensation? If so, please describe.* Annotators were compensated with an hourly wage set by the vendor. The vendor is a Certified B Corporation.

Dataset Analysis and Evaluation

1. *How do you define the quality of annotations in your context, and how did you assess the quality in the dataset you constructed?* Annotators were first placed into training. They followed a 1-day training session led by the vendor and then were asked to annotate a large number of examples from a training queue. Annotators graduated from training to production after the vendor QA team, in collaboration with the research team, manually spot-checked the annotator's masks to ensure quality. On average, annotators spent one week in training before graduating. Production quality assessment followed a similar process: the vendor QA team and the research team manually reviewed the annotations weekly, sharing feedback weekly.
2. *Have you conducted any analysis on disagreement patterns? If so, what analyses did you use and what were the major findings? Did you analyze potential sources of disagreement?* We pointed out common mistakes during weekly meetings with the annotators.
3. *How do the individual annotator responses relate to the final labels released in the dataset?* The annotations were only used to train early versions of the SAM model and we do not currently plan to release them.

Dataset Release and Maintenance

1. *Do you have reason to believe the annotations in this dataset may change over time? Do you plan to update your dataset?* No, except to remove objectionable images.
2. *Are there any conditions or definitions that, if changed, could impact the utility of your dataset?* We do not believe so.
3. *Will you attempt to track, impose limitations on, or otherwise influence how your dataset is used? If so, how?* The SA-1B dataset will be released under a license agreement allowing use for certain research purposes and protections for researchers. Researchers must agree to the terms of the license agreement to access the dataset.
4. *Were annotators informed about how the data is externalized? If changes to the dataset are made, will they be informed?* No, we do not plan to release the manual annotations at the moment.
5. *Is there a process by which annotators can later choose to withdraw their data from the dataset? If so, please detail.* No.

GitHub）？为了帮助使用 SA–1B 进行研究的可重复性，唯一的更新是删除报告的图像。

5. 如果数据集与人相关，与实例相关的数据的保留是否有适用的限制（例如，相关个人是否被告知他们的数据将保留一段固定的时间，然后被删除）？如果是这样，请描述这些限制并解释如何执行它们。数据保留没有限制。我们采取措施从任何人物图像中删除个人身份信息。用户可以在此处报告可能删除的内容：segment–

移植/托管/维护？如果是这样，请描述如何。如果没有，请描述如何将其过时信息传达给数据集使用者。不会，因为唯一的更新是删除潜在有害的内容，我们不会保留旧版

7. 如果其他人想要扩展/增强/构建/贡献数据集，有吗他们这样做的机制是什么？如果是，请提供描述。这些贡献会被验证/验证吗？如果是这样，请描述如何。如果没有，为什么不呢？是否有一个流程可以将这些贡献传达/分发给数据集使用者？如果是，请提供描述。我们鼓励用

8. 还有其他意见吗？不。

F.2。数据标注卡

任务制定

图像中存在的物体本质上是一项主观任务。例如，一个注释者可以将两个靴子分割为一个掩码，而另一个注释器可以单独分割每个靴子。根据注释者的技能，注释者之间的掩模质量和每个图像的掩模数量有所不同。尽管任务存在这些主观方面，但我们相信有效的注释是可能的，因为数

2.您对注释者有何假设？我们的注释员全职从事我们的注释任务，流失率非常低。这使得培训注释者定期提供反馈和回答问题成为可能。具体来说：（1）通过清楚地理解这项工作的目标并提供明确的指导方针，包括任务的视觉效果和视频记录，注释者有足够的背景来合理地理解和执行任务。(2) 分享目标和关键结果以及每周与注释者举行会议增加了注释者随着时间的推移提高注释质量和数量的可

是否采取了步骤（如果有）来验证任务说明和注释者措辞的清晰度？由于我们的任务是注释图像，注释指南包括视觉示例。我们的研究团队完成了 30 项注释任务，以使用注释工具识别任何明显的挑战，共同决定如何处理复杂的案例，并完善指南。研究团队每周与注释者会面，进行反馈会议。研究团队执行任务的视频与注释者实时分享，随后进行了问答环节。注释者能够在反馈会话期间和异步方

