🎭

Segment Anything

Abstract & Introduction

- 웹 데이터로 pre-trained된 LLMs는 zero-shot/few-shot 일반화 능력을 보임. 이러한 foundation model은 prompt engineering을 통해 훈련에 없었던 task나 데이터 분포에도 일반화가 가능해짐 → 이를 vision 분야로. 기존의 CLIP, ALIGN은 웹에서 수집된 텍스트와 이미지 쌍을 사용해 정렬된 encoder를 학습시키는 방식으로 프롬포트를 잘 설계하여 일반화가 가능해짐 + downstream 작업에도 but cv 분야에서는 텍스트-이미지 쌍으로는 해결할 수 없는 다양한 과제가 있고 대규모 훈련 데이터가 부족한 상황
    - 결국 목표는, image segmenation을 위한 foundation model을 만드는 것!

  → 프롬포트가 가능한 모델, 그리고 강력한 일반화를 가능하게 하는 작업을 통해 광범위한 데이터셋에서의 pre-train을 시도함 →이때 **3가지 요소**가 중요

1) 어떤 작업이 zero-shot 일반화를 가능하게 하는지

- promptable segmentation task 을 제안 → 어떤 형태의 prompt (point, box, mask, text 등 무엇을 분할할지를 알려주는 모든 정보를 포함함)을 입력 받아 valid한 segmentation mask(프롬포트가 모호하더라도, 그 중 하나에 대해 타당한 마스크를 생성해야 한다는 것)를 출력하는 것이 목표(pretraining objective)임.

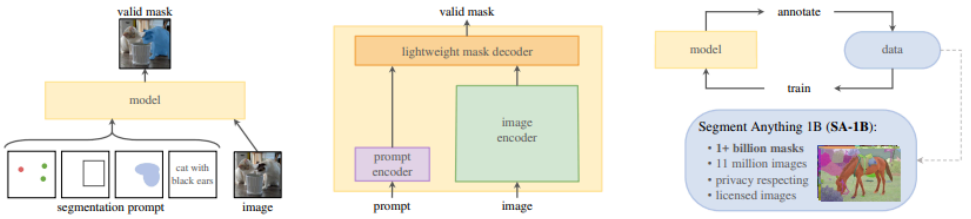
→ 이는 pre-trained에 사용되며 후에 downstream 분할 작업 프롬포트 기반으로 해결하는데 도움이 됨

> → 이때 downstream vs distilation? : downstream은 pre-trained model로 수행하는 실제 응용 작업으로 모델의 일반화 능력으로 다양한 실제 task에 적용시키는 것. 일반적으로 별도의 fine-tuning없이, prompt만 조절해서 사용함. disilation은 큰 모델 teacher의 지식을 student인 작은 모델로 옮기는 학습 기법으로 연산 효율이 좋은 모델을 만들기 위해 성능을 압축시키는 것. 보통 supervised 방식이고, teacheer의 ouput과 loss를 기반으로 학습시킴 (모델 경량화를 위한!)

2) 이에 적합한 모델 아키텍터가 무엇인지

- 위의 task를 위해선, 유연한 prompt를 지원하는 모델이 필요함. 또한 실시간으로 mask를 출력할 수 있어야 하고, interactive한 사용이 가능해야 함
- promptable → prompt 입력에 반응할 수 있도록 설계됨. 새로운 이미지 분포나 작업에도 zero-shot으로 전이가 가능해짐

- 이는 간단한 디자인으로 해결이 가능한데, **강력한 image encoder가 image embedding을 계산하고, prompt encoder가 prompt를 embedding 하여 → 경량 mask decoder에서 결합되어 분할 마ㅅ크를 예측함 : SAM**
    - 이미지 인코더, 그리고 프롬퍼트 인코더, 마스크 디코더를 분리해두었기 때문에 한 번 생성한 이미지 임베딩을 다양한 prompt에 재사용이 가능하고, 연산 비율을 줄일 수 있게 됨
    - 이때 주요 입력 prompt는 point, box, mask이며 자유형 텍스트에 대한 초기 실험 결과도 제시
    - 또한 SAM이 모호한 prompt를 처리할 수 있도록, 하나의 prompt에 대해 다양한 mask를 예측할 수 있도로 설계됨 ex. 셔츠 위의 점이 셔츠인지, 사림인지 모를 때, 두가지에 대한 마스크를 모두 변환함으로써 자연스럽게 해결함



(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 prompt-able segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

