SEMINAR PAPER TALK

NAIL052 2026

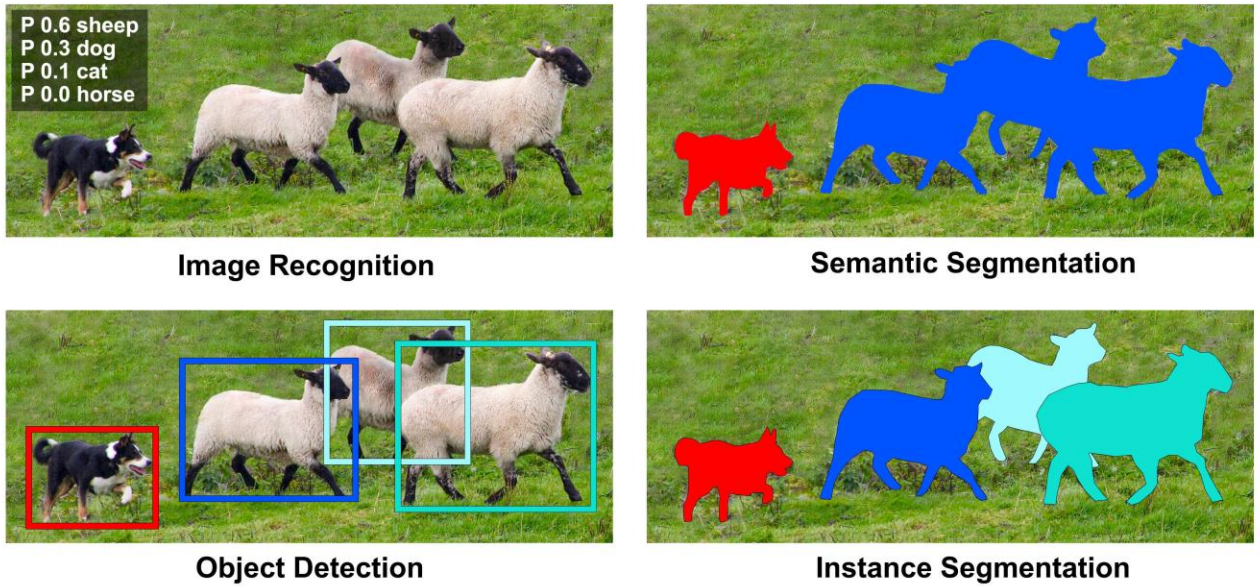
SEGMENT ANYTHING MODEL

Presenter:
Vladyslav Furda

Paper authors:
Kirillov et al., 2023.
Meta FAIR

SEGMENTATION LANDSCAPE CHECK

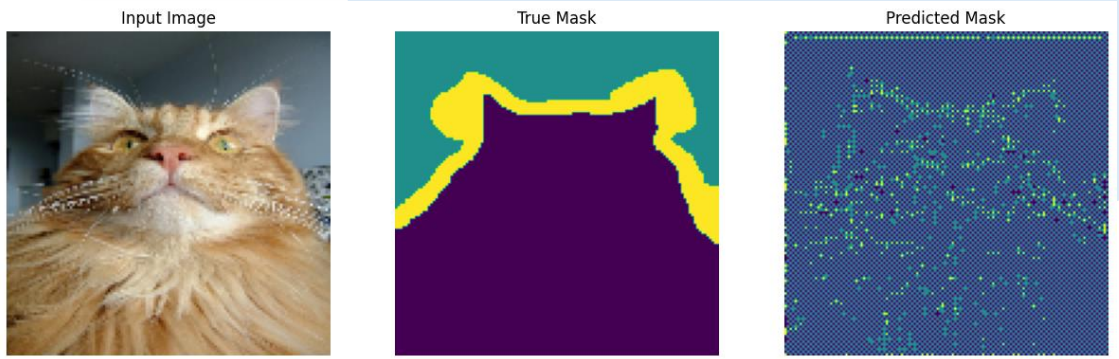
Historically, segmentation evolved from fixed-task to more flexible and now to promptable. Earlier models were trained for specific datasets + fixed classes and required retraining (or SFT) for new domains.



Tedrake (2024), *Robotic Manipulation: Perception, Planning, and Control* as MIT 6.421 course notes, online at manipulation.mit.edu.

For segmentation, on menu we have:

- Semantic segmentation (U-Net, DeepLab)
- Instance segmentation (Mask R-CNN)
- Panoptic segmentation
- Interactive segmentation (RITM, GrabCut)
- Foundation / promptable (SAM family)

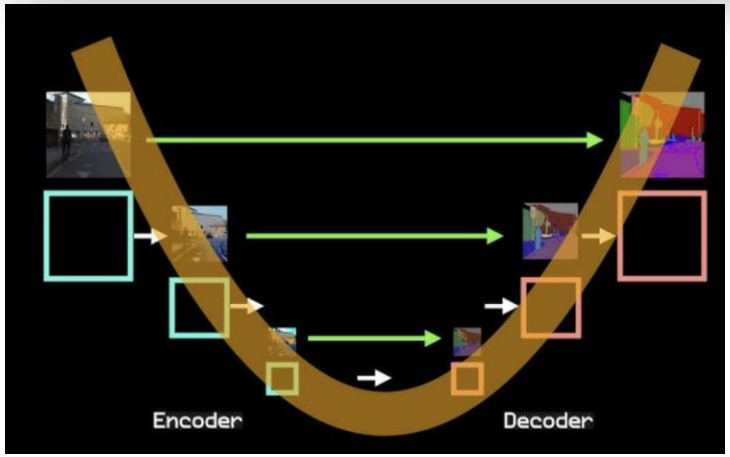


This slide shows an example of image segmentation from TensorFlow Core documentation.

Kirillov et al., 2023 (SAM)

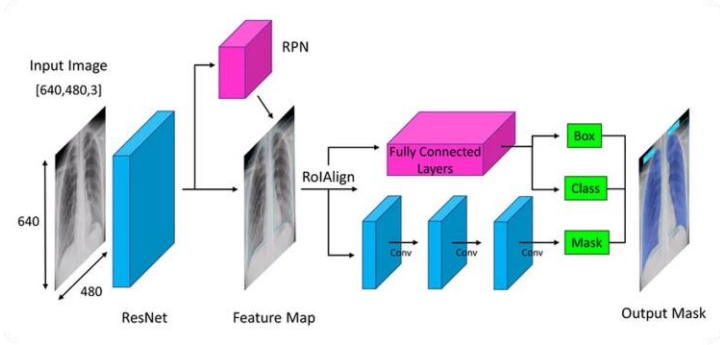
TRADITIONAL SEGMENTATION

CNN-based (U-Net, DeepLab). They feature pixel-wise classification and encoder–decoder architecture. U-Net is still considered SOTA for biomedical applications.



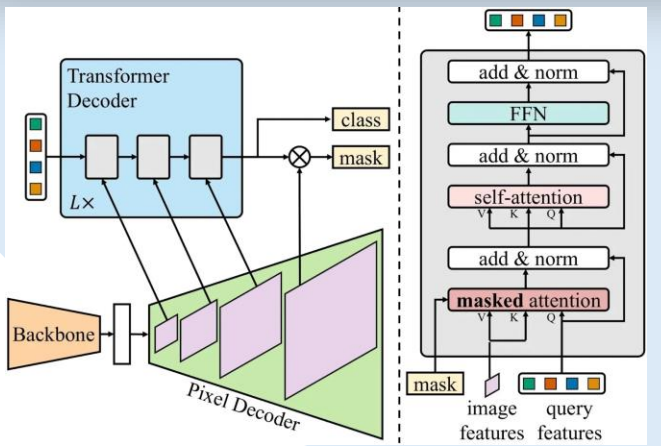
Picture from the video The U-Net (actually) explained in 10 minutes by Rupert AI on YouTube, illustrating the encoder–decoder architecture used in image segmentation

Instance segmentation (Mask R-CNN).
Detect -> classify -> segment and use bounding boxes. Those add instance-level understanding but depend on predefined classes.



Picture is based on an illustration from Ultralytics explaining the Mask R-CNN architecture, including region proposal, classification, and segmentation components.