4. 您的任务给注释者带来了哪些风险（如果有）？他们在参与该任务之前是否被告知了这些风险？没有发现的风险。在注释阶段之前，图像会被过滤掉令人反感的内容。

仅提供高级指令：给定图像，我们的目标是分割每个可能的对象。注释者为他们可以识别的每个潜在对象生成一个掩码。可以使用我们的交互式分割工具来分割对象，方法是使用校正前景/背景单击来添加/删除蒙版的一部分，或者在对象周围绘制边界框。可以使用像素精度手动细化

选择注释

寻求这些观点？我们选择与之前从事过其他视觉注释任务的注释者合作。

2. 是否包含某些有害的观点？如果是这样，您是如何筛选出这些观点的？不。

3. 是否使用社会人口特征来为您的任务选择注释者？如果是的话，请详细说明过程。不。

塔托池，请描述。您是否有理由相信注释者的社会人口特征可能会影响他们注释数据的方式？为什么或者为什么不？我们与 130 名注释者合作。注释者均位于肯尼亚。我们认为注释者的社会人口特征不会对注释数据产生有意

5. 考虑数据集的预期使用环境以及可能受此数据集训练的模型影响的个人和社区。这些社区是否出现在您的注释者池中？ Segment Anything 1B (SA–1B) 数据集仅用于研究目的。 SA–1B 数据集是地理上最多样化的分割数据集之一，如第 6 节中所述。此外，我们分析了在第 6 节中的数据集上训练的模型的负责 AI 轴。

平台和基础设施选择

您决定选择这个平台的原因是什么？所选平台是否足以满足您为注释器池列出的要求？是否还有哪些方面没有涵

2. 您选择的平台提供哪些沟通渠道（如果有）促进与注释者的沟通？这种沟通渠道如何影响注释过程和/或生成的注释？我们每周手动审核注释并与注释者共享反馈。我们传达了常见错误或不一致之处以及相应的更正。此外，注释 QA 团队每天都会向注释者提供改进反馈。在每周的反馈会议之外，注释者可以访问电子表格和聊天组，以促进与研究团队的沟通。这个过程极大地提高了注释的平均速度和质量

3. 注释者的报酬是多少？你有没有考虑过任何参与–确定其报酬时的工资标准？如果有，请描述。注释者按供应商规定的小时工资获得报酬。该供应商是经过 B 公司

数据集分析与评估

您评估您构建的数据集的质量吗？注释者首先接受培训。他们参加了由供应商主持的为期 1 天的培训课程，然后被要求对培训队列中的大量示例进行注释。在供应商 QA 团队与研究团队合作，手动抽查注释者的蒙版以确保质量后，注释者从培训过渡到生产。平均而言，注释者在毕业前要接受一周的培训。生产质量评估遵循类似的流程：供应商 QA 团队和研究团队每周手动审核注释，每周共享反

2. 您是否对分歧模式进行过分析？如果是这样，您使用了哪些分析以及主要发现是什么？您是否分析了分歧的潜在来源？我们在与注释者的每周会议中指出了常见错误。

3. 各个注释者的响应与数据集中发布的最终标签有何关系？这些注释仅用于训练 SAM 模型的早期版本，我们目前不打算发布它们。

数据集发布与维护

1. 您是否有理由相信该数据集中的注释可能会随着时间的推移而改变？您打算更新您的数据集吗？不，除了删除令人反感的图像。

2. 是否存在任何条件或定义如果发生更改可能会影响数据集的实用性？我们不这么认为。

3. 您是否会尝试跟踪、限制或以其他方式影响数据集的使用方式？如果是这样，怎么办？ SA–1B 数据集将根据许可协议发布，允许用于某些研究目的并保护研究人员。研究人员必须同意许可协议的条款才能访问数据集。

Model Overview

Name	SAM or Segment Anything Model
Version	1.0
Date	2023
Organization	The FAIR team of Meta AI
Mode type	Promptable segmentation model
Architecture	See §3
Repository	https://github.com/facebookresearch/segment-anything
Citation	https://research.facebook.com/publications/segment-anything
License	Apache 2.0

Intended Use

Primary intended uses	SAM is intended to be used for any prompt-based segmentation task. We explored its use in *segmenting objects from a point* (§7.1), *edge detection* (§7.2), *segmenting all objects* (§7.3), and *segmenting detected objects* (§7.4). We explored how SAM can integrate with other vision models to *segment objects from text* (§7.5).
Primary intended users	SAM was primarily developed for research. The license for SAM can be found at https://github.com/facebookresearch/segment-anything.
Out-of-scope use cases	See terms of use for SAM found at https://github.com/facebookresearch/segment-anything. See *Use Cases* under *Ethical Considerations*.
Caveats and recommendations	SAM has impressive zero-shot performance across a wide range of tasks. We note, however, that in the zero-shot setting there may be multiple valid ground truth masks for a given input. We recommend users take this into consideration when using SAM for zero-shot segmentation. SAM can miss fine structures and can hallucinate small disconnected components. See §8 for a discussion of limitations.