3) 이 작업과 모델을 지원할 수 있는 데이터가 무엇인지

- 모델을 훈련시키기 위해선, 다양하고 대규모의 데이터가 필요함. 이때 segmentation에 적합한 web-scale의 데이터셋을 아직 존재하지 않기에, data engine을 구축함

→ 이 data engine은 anoatation 과정을 동시에 개발하는 방식으로 1) assitant-manual에 의해 SAM이 anotator가 mask를 labeling 하는데 실시간으로 도움을 주고, 2) semi-automatic 방식으로 SAM이 일부 객체의 마스크를 자동으로 생성하고 나머지는 사람이 직접 라벨링해서 다양성을 확보하고, **3) fully automatic으로 정해진 격자 위치에 point prompt를 주고 SAM이 이미지 당 평균 100개의 고품질 마스크를 생성하는 방식**

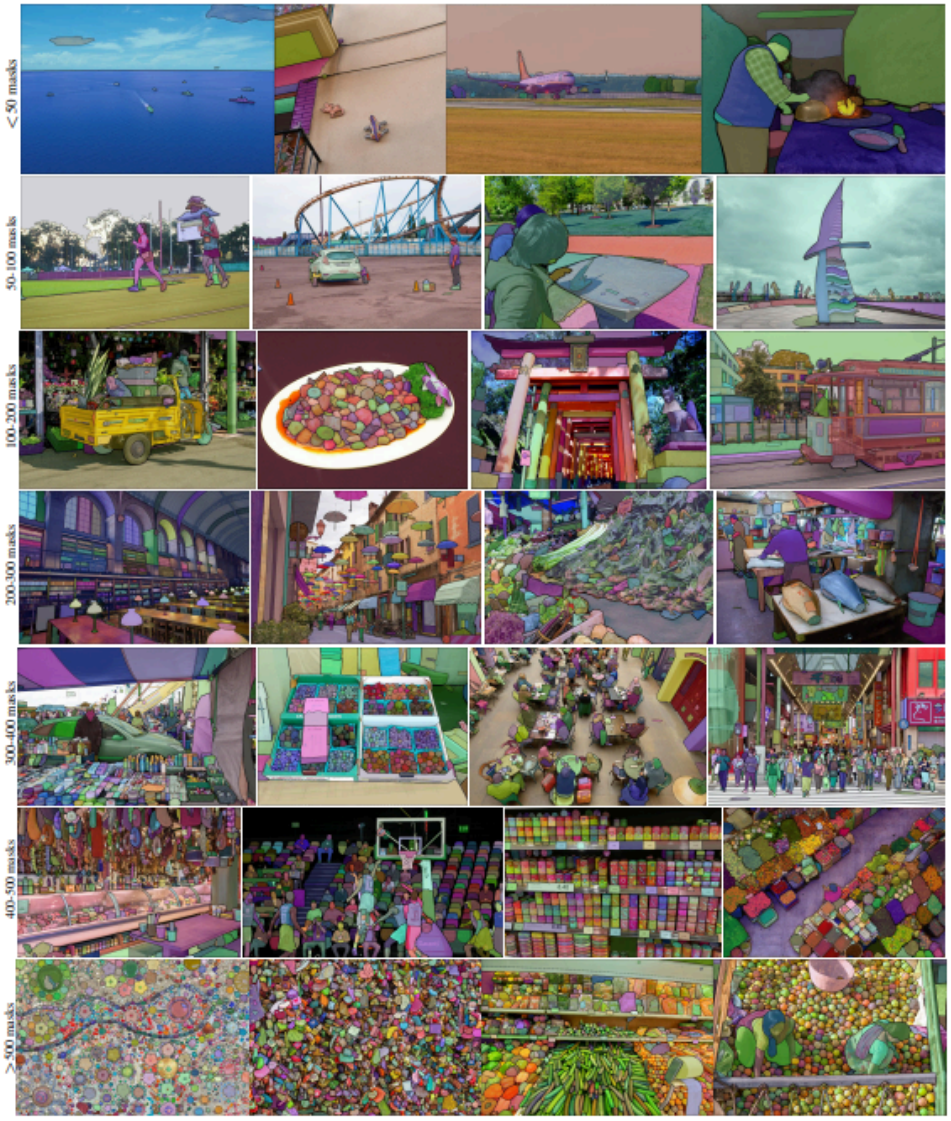


Figure 2: Example images with overlaid masks from our newly introduced dataset, **SA-1B**. SA-1B contains 11M diverse, high-resolution, licensed, and privacy protecting images and 1.1B high-quality segmentation masks. These masks were annotated *fully automatically* by SAM, and as we verify by human 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).