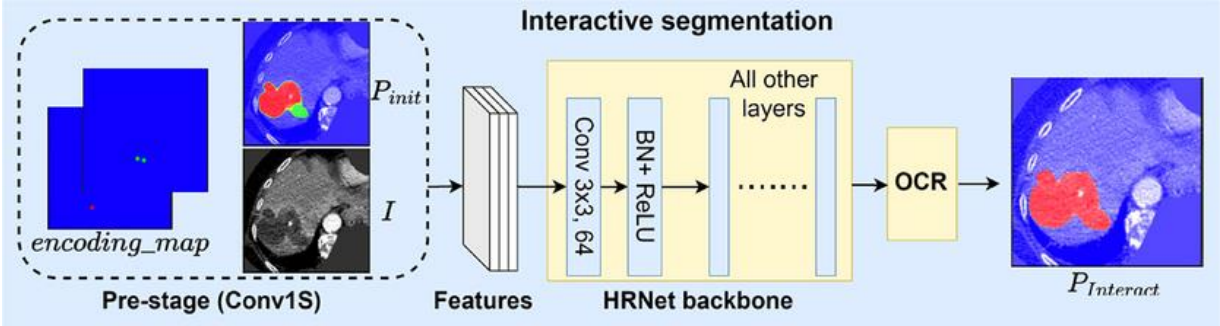
Transformer-based (Mask2Former, DETR). Generally considered more modern, unified frameworks with strong benchmarks but hard to train and adapt.



Picture is an illustration from the Roboflow Blog's guide on Mask2Former, illustrating the transformer-based decoder and pixel-level segmentation architecture.

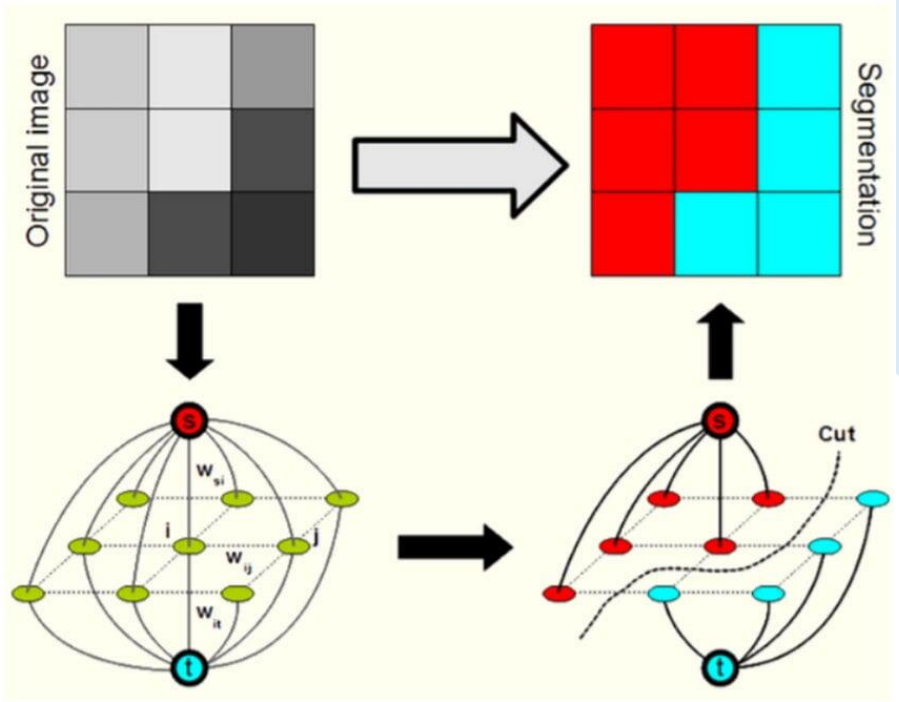
Kirillov et al., 2023 (SAM)

INTERACTIVE SEGMENTATION



Picture is an illustration of the RITM architecture from a ResearchGate figure, showing how user click points are encoded for interactive segmentation (accessed March 30, 2026).

Interactive methods like GrabCut and RITM are closest to SAM conceptually. But they require multiple corrections and do not generalize well.

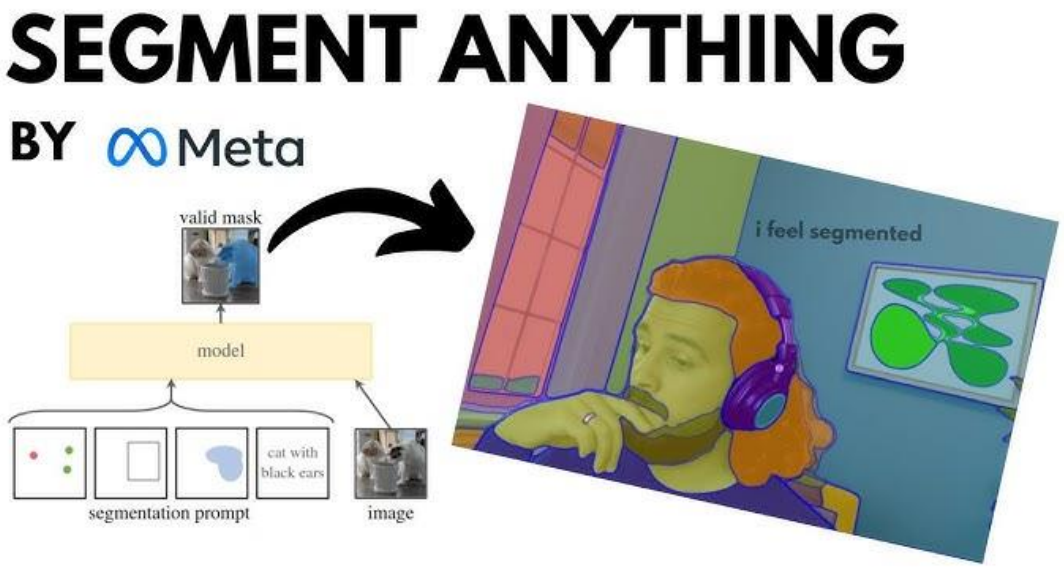


Picture is an illustration of the unsupervised GrabCut algorithm sourced from ResearchGate (accessed March 30, 2026).

Kirillov et al., 2023 (SAM)

WHY SAM EVEN MATTERS?

In NLP, promptable foundation models (LLMs) already showed strong transfer learning. In vision, CLIP showed that prompt-based transfer can work for image-text alignment too.



Picture references the video Segment Anything Model (SAM) – Foundational Model Deep Dive by Deep Learning with Yacine on YouTube.

Classical segmentation is very task-specific, we often tend to train some suitable model to fit our domain and data.

Zero-shot generalization in segmentation was weak + no web-scale segmentation dataset, like we have for LLMs.

Paper asks 3 questions:

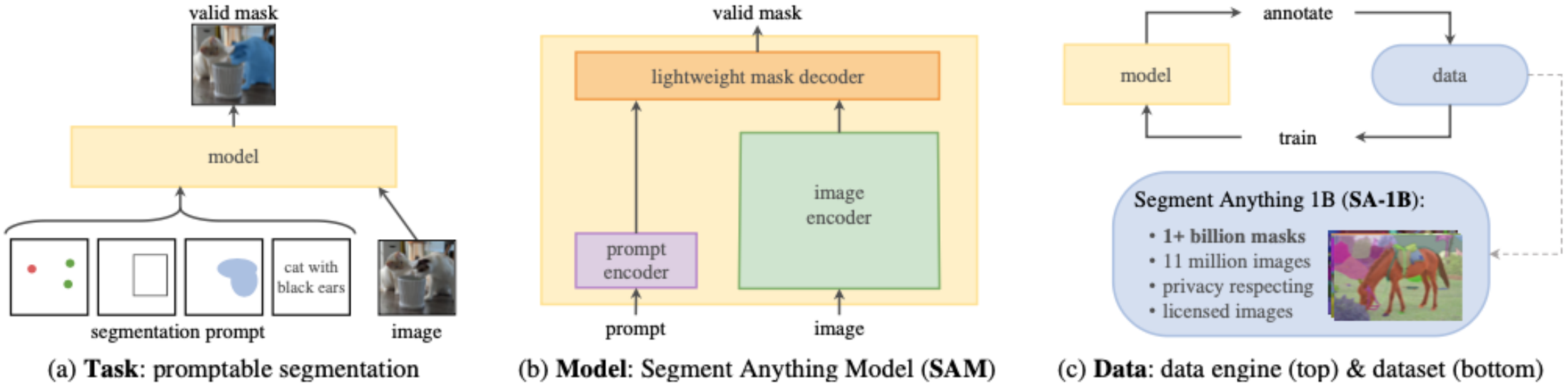
1. What task enables zero-shot segmentation?
2. What model supports that task?
3. Where do we get enough data?