Relevant Factors

Groups	SAM was designed to segment any object. This includes *stuff* and *things*.
Instrumentation and environment	We benchmarked SAM on a diverse set of datasets and found that SAM can handle a variety of visual data including *simulations, paintings, underwater images, microscopy images, driving data, stereo images, fish-eye images*. See §D.1 and Table 7 for information on the benchmarks used.

Metrics

Model performance measures	We evaluated SAM on a variety of metrics based on the downstream task in our experiments. • *mIoU*: We used the mean intersection-over-union after a given number of prompts to evaluate the segmentation quality of a mask when prompted with points. • *Human evaluation*: We performed a human study (detailed in §E) to evaluate the real world performance of SAM. We compared the masks generated by SAM to a baseline state-of-the-art interactive segmentation model, RITM [92], using a perceptual quality scale from 1 to 10. • *AP*: We used average precision to evaluate instance segmentation for a given box and edge detection. • *AR@1000*: We used average recall to evaluate object proposal generation. • *ODS, OIS, AP, R50*: We used the standard edge detection evaluation metrics from BSDS500 [72, 3].

Evaluation Data

Data sources	See §D.1.

Training Data

Data source	See Data Card in §F.1.

Ethical Considerations

Data	We trained SAM on licensed images. The images were filtered for objectionable content by the provider, but we acknowledge the possibility of false negatives. We performed a geographic analysis of the SA-1B dataset in §6. While SA-1B is more geographically diverse than many of its predecessors, we acknowledge that some geographic regions and economic groups are underrepresented.
Cost and impact of compute	SAM was trained on 256 A100 GPUS for 68 hours. We acknowledge the environmental impact and cost of training large scale models. The environmental impact of training the released SAM model is approximately 6963 kWh resulting in an estimated 2.8 metric tons of carbon dioxide given the specific data center used, using the calculation described in [77] and the ML CO_2 Impact calculator [61]. This is equivalent to $\sim$7k miles driven by the average gasoline-powered passenger vehicle in the US [101]. We released the SAM models to both reduce the need for retraining and lower the barrier to entry for large scale vision research.
Risks and harms	We evaluated SAM for fairness in §6. Downstream use cases of SAM will create their own potential for biases and fairness concerns. As such we recommend users run their own fairness evaluation when using SAM for their specific use case.
Use cases	We implore users to use their best judgement for downstream use of the model.

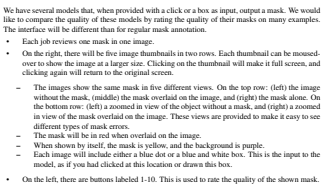
Table 9: Model Card for SAM, following the procedure detailed in [75].