- **SA-1B** → 효율적인 모델을 데이터 수집 루프에 활용하여 가장 큰 규모의 분할 데이터셋을 구축함. 11백만장의 개인정보 보호 이미지에 대해 10억개 이상의 마스크를 생성) 을 제안함 (이건 위의 3) 방식으로 생성된 것) 기존의 segmenation보다 400배 많은 mask를 포함하고 있음

→ 그럼 이때, online model(data를 streaming 처럼 받으면서 모델이 계속 학습하는 방식)인지? No. 한번 pre-trained된 offline 사전학습 모델로 학습된 후에는 실시간으로 프롬포트를 입력받아 마스크를 예측할 수 있음. 따라서 web이나 client에서 실시간으로 입력을 받고 예측만 수행하는 것!

→ **즉, 본 논문에서 제시한 효율적인 모델을 이용해 데이터를 수집하고, 새로 수집한 데이터로 다시 모델 개선**

Method

Segment Anything Task

- NLP의 next token prediction 에서 영감을 받은 것 → 모호한 프롬프트에 대해 언어 모델이 말이 되는 답변을 주는 것과 유사하도록.
- 이때 prompt란, foreground/background points + box + mask + free-form text 등 '무엇을 분할할지'를 나타내는 정보임
- pre-training 방식은, training sample 마다 여러가지 프롬포트를 시뮬레이션을 한 수, 모델이 생성한 마스크를 ground-truth와 비교하아여. 매 프롬포트 마다 항상 valid한 mask를 예측할 수 있도록 함

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

- zero-shot transfer로 위 과정을 거치면, 모델을 어떤 프롬포트에도 적절하게 반응할 수 있기에, downstream 작업에도 적절한 prompt를 만들면 됨.
- 이때 하나의 모델이 미리 정의된 여러 분할 작업을 수행하는 multi-task 분할과 달리, 훈련되지 않은 새로운 task에서도 prompt만 주면 수행이 가능해짐. 따라서 훈련시 보지 못한 inference에도 prompt composition으로 해결이 가능해짐

Segment Anything Model

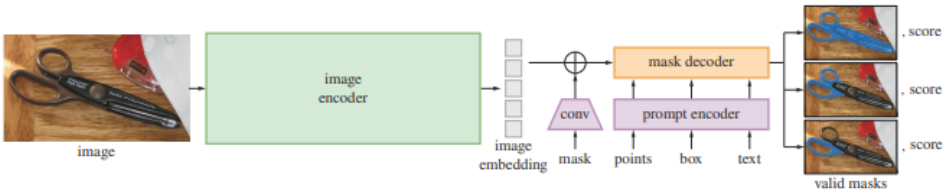


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output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

3가지 구성요소: Image Encoder, Prompt Encoder, Mask Decoder

- **Transformer 기반 비전 모델들** 위에 구축하되, amortized real-time performance을 고려

[Image Encoder]

- 확장성과 pre-training 방식을 고려하여 **MAE방식으로 사전학습된 Vision Transformer (ViT)**
- 고해상도 입력도 처리할 수 있도록 조정, 이 인코더는 이미지당 한번만 실행되고, prompt를 입력하기 전에 미리 계산이 가능함

[Prompt Encoder]

- prompt는 2가지 종류로 나눠짐 → 희소형(point, box, text), 조밀형(mask)
- 점과 박스는 위치 인코딩과 학습된 임베딩을 더해 표현함
- 자유형 텍스트는 CLIP의 text encoder를 사용함
- 마스크와 같은 조밀한 프롬프트는 합성곱으로 임베딩하여 이미지 임베딩과 요소별로 더해짐

[Mask Decoder]

- **이미지 임베딩**, **프롬프트 임베딩**, 그리고 output token을 입력으로 받아 마스크를 생성
- 구조는 Transformer decoder block을 변형한 형태로, 이후 dynamic mask prediction head가 붙음
- 이때 prompt간 self-attention, 그리고 이미지 임베딩 간 양방향 cross-attention이 포함됨
- 디코더 블록을 두 번 실행한 후, 이미지 임베딩을 업샘플링하고
  MLP를 통해 output token을 선형 분류기로 매핑, 이 분류기가 각 위치에서 마스크의 확률을 계산

[모호성 처리]