Kirillov et al., 2023 (SAM)

TASK, MODEL, DATA

Authors try to build a foundation model for image segmentation. This is a large-scale AI model trained on a vast, diverse dataset that can be fine-tuned to a wide range of downstream tasks, fitting your domain.

> The success of this plan hinges on three components: task, model, and data itself.



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

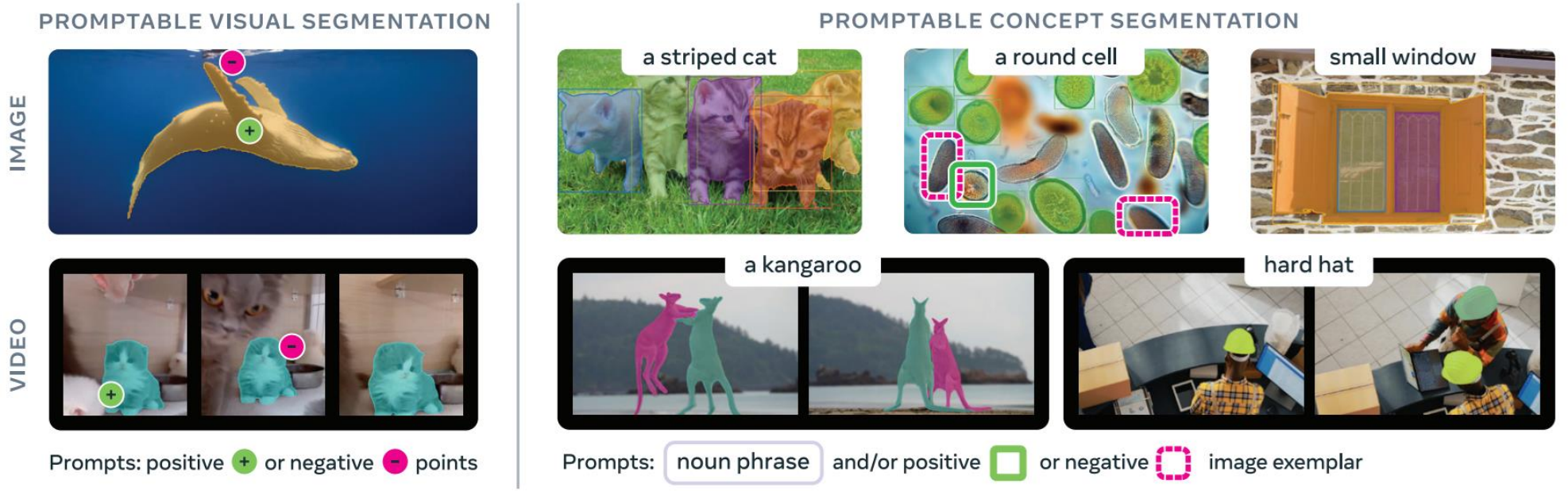
Kirillov et al. (2023) illustrate this in a figure from Segment Anything, an arXiv paper on general-purpose image segmentation.

Kirillov et al., 2023 (SAM)

NOW MORE ABOUT THE TASK

One cool thing about SAM is so-called promptable segmentation, where the goal is to return a valid segmentation mask (or masks) given any segmentation prompt. This is important because prompts can be ambiguous. A point on a wristwatch might refer to the watch itself, hand with a watch or the person.

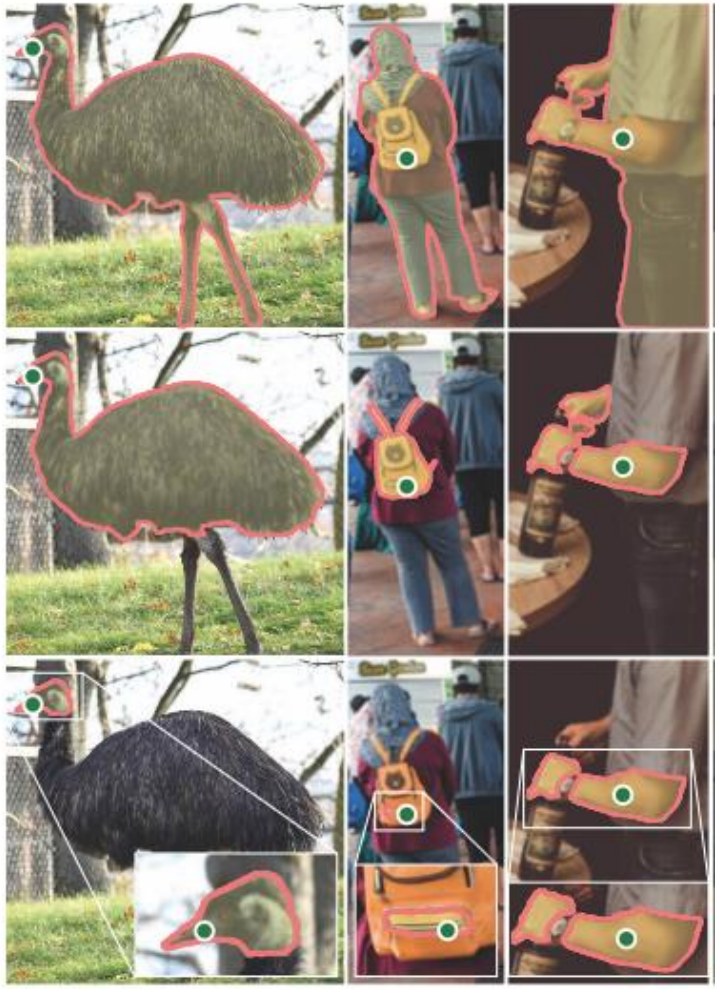
> Prompts can be boxes, points, free-form text, other masks etc. Basically any information telling model what to segment.



Roboflow Blog (2025), What Is Promptable Concept Segmentation (PCS)?

Kirillov et al., 2023 (SAM)

WHY AMBIGUITY MATTERS



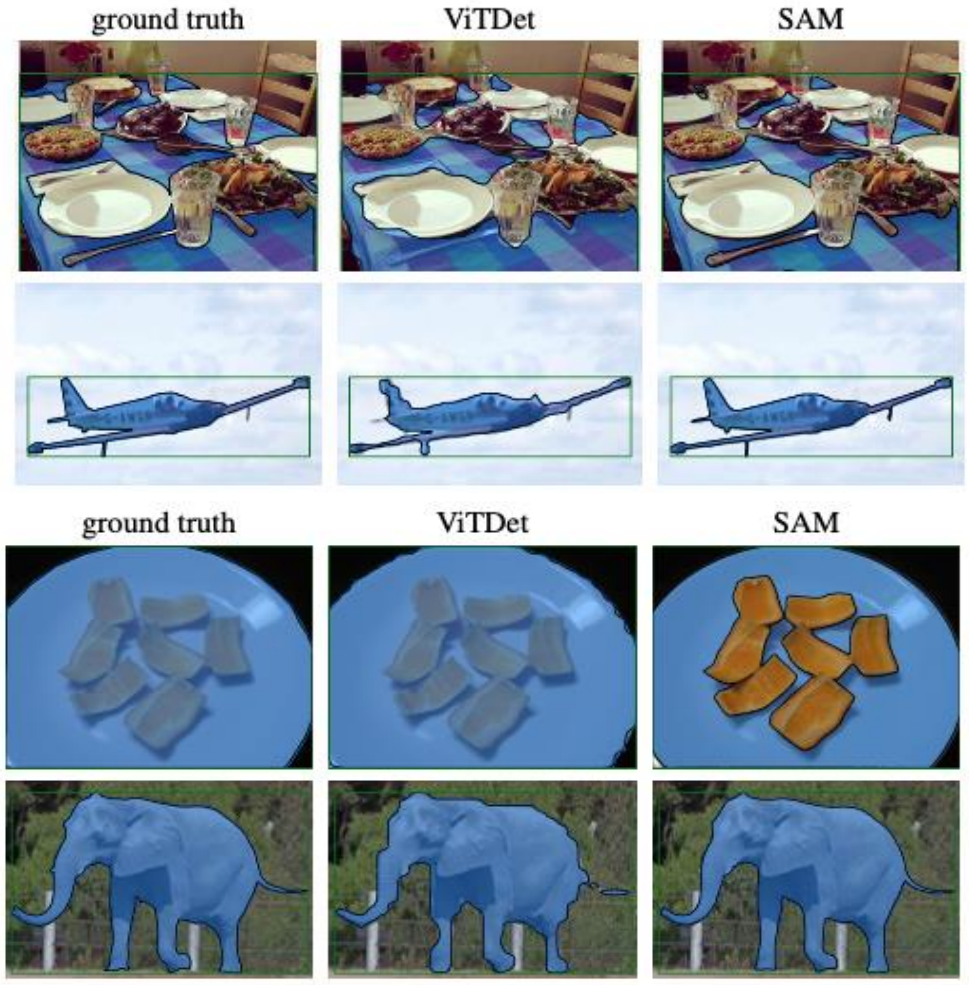
Kirillov et al. (2023) illustrate this in a figure from Segment Anything. Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

> Ambiguous prompts are common and single-point prompting is ill-posed by default. SAM is designed to return one of several valid interpretations.