型号概览	
姓名	SAM 或分段任意模型
版本	1.0
日期	2023
组织	Meta AI FAIR团队
模式类型	及时的细分模型
建筑学	参见§3
存储库	
引文	
执照	阿帕奇2.0
有可能的使用	
主要预期用途	SAM 旨在用于任何基于提示的分割任务。我们探索了它在从点分割对象（第 7.1 节）、边缘检测（第 7.2 节）、分割所有对象（第 7.3 节）和分割检测到的对象（第
主要目标用户	SAM 主要是为了研究而开发的。 SAM 的许可证可以在以下位置找到：
超出范围的用例	请参阅 https://github.com/facebookresearch/segment-anything 上的 SAM 使用条款，请参阅道德考虑下的用例。
注意事项和建议	SAM 在各种任务中都具有令人印象深刻的零样本性能。然而，我们注意到，在零样本设置中，给定输入可能有多个有效的地面实况掩码。我们建议用户在使用 SAM 进行零样本分割时考虑这一点。 SAM 可能会错过精细结构，并可能产生幻觉小的不连贯
相关因素	
团体	SAM 旨在分割任何对象。这包括东西和东西。
仪器仪表及环境	我们在一组不同的数据集上对 SAM 进行了基准测试，发现 SAM 可以处理各种视觉数据，包括模拟、绘画、水下图像、显微镜图像、驾驶数据、立体图像、鱼眼图像。有
指标	
模型绩效衡量	我们根据实验中的下游任务，根据各种指标评估了 SAM。 • mIoU：我们在给定数量的提示后使用平均交集来评估点提示时掩模的分割质量。 萨姆。我们使用从 1 到 10 的感知质量等级，将 SAM 生成的掩模与最先进的基线交互式分割模型 RITM [92] 进行比较。 • ODS、OIS、AP、R50：我们使用BSDS500 [72, 3] 中的标准边缘检测评估指标。
评估数据	
数据源	参见§D.1。
训练数据	
数据源	请参阅§F.1 中的数据卡。
道德考虑	
数据	我们使用许可图像对 SAM 进行了训练。图像由提供商过滤掉令人反感的内容，但我们承认存在误报的可能性。我们在第 6 节中对 SA-1B 数据集进行了地理分析。虽然 SA-1B 比其许多前任在地理上更加多样化，但我们承认某些地理区域和经济群体的代
计算的成本和影响	
SAM 在 256 个 A100 GPU 上进行了 68 小时的训练。我们承认培训的环境影响和成本	

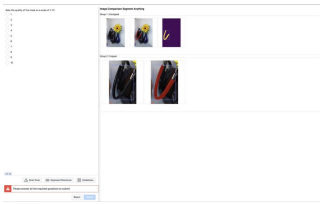
风险和危害 我们在第 6 节中评估了 SAM 的公平性。 SAM 的下游用例将产生潜在的偏见和公平问题。因此，我们建议用户在将 SAM 用于其特定用例时运行自己的公平性评估。

用例我们恳请用户对模型的下游使用做出最佳判断。

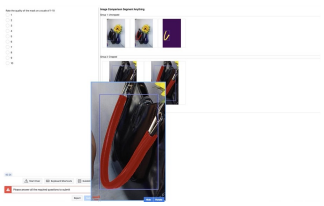
表 9：SAM 的模型卡，遵循[75]中详细说明的过程。



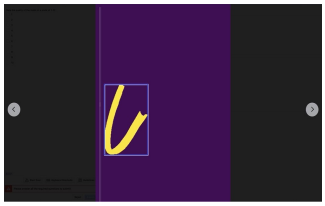
Objective and Setup



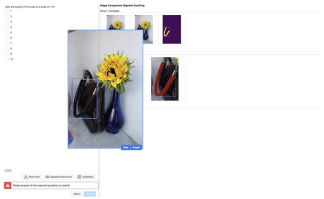
Example interface page. There will be five images on the right and a question box on the left.



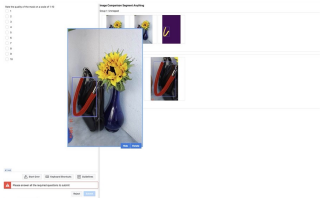
Mouse over an image to show the full image.



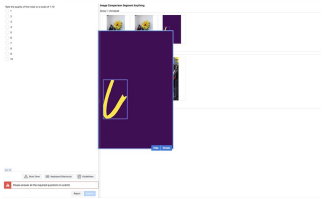
Click on an image to make it full screen. The arrows will cycle between images. Click again to return to previous view.



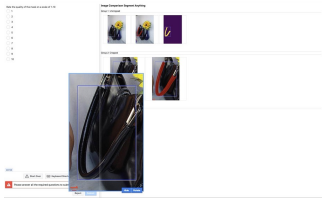
The first image on the top row shows the image without a mask. A blue point will be on the object of interest, or a blue and white box will surround it.



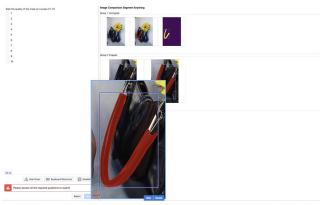
The second image on the top row shows the mask for the object in red.



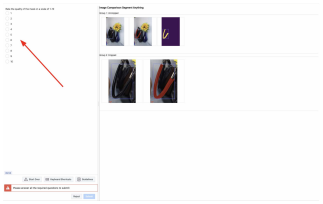
The third image on the top row shows the mask only. The mask is in yellow and the background is purple.