- 하나의 출력만 갖는다면, 모호한 프롬프트에 대해 **여러 가능한 마스크가 섞인 평균값이 나옴 → SAM은 하나의 프롬프트에 대해 여러 개의 마스크를 출력**하도록 변경
- 보통 **3개의 마스크 출력**으로 대부분의 경우를 처리할 수 있었음
- 학습 중에는 3 마스크 중 손실이 가장 낮은 것 하나만 역전파를 수행함
- 마스크를 순위 매기기 위해, 각 마스크에 대해 예상 IoU도 함께 예측

Segment Anything Data Engine

[1단계: 모델 보조 수동 어노테이션 (Assisted-Manual Stage)]

- 초기에는 기존의 공개 세그멘테이션 데이터셋으로 SAM을 훈련한 후, 수작업 어노테이션을 시작.
- SAM은 브라우저 내 실시간 보조 도구로 사용되며, 어노테이터는 전경/배경 포인트를 클릭해 마스크를 생성.
- 어노테이션 도구는 "브러시", "지우개" 기능도 지원해 정교한 수정이 가능.
- 모델의 반복적 개선과 함께, 이미지 인코더는 ViT-B에서 ViT-H로 확장되었고 총 6회 재학습 진행.
- 마스크 하나당 평균 어노테이션 시간은 **34초 → 14초**로 감소.
- COCO 기준 대비 6.5배 빠르며, 바운딩 박스 수준의 효율에 근접.
- **총 12만 장의 이미지에서 약 430만 개의 마스크** 수집.

[2단계: 반자동 어노테이션 (Semi-Automatic Stage)]

- 목표: **마스크 다양성 향상**.
- SAM으로 **자동으로 일부 객체 마스크를 예측**, 사람이 **나머지 덜 눈에 띄는 객체**를 라벨링.
- 자동 마스크는 1단계 데이터를 이용해 학습한 **박스 탐지기(object detector)**로부터 유도됨.
- 총 18만 장의 이미지에서 **590만 개의 추가 마스크** 생성 → 누적 1,020만 개.
- 어려운 객체가 늘어남에 따라 수동 어노테이션 시간은 다시 평균 **34초**로 증가.
- 이미지당 평균 마스크 수: **44개 → 72개**로 증가 (자동 마스크 포함).

[3단계: 완전 자동 어노테이션 (Fully Automatic Stage)]

- SAM이 충분히 개선되고 **모호성 처리 능력(ambiguity-aware)**을 갖춘 덕분에 **전면 자동화** 가능.

- 32×32 점 격자(grid)를 기반으로 SAM이 이미지 전역에 마스크를 예측.
- **하나의 점에서 전체, 부분, 세부(subpart) 객체까지 동시에 예측** 가능.
- **IoU 기반 confidence score와 마스크 안정성 검사**를 통해 신뢰도 높은 마스크만 선택.
- 중복 제거를 위한 NMS(non-maximal suppression) 적용.
- 작은 객체 품질 향상을 위해 **다중 줌-인 이미지 패치**도 처리.
- 이 단계에서 **1.1B개의 마스크를 1,100만 장의 이미지에 대해 자동 생성**하여 SA-1B 완성.

Segment Anything Dataset

- 총 **1,100만 장의 이미지**는 고해상도(평균 3300×4950)이며, **전문 사진작가로부터 라이선스를 정식 확보**.
- 개인정보 보호를 위해 얼굴, 차량 번호판 등은 블러 처리됨.
- 공개 시에는 짧은 변 길이가 1500픽셀로 축소된 이미지가 배포됨.
- **mask의 경우, 99.1%는 완전 자동 생성**, 따라서 **자동 마스크 품질이 데이터셋 신뢰성의 핵심**.
- 전문 어노테이터들이 무작위 500장의 이미지(약 5만 개 마스크)를 수정해 평가한 결과:
  - **94%는 90% 이상의 IoU**
  - **97%는 75% 이상의 IoU**
- 이는 사람 간 어노테이션 일치도(85~91% IoU)보다 높거나 비슷한 수준.
- 실험 결과(§7)에서도 다양한 데이터셋 대비 **높은 주관적 품질 평점**을 기록.
- 따라서 **완전 자동 마스크만으로도 모델 훈련에 충분히 효과적**임이 입증됨.