Traditional evaluation assumes one ground-truth mask. But in real use, a single point can correspond to multiple sensible objects. The authors embrace ambiguity instead of forcing the model to average it away. This task is handy because it leads to a natural pre-training algorithm and a general method for zero-shot transfer to downstream segmentation tasks via prompting.

Kirillov et al., 2023 (SAM)

WHY THIS HELPS FOR 0-SHOT



Kirillov et al. (2023) illustrate zero-shot instance segmentation on LVIS v1. SAM produces higher quality masks than ViTDet. As a zero-shot model, SAM does learn specific training data biases

Prompt engineering as the transfer mechanism:

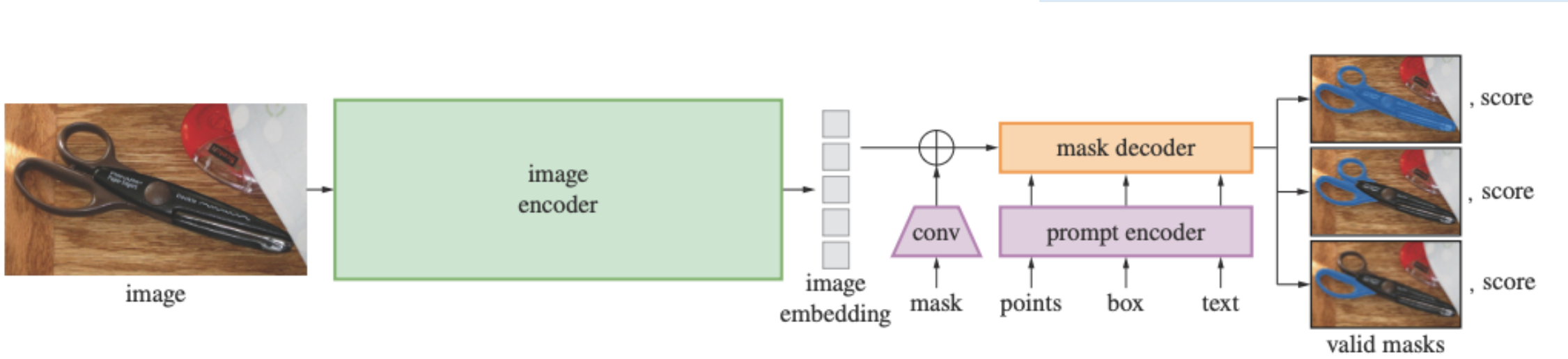
Detector box + SAM → instance segmentation
Grid points + SAM → segment everything / proposals
Text embedding + SAM → text-to-mask
This is **task generalization**, not just multi-task learning

Authors claim that many downstream tasks can be reframed as prompting. And that is quite different from a classic multi-task model trained on a fixed number of tasks.

Here, the same segmentation model can become part of larger systems during inference. This is one reason the paper was so impactful: it made SAM a reusable module, not just a benchmark model.

Kirillov et al., 2023 (SAM)

SAM ARCHITECTURE OVERVIEW



Kirillov et al. (2023) illustrate 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.

SAM has three components: an image encoder, a flexible prompt encoder, and a fast mask decoder.

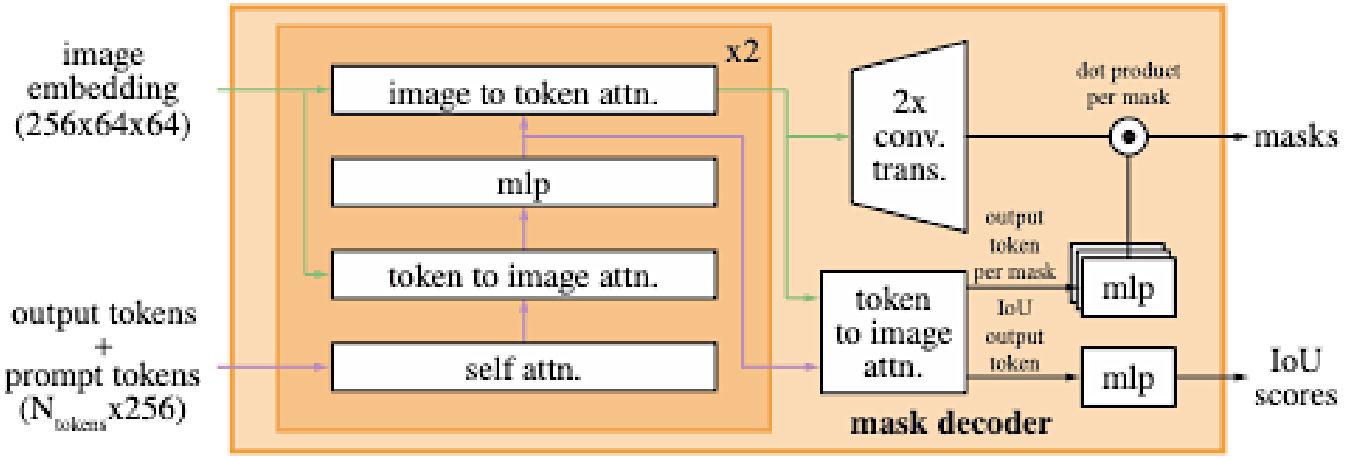
Authors consider two sets of prompts: sparse (points, boxes, text) and dense (masks). Sparse are represented by positional encodings with learned embeddings and CLIP text encoder. Dense prompts are processed using convolutions.

SAM separates expensive image processing from cheap prompt-conditioned prediction.

The image encoder (MAE pre-trained ViT) computes an embedding once. Then different prompts can query that embedding efficiently. This is the trick that makes interactive use feasible.

Kirillov et al., 2023 (SAM)

SAM ARCHITECTURE OVERVIEW



Kirillov et al. (2023) illustrate 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.)

Mask decoder efficiently maps the image embedding, prompt embeddings, and an output token to a mask. This design, inspired by Per-Pixel Classification Is Not All You Need for Semantic Segmentation (Cheng et al., 2021) and End-to-End Object Detection with Transformers (Carion et al., 2020), employs a modification of a Transformer decoder block [103] followed by a dynamic mask prediction head.

That modified decoder uses prompt self-attention and cross-attention in two directions (prompt-to-image embedding and vice-versa) to update all embeddings. After we up-sample 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.

Kirillov et al., 2023 (SAM)