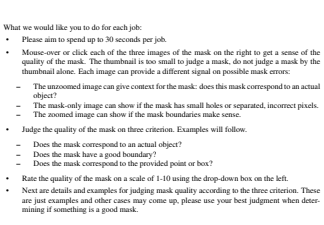
The first image on the bottom row shows a zoomed in view of the object without a mask.



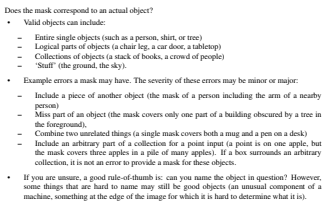
The second image on the bottom row shows a zoomed in view of the object with a mask. The mask is in red.



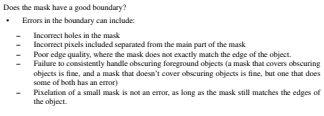
On the left are buttons to rate the mask quality, with selections 1-10.



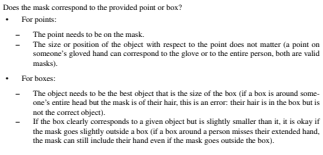
Task



Judging Mask Quality (1 of 3)



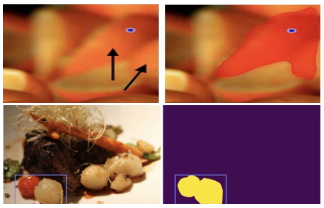
Judging Mask Quality (2 of 3)



Judging Mask Quality (3 of 3)

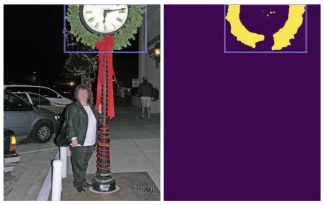
Example error of 'Include a piece of another object': The elephant mask contains a piece of another nearby elephant.

Example error of 'Missing a part of an object': the mask is missing a disconnected part of the object: the back half of the zebra, and the right portion of the plate.



Example error of 'Include an arbitrary part of a collection': In top top image, the point is on one orange rind, but the mask covers two orange rinds. This is a mask error: the mask covers an arbitrary number of objects in the collection, and should either cover one orange rind or all of them. In the bottom image, the box is around both vegetables. Since this is the best match to the box, this is not a mask error.

Example error for 'Incorrect holes in the mask': This mask has holes in the upper left and on the left sides (black arrows). These holes are much easier to see on the 'mask only' image.



Example error for 'Incorrect pixels included separated from the main part of the mask': The 'mask only' view reveals a few stray incorrect pixels on the clock face.

Example error for 'Poor edge quality': The mask has poor edge quality, both along the edge of the umbrella, as well as along the thin pole.

Figure 19: Here we provide the complete guidelines given to annotations for the human review of mask quality. Some images been edited slightly and faces have been blurred to enable release. Best viewed with zoom (part 1 of 2).

G. Annotation Guidelines

We provide the complete guidelines given to annotations for the human review of mask quality in Fig. 19 and Fig. 20.

我们有几个模型，当提供点击或框作为输入时，输出一个掩码。我们希望通过在许多示例中对这些模型的掩模质量进行评级来比较这些模型的质量。该接口将与常规掩码注释的接口不同。

目标和设置

示例界面页面。右侧有五张图片，左侧有一个问题框。

将鼠标悬停在图像上可显示完整图像。

单击图像可使其全屏显示。箭头将在图像之间循环。再次单击可返回到上一视图。

顶行的第一张图像显示了没有蒙版的图像。感兴趣的对象上将出现一个蓝点，或者它周围会出现一个蓝白框。

顶行的第二个图像显示了红色对象的蒙版。

顶行的第三张图像仅显示掩模。面具为黄色，背景为紫色。

底行的第一张图像显示了没有蒙版的对象的放大视图。

我们希望您为每项工作做什么：
- 每个作业的时间请控制在 30 秒以内。
- 将鼠标悬停或单击右侧蒙版的三个图像中的每一个，即可了解蒙版的质量。缩略图太小，无法判断蒙版，请勿仅通过缩略图来判断蒙版。每个图像可以针对可能的掩模错误提供不同的信号：