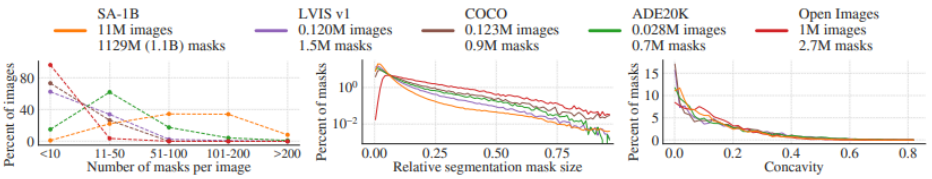


Figure 6: Dataset mask properties. The legend references the number of images and masks in each dataset. Note, that SA-1B has 11× more images and 400× more masks than the largest existing segmentation dataset Open Images [60].

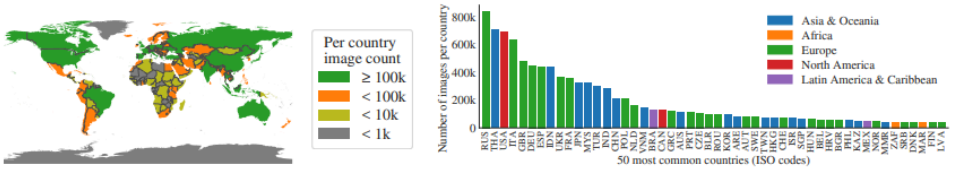


Figure 7: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.

Result & Analysis

	# 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.

	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.

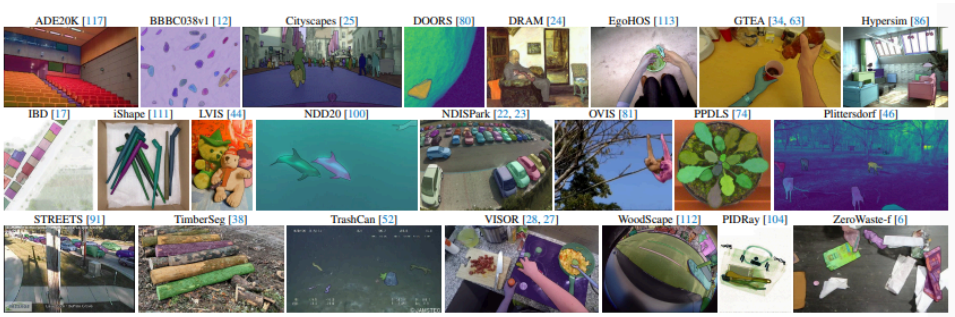
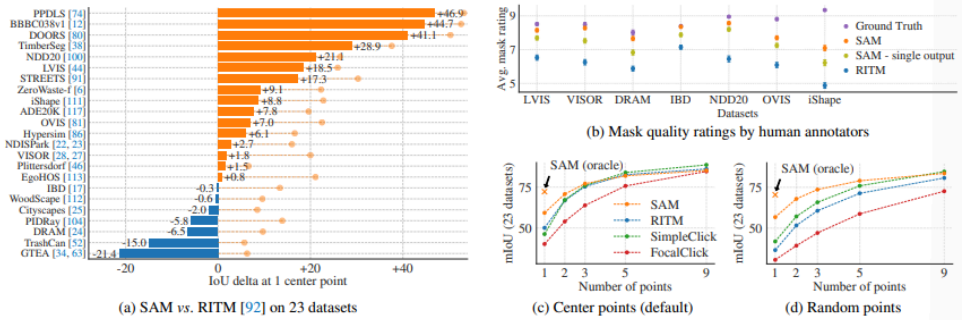


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.



(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.

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.

	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.

	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).

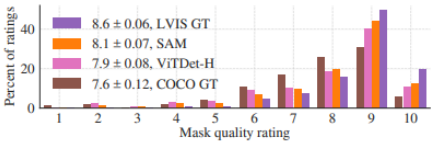


Figure 11: Mask quality rating distribution from our human study for ViTDet and SAM, both applied to LVIS ground truth boxes. We also report LVIS and COCO ground truth quality. The legend shows rating means and 95% confidence intervals. Despite its lower AP (Table 5), SAM has higher ratings than ViTDet,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

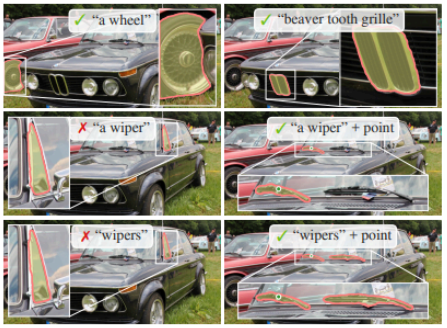


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

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