AMBIGUITY-AWARE + EFFECTIVE

whole

part

subpart

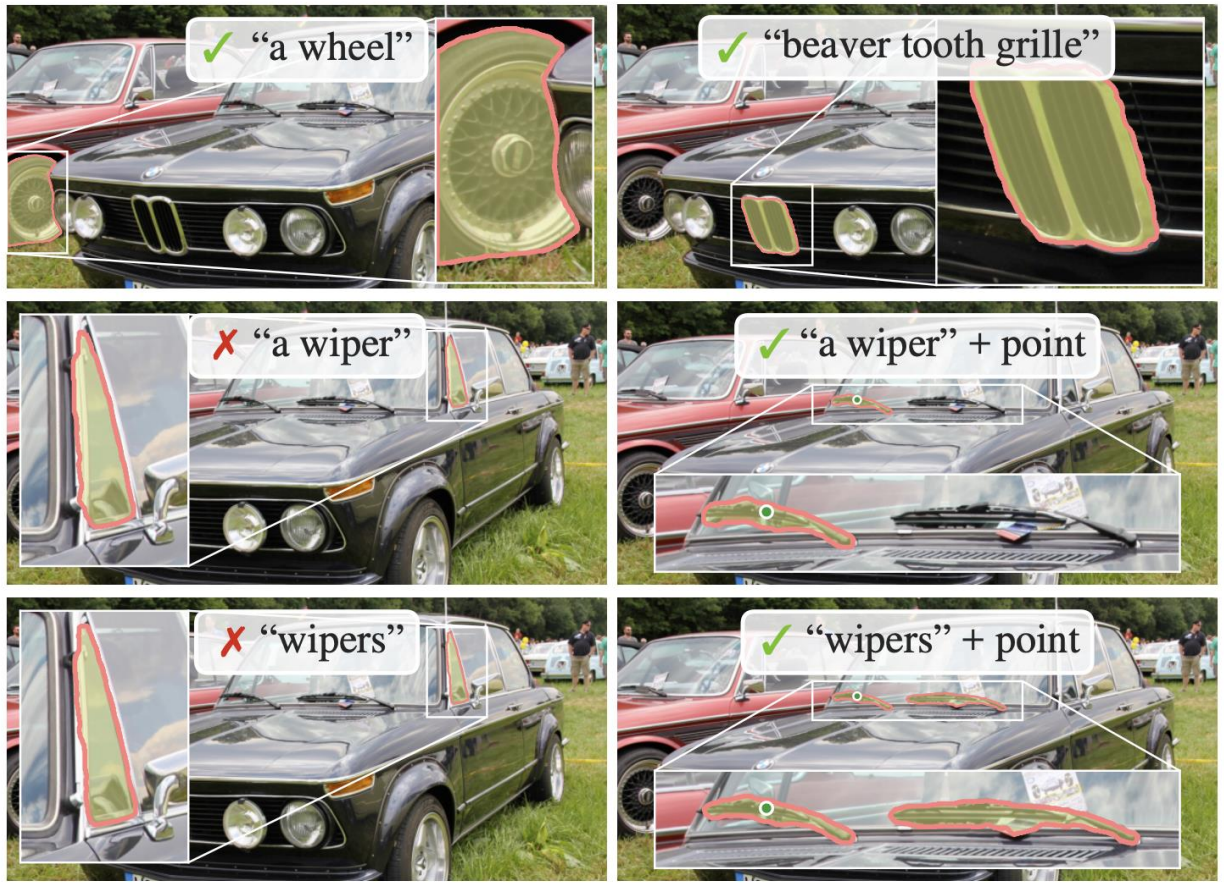
Kirillov et al. (2023) illustrate whole, part and subpart predicted masks.

SAM predicts 3 masks for a single ambiguous prompt.

It also predicts confidence / estimated IoU for ranking. Prompt encoder + decoder run in 50 ms in browser after image embedding is computed

If SAM predicted only one mask, it would often average multiple valid interpretations. Instead, it predicts multiple candidates. The paper says 3 masks are usually enough for common nested structures: whole, part, subpart. After the image embedding is ready, prompt-time inference is very fast, which is crucial for annotation tools and interactive demos.

AMBIGUITY-AWARE + EFFECTIVE



Kirillov et al. (2023) illustrate 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

For Zero-Shot Text-to-Mask task authors for each manually collected mask with area larger than 100^2 extracted the CLIP image embedding. Then, during training, they prompted SAM with the extracted CLIP image embeddings as its first interaction.

It shows that SAM can segment objects based on simple text prompts like ”a wheel” as well as phrases like “beaver tooth grille”.

Kirillov et al., 2023 (SAM)

WHAT ABOUT TRAINING?

$$\mathcal{L}_{\text{focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t).$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i p_i g_i}{\sum_i p_i + \sum_i g_i + \epsilon}$$

Source: my bachelor's work

$$\mathcal{L}_{\text{Tversky}} = 1 - \frac{\sum_i p_i g_i}{\sum_i p_i g_i + \alpha \sum_i p_i (1 - g_i) + \beta \sum_i (1 - p_i) g_i + \epsilon}$$

I personally used Tversky Loss - generalization of Dice Loss that introduces parameters α and β to control the penalties for FP and FN.

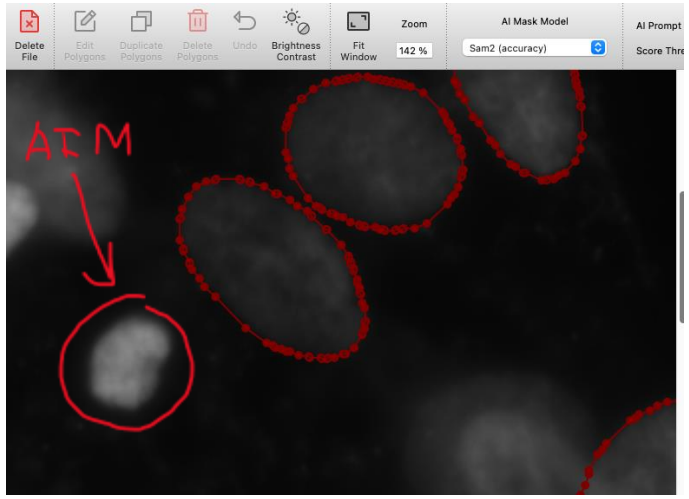
Loss = focal loss + dice loss. It works very nice because it balances regional accuracy with pixel-wise precision, specifically addressing severe class imbalances. Dice loss focuses on maximizing overlapping regions (global structure), while Focal loss forces the model to focus on hard-to-classify pixels (local details, boundaries).

The training setup imitates real interactive segmentation. The model sees an initial prompt, predicts a mask, then receives corrective prompts from the error region. This makes SAM naturally suitable for annotation workflows and iterative refinement. The paper's framing is elegant: train the model in the same style it will later be used.

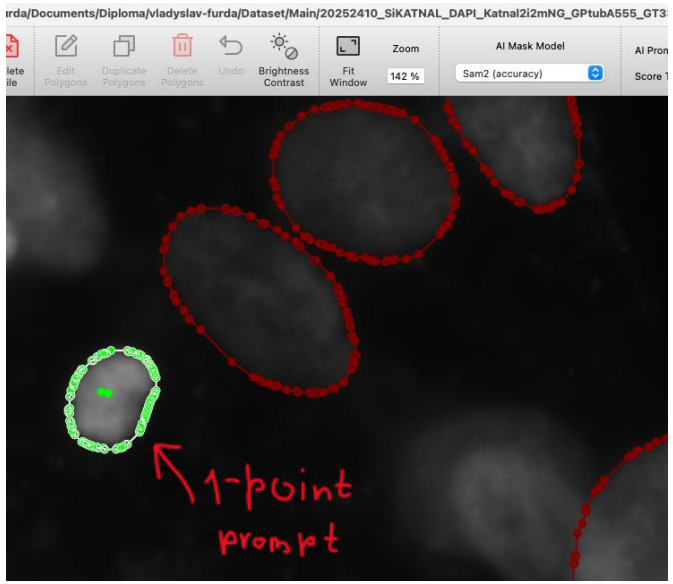
Kirillov et al., 2023 (SAM)

DATA ENGINE: stage 1

1.



2.



3.

Screenshots demonstrating SAM2 performance for annotated segmentation of nucleus in microscopy image. Software: LabelMe.