任务

判断面罩质量（第 1 项，共 3 项）

面具与实际物体相对应吗？
- 有效对象可以包括：
  - 整个单个对象（例如人、衬衫或树）
  - 物体的逻辑部分（椅腿、车门、桌面）
  - 物品集合（一摞书、一群人）
  - “东西”（地面、天空）。
- 包括另一个物体的一部分（一个人的面具，包括附近人的手臂）
  - 丢失对象的一部分（遮罩仅覆盖被前景中的一棵树遮挡的建筑物的一部分）
  - 将两个不相关的东西组合起来（一个面具覆盖了桌子上的杯子和钢笔）
- 包含点输入集合的任意部分（一个点位于一个苹果上，但蒙版覆盖了一堆苹果中的……）
- 如果……些难以命名的东西可能仍然是好对象（机器的不寻常组件，位于图像边缘的东西，很难确定它是什么）。

底行的第二张图像显示了带有蒙版的对象的放大视图。面具是红色的。

左侧是用于评价蒙版质量的按钮，可选择 1–10。

判断面罩质量（3 中的 2）

掩模有良好的边界吗？
- 边界错误可能包括：
  - 面罩上的孔不正确
  - 包含与蒙版主要部分分离的不正确像素
  - 边缘质量差，遮罩与对象的边缘不完全匹配。
  - 无法一致地处理模糊的前景……种操作的蒙版会出错）

判断面罩质量（3 / 3）

掩模是否对应于提供的点或框？
- 对于积分：
  - 点必须在面具上。
  - 对象相对于点的大小或位置并不重要（某人戴手套的手上的点可以对应于手套或整个人，两者都是有效的掩模）。
- 对象必须是盒子大小的最佳对象（如果盒子围绕某人……也没关系（如果一个人周围的盒子错过了他们伸出的手，面具仍然可以包括他们的手，即使面具走出盒子）。

“包含另一个物体的一块”错误示例：大象面具包含附近另一头大象的一块。

“缺少对象的一部分”的示例错误：面罩缺少对象的断开部分：斑马的后半部分和板的右侧部分。

“包含集合的任意部分”的示例错误：在顶部图像中，该点位于一个橙皮上，但蒙版覆盖了两个橙皮。这是一个掩码错误……橙皮或全部橙皮。在下图中，盒子围绕着两种蔬菜。由于这是与框的最佳匹配，因此这不是掩码错误。

“面罩上的孔不正确”的错误示例：该面罩的左上角和左侧有孔（黑色箭头）。这些孔在“仅掩模”图像上更容易看到。

“包含与蒙版主要部分分离的不正确像素”的示例错误：“仅蒙版”视图显示钟面上有一些杂散的不正确像素。

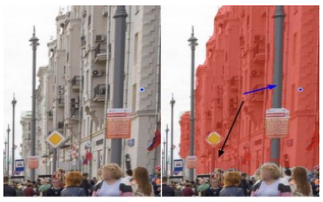
“边缘质量差”的错误示例：面罩的边缘质量很差，无论是沿着伞的边缘还是沿着细杆。

图 19：在这里，我们提供了用于人工审查掩模质量的注释的完整指南。一些图像经过轻微编辑，脸部已模糊以方便发布。缩放效果最佳（第 1 部分，共 2 部分）。

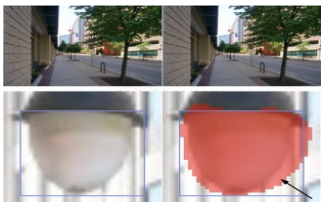
G. 注释指南

我们在图 19 和图 20 中提供了用于人工审查掩模质量注释的完整指南。

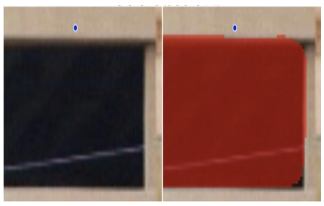
Example for ‘Combine two unrelated things’: The point indicates the lizard, but the mask covers both the lizard and a bird. This is a mask error.



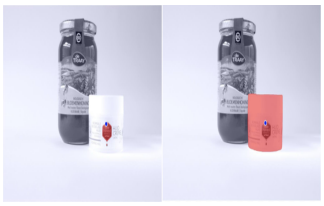
Example error for ‘Failure to consistently handle obscuring foreground objects’: The pole on the right (blue arrow) is excluded from the mask, while the pole on the left is included in the object (black arrow). The mask should either include or exclude both of these.



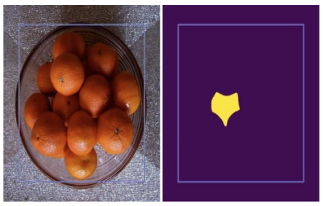
Example of ‘Pixelation of a small mask’: this mask has an imperfect boundary, since it extends beyond the object at the black arrow. However, the ‘blocky’ pattern of the mask is not an error, since, when zoomed in this much, the image is also blocky the same way.



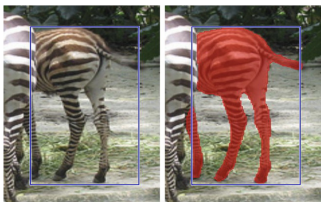
Example error for consistency with the provided point: The mask does not agree with the blue point, so this is a mask error.



Example for consistency with the provided point: For this input point, but the logo (left) and the container (right) are valid objects, since the blue point lies on both of them. Neither mask has a mask error.



Example for consistency with a box: The box surrounds the bowl of oranges, but the mask is only of a single orange. This is a mask error.



Example for consistency with a box: The box’s shape fits the zebra. Even though the mask extends slightly outside the box to include the zebra’s left leg, this is not an error.

Overall mask quality is subjective, each of the above errors may hurt mask quality only a little or a lot, depending on how large the error is. Please use your best judgment when choosing mask scores, and try to stay consistent from mask-to-mask. Here are some general guidelines for what different scores should correspond to:

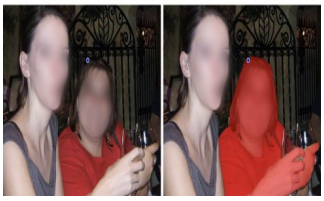
- A score of 1: It is not possible to tell what object this mask corresponds to. This includes the case that there is no mask visible at all.
- A low score (2-4): The object is mostly identifiable, but the mask quality is extremely poor (*e.g.* large regions of the mask cover other objects; large regions of the object missing; extremely splotchy mask boundaries that cut through the middle of the object).
- A mid score (5-6): The object is identifiable and the boundary is mostly correct, but there are major errors (missing a significant disconnected part of the object; containing a significant part of another object; very poor boundary quality in one area of the object but not the entire object).
- A high score (7-9): The object is identifiable and errors are small and rare (missing a small, heavily obscured disconnected component, having small regions where the mask boundary does not quite match the object boundary).
- A score of 10: The mask is pixel-perfect; it has no identifiable errors at all.

Mask Scoring

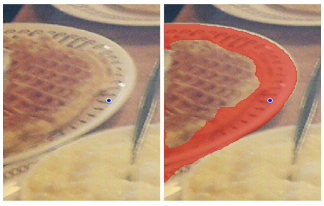
Example of a mask with a score of 1: It is not clear what object this mask corresponds to.

Example of a mask with a low score (2-4): The main object is identifiable, but the mask includes a large, incorrect portion of another object.

Example of a mask with a low score (2-4): The main object is identifiable, but a large, random part of the object is missing.



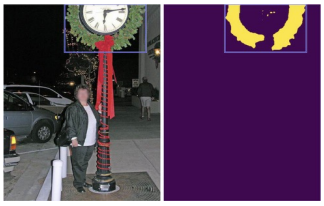
Example of a mask with a low-to-medium score (4-5): The object is identifiable and the edges are all correct, but the mask incorrectly includes the hand of the person on the left.



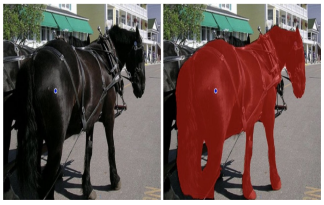
Example of a mask with a medium score (5-6): The mask clearly corresponds to the plate, but the boundary with the waffle is quite poor.

Example of a mask with a medium score (5-6): the object is easy to identify, and most of the edges make sense. However, there is a significant disconnected part (their arm inside the frame) that is mostly missing, as well as splotchy pixels in this region.

Example of a mask with a medium-to-high score (6-8): the mask has two small-ish regions of poor boundary, at the top of the mask and on the bottom right.