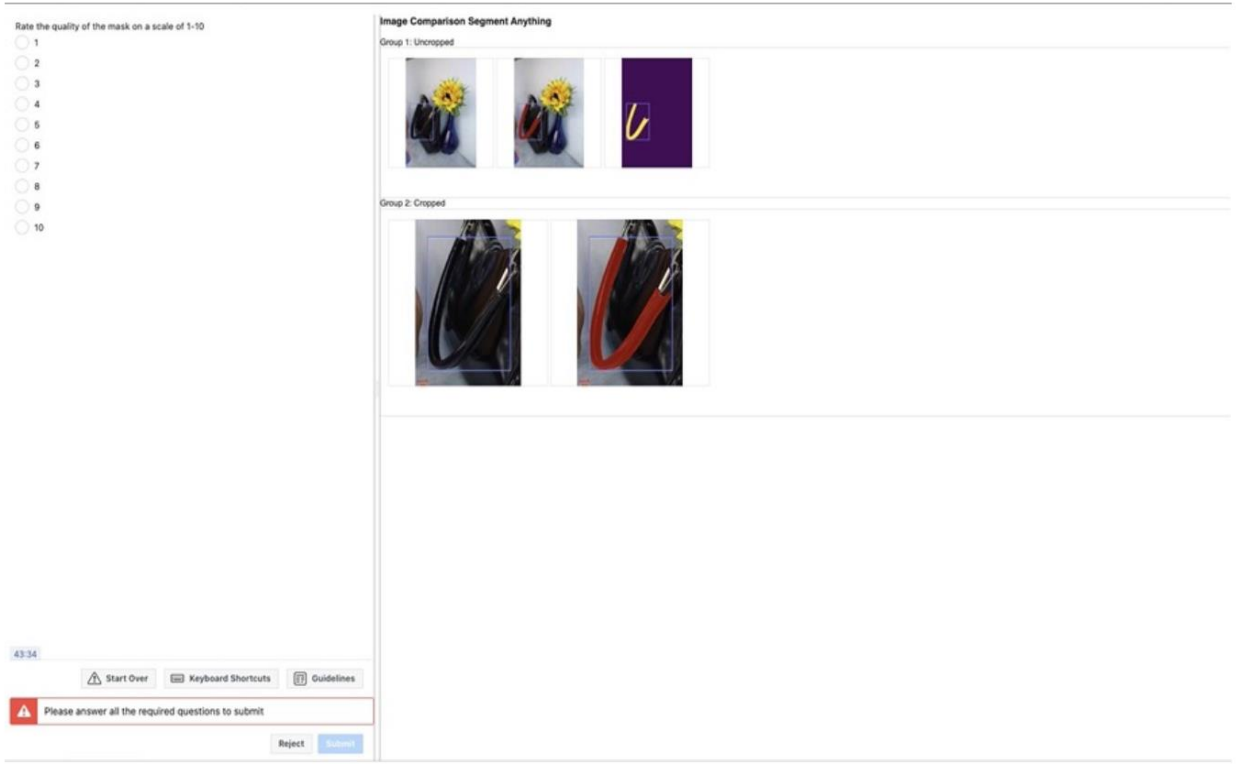
Professional annotators used browser tool powered by SAM to segment objects on pictures. They used foreground/background clicks + brush/eraser tools. Annotation time dropped significantly from 34 s to 14 s per mask.

Stage output: 4.3M masks from 120k images.

So, at this stage the model is already useful early on, and annotators use it to label faster. As the model improves, annotation gets faster and mask count per image increases. The annotation loop is central to the whole project.

Kirillov et al., 2023 (SAM)

DATA ENGINE: stage 2&3



Example interface page of SAM automatic mask evaluation by human annotator. There were five images on the right and a question box on the left.

Stage 2:

SAM automatically detects confident masks and annotators only fill in missing objects. Results in +5.9M masks from 180k images

Stage 3:

Is fully automatic!

Authors sample 32×32 grid of prompt points. Further refine masks by using stable-mask filtering + NMS + zoomed crops.

Stage 2 improves diversity by pushing annotators toward less obvious objects.

Stage 3 becomes possible only after the model is good enough and ambiguity-aware.

The fully automatic pipeline selects confident, stable masks and suppresses duplicates. This final stage is what scales the dataset to the billion-mask level.

Kirillov et al., 2023 (SAM)

SA-1B DATASET

Kirillov et al. (2023) illustrate example images with overlaid masks from our newly introduced dataset. Those masks were annotated automatically using SAM.

11M images, 1.1B masks with average of 100 masks per image. High-resolution images, faces and license plates blurred.

This was by far the largest segmentation dataset at the time. The released images are downsampled a bit, but still much higher resolution than datasets like COCO. Importantly, 99.1% of masks are fully automatic in the final release. That sounds risky, but 94% of automatic vs corrected mask pairs had IoU > 90% and 97% had IoU > 75%.

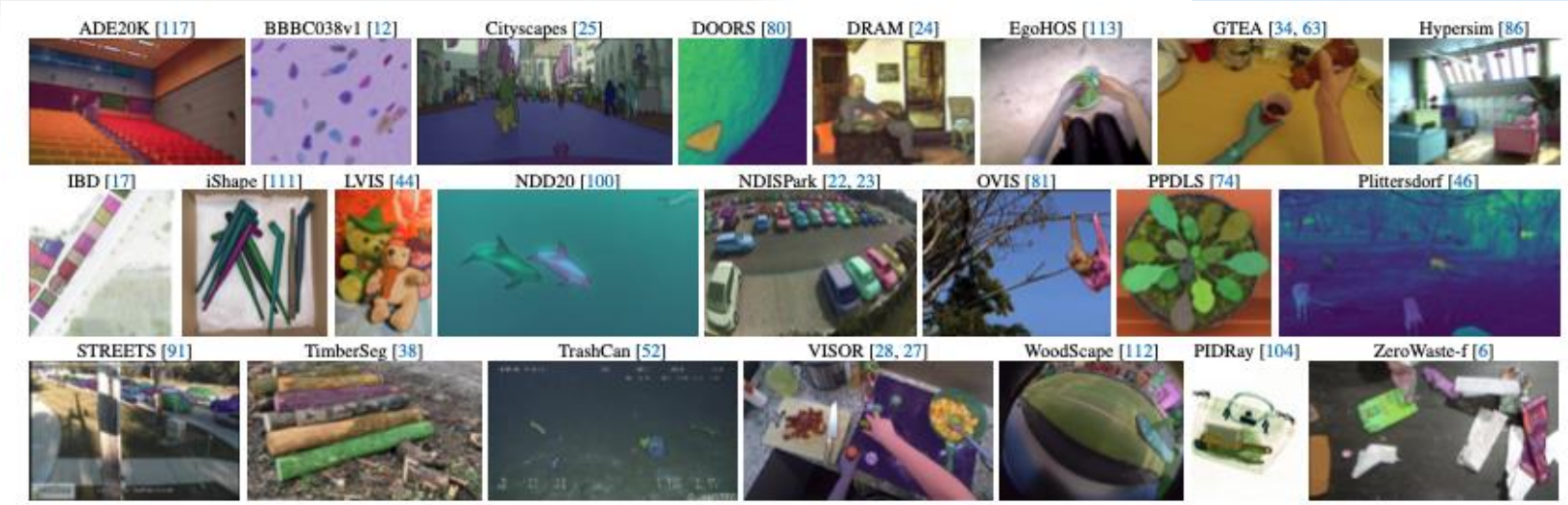
The authors validate mask quality by asking professionals to correct masks and measuring agreement.

Kirillov et al., 2023 (SAM)

EVALUTATION SETUP

23 diverse datasets. Domains include microscopy, underwater, X-ray, paintings, driving, egocentric, drones, synthetic scenes. Main tasks were single-point valid mask , edge detection, object proposals, instance segmentation, text-to-mask etc.

A strong part of the paper is breadth of evaluation. They do not just test on COCO-like images. The benchmark spans very different domains, which is indeed what a foundation-model claim should be tested on. Human evaluation is important here because with ambiguous prompts, a low IoU does not always mean the mask is bad.

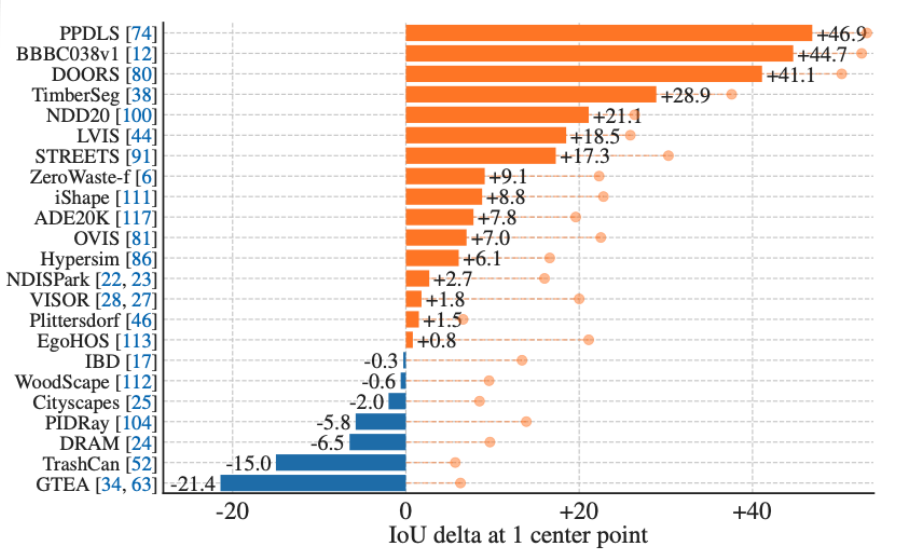


Kirillov et al. (2023) illustrate samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.

Kirillov et al., 2023 (SAM)

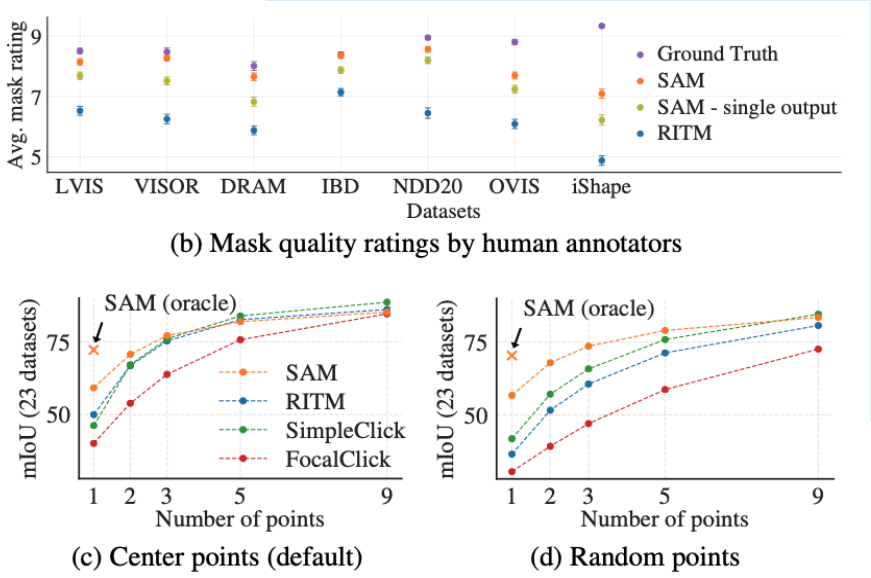
MAIN RESULT

SAM beats RITM on 16/23 datasets by standard mIoU. “Oracle” version (always select mask with higher IoU with GT) beats RITM on all 23. Human ratings consistently favor SAM. Mean ratings are often in the 7/10–9/10 range.



Kirillov et al. (2023) illustrate SAM vs. RITM on 23 datasets

This is probably the core experimental result. The oracle result is very revealing: it shows that ambiguity, and not poor mask generation, often explains lower automatic scores. The paper argues that SAM learns to output valid masks from weak prompts, even when the dataset provides only one canonical answer.



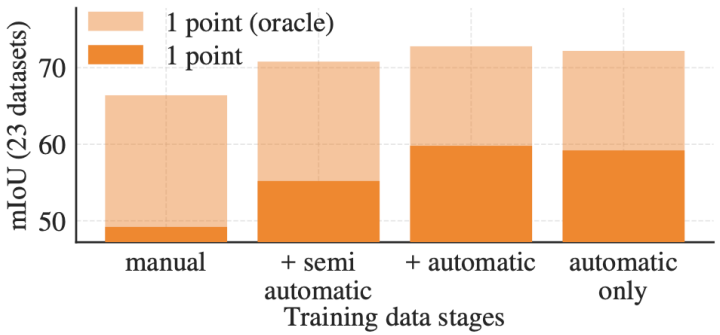
Kirillov et al. (2023) illustrate per-dataset comparison of mask quality ratings.

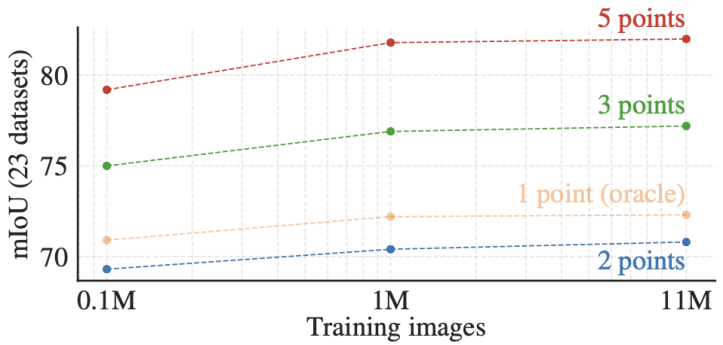
Kirillov et al., 2023 (SAM)

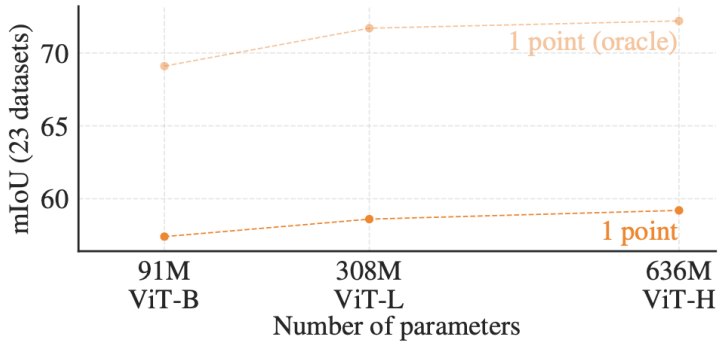
ABLATIONS, LIMITATIONS ETC.

Every data-engine stage helps and using only automatic data from last stage is only 0.5 mIoU worse than using all stages. From their studies, 1M images gives results comparable to full 11M in ablation, but 100k lowers performance significantly. ViT-H improves over ViT-B, but gains saturate vs ViT-L.

One especially practical result is that about 10% of the full dataset already gets close to full-dataset performance. SAM is definetely not the final answer for every segmentation problem, but it kind of changed how the field thinks about segmentation interfaces and data scaling.







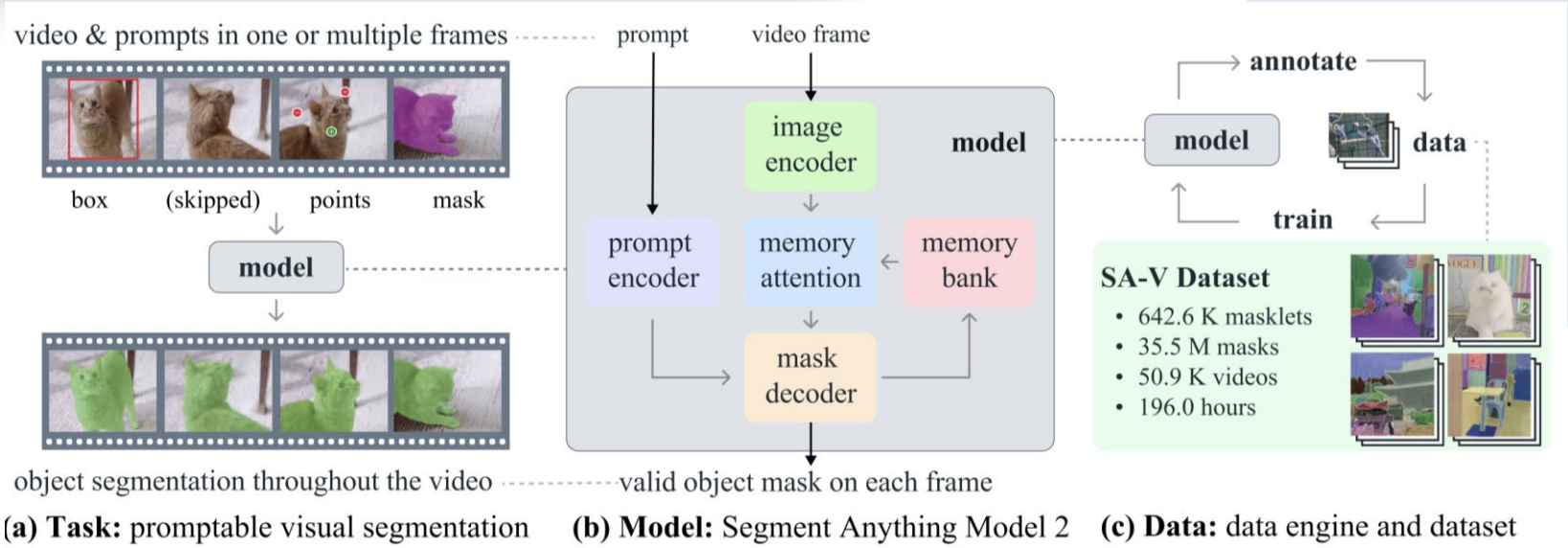
Kirillov et al. (2023) illustrate ablation studies of data engine stages, image encoder scaling, and training data scaling.

Kirillov et al., 2023 (SAM)

SAM2

Original SAM is stateless — it processes each image independently. This limits it in real-world scenarios like video. SAM2 extends original SAM by introducing memory across frames that support video segmentation and object tracking.

Basically, SAM2 treats an image as a 1-frame video. It handles motion, occlusion and appearance changes. New SA-V dataset has 51k videos and 643k masks. Reported several efficiency improvements such as model being 6x faster and requiring 3x less user interactions for correct segmentation.

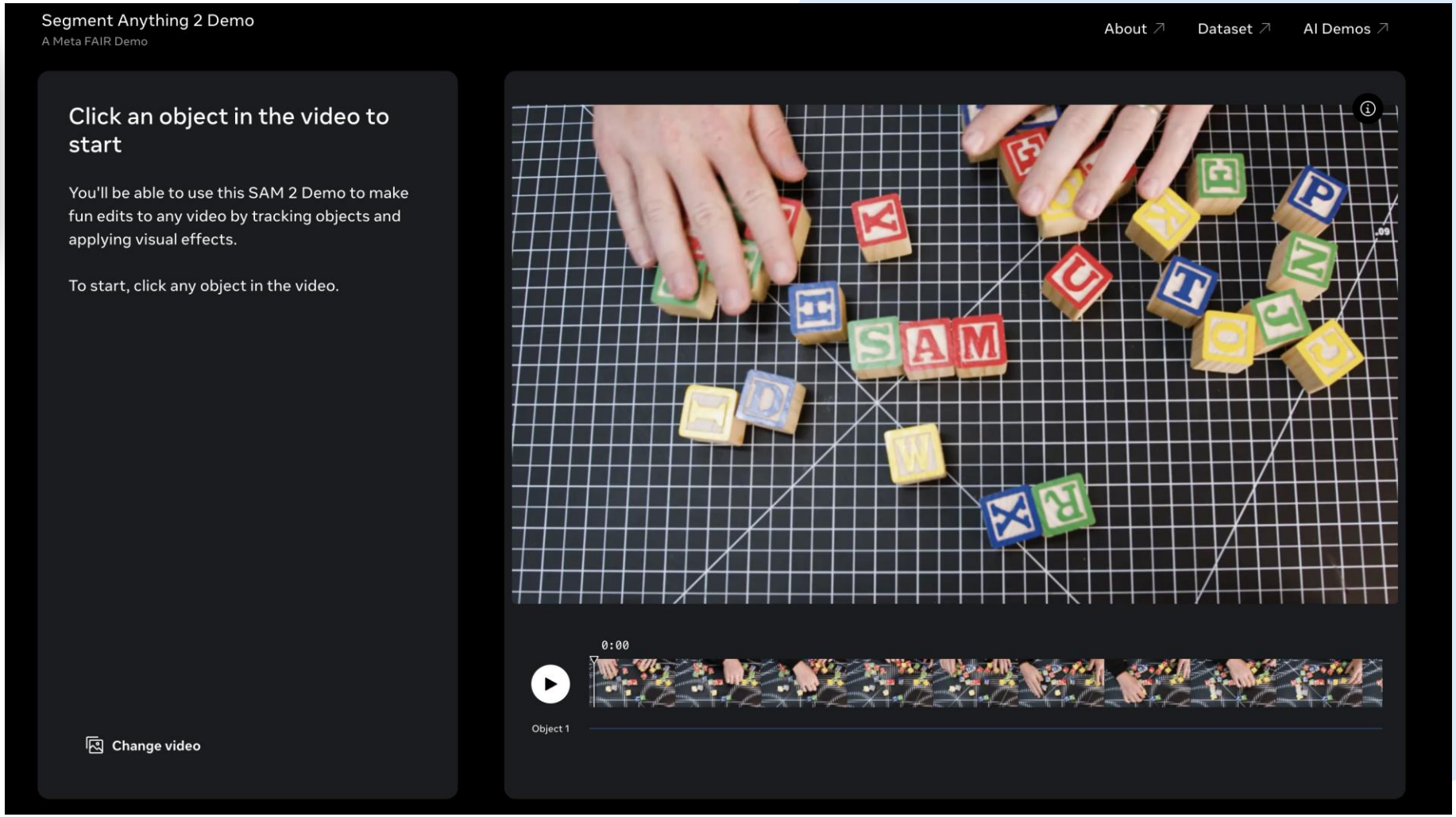


This figure from the paper SAM 2: Segment Anything in Images and Videos by Nikhila Ravi et al., Meta AI.

Kirillov et al., 2023 (SAM)

SAM2

We can select any video we want to analyze

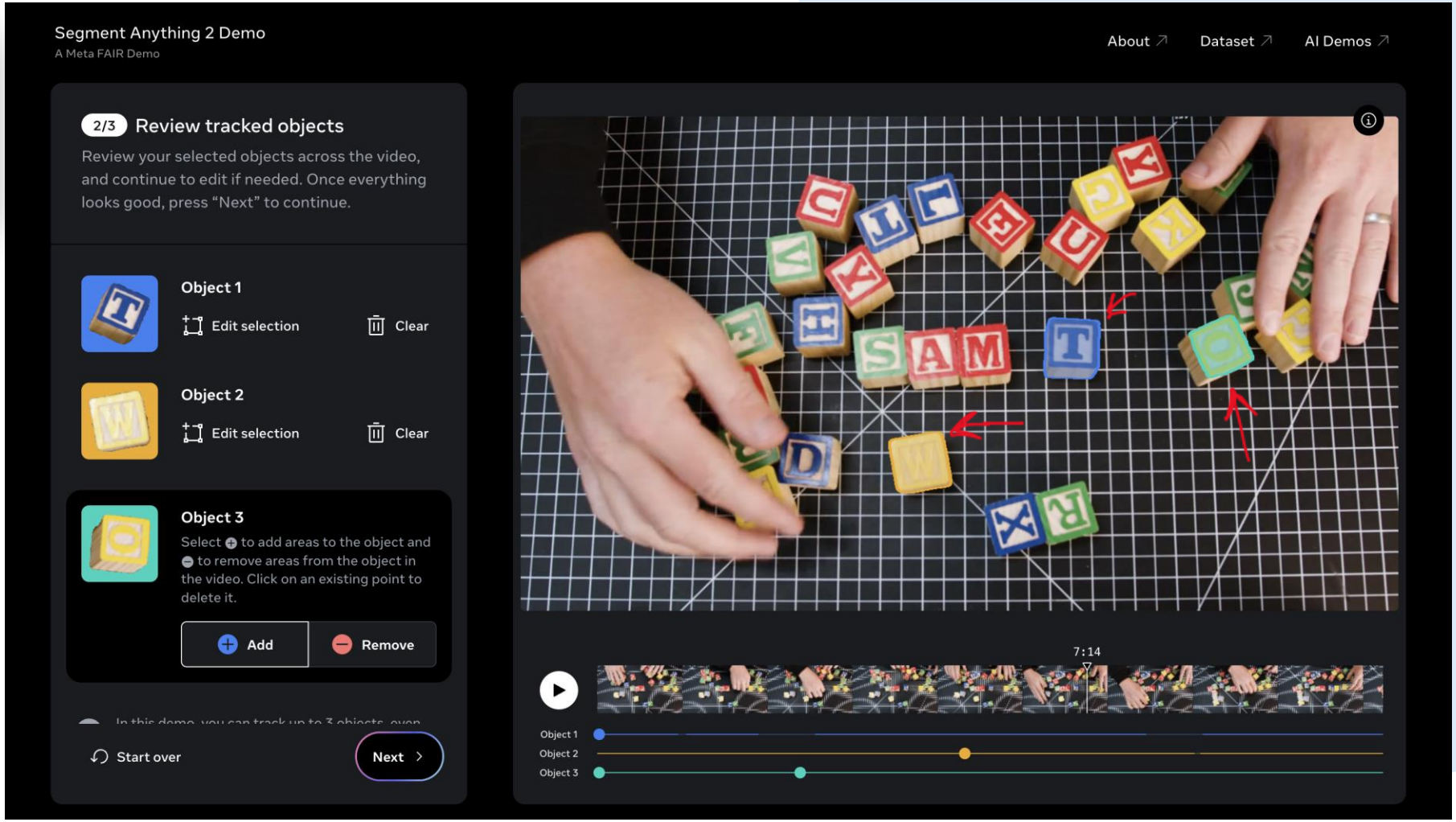


Demo from https://sam2.metademolab.com/demo

Kirillov et al., 2023 (SAM)

SAM2

We may select any object that we would like to track and segment in our selected video



Demo from https://sam2.metademolab.com/demo

Kirillov et al., 2023 (SAM)

SAM2

I applied desaturation on a background to visually highlight segmented toy cubes