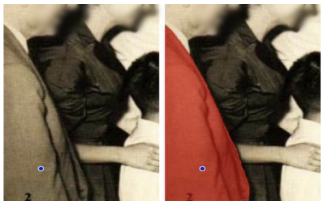


Example of a mask with a medium-to-high score (6-8): The wreath is a valid object that is the size of the box (the entire wreath + clock would also be a valid object). However, there are incorrect stray mask pixels on the clock.



Example of a mask with a high score (7-9): The boundary of the horse is almost entirely correct, except for the right side of its back leg. The mask consistently includes all of the equipment that horse is wearing, and has logical boundaries.

Example of a mask with a very high score (∼9): There are only minor errors around the edge of the mask. The blocky ‘pixelation’ is not an error, since the image is also blocky at this scale.



Example of a mask with a very high score (9-10): the mask has only very minor errors in the edge on the bottom right.

Example of a mask with a very high score (9-10): There are only minor errors around the edge of the mask.

Figure 20: Here we provide the complete guidelines given to annotations for the human review of mask quality. Some images been edited slightly and faces have been blurred to enable release. Best viewed with zoom (part 2 of 2).

“组合两个不相关的事物”的示例：点表示蜥蜴，但面具同时覆盖了蜥蜴和鸟。这是掩码错误。

“未能一致地处理模糊前景对象”的错误示例：右侧的极点（蓝色箭头）被排除在蒙版之外，而左侧的极点包含在对象中（黑色箭头）。掩码应包含或排除这两者。

“小蒙版的像素化”示例：该蒙版的边界不完美，因为它延伸到黑色箭头处的对象之外。然而，掩模的“块状”图案并不是错误，因为当放大这么多时，图像也会以同样的方式呈现块状。

与提供的点一致性的示例错误：掩模与蓝点不一致，因此这是掩模错误。

总体掩模质量是主观的，上述每个错误都可能对掩模质量产生轻微或很大的影响，具体取决于误差有多大。选择掩模分数时请使用您的最佳判断，并尽量在掩模之间保持一致。以下是不同分数应对应的一些一般准则：

掩模评分

与提供的点保持一致的示例：对于此输入点，但徽标（左）和容器（右）是有效对象，因为蓝点位于它们两者上。两个掩模都没有掩模错误。

与盒子保持一致的示例：盒子包围着一碗橙子，但面具只有一个橙子。这是掩码错误。

与盒子保持一致的示例：盒子的形状适合斑马。尽管遮罩稍微延伸到框外以包含斑马的左腿，但这并不是错误。

得分为 1 的掩码示例：不清楚该掩码对应于什么对象。

低分 (2–4) 的蒙版示例：主要对象是可识别的，但蒙版包含另一个对象的较大且不正确的部分。

低分 (2–4) 的掩码示例：主要对象是可识别的，但对象的很大一部分随机部分丢失。

中低分 (4–5) 的蒙版示例：对象可识别且边缘全部正确，但蒙版错误地包含了左侧人的手。

中等分数 (5–6) 的掩模示例：掩模明显对应于盘子，但与华夫饼的边界相当差。

中等分数 (5–6) 的掩模示例：对象易于识别，并且大部分边缘有意义。然而，有一个显着的断开部分（他们的手臂在框架内）大部分缺失，并且该区域中存在斑点像素。

具有中到高分数 (6–8) 的蒙版示例：蒙版有两个边界较差的小区域，位于蒙版顶部和右下角。

中高分 (6–8) 的掩码示例：花环是一个有效对象，其大小与盒子一样（整个花环 + 时钟也将是一个有效对象）。然而，时钟上存在不正确的杂散掩模像素。

高分 (7–9) 的面具示例：马的边界几乎完全正确，除了后腿的右侧。面具始终包含马所佩戴的[illegible]，并且[illegible]边界。

具有非常高分数 (~9) 的掩模示例：掩模边缘周围仅存在微小错误。块状“像素化”并不是[illegible]，[illegible]图像也是块状的。

具有非常高分数 (9–10) 的蒙版示例：该蒙版在右下角边缘仅存在非常小的错误。

具有非常高分数 (9–10) 的掩模示例：掩模边缘周围仅存在微小错误。

图 20：在这里，我们提供了用于人工审查掩模质量的注释的完整指南。一些图像经过轻微编辑，脸部已模糊以方便发布。使用缩放效果最佳（第 2 部分，共 2 部分）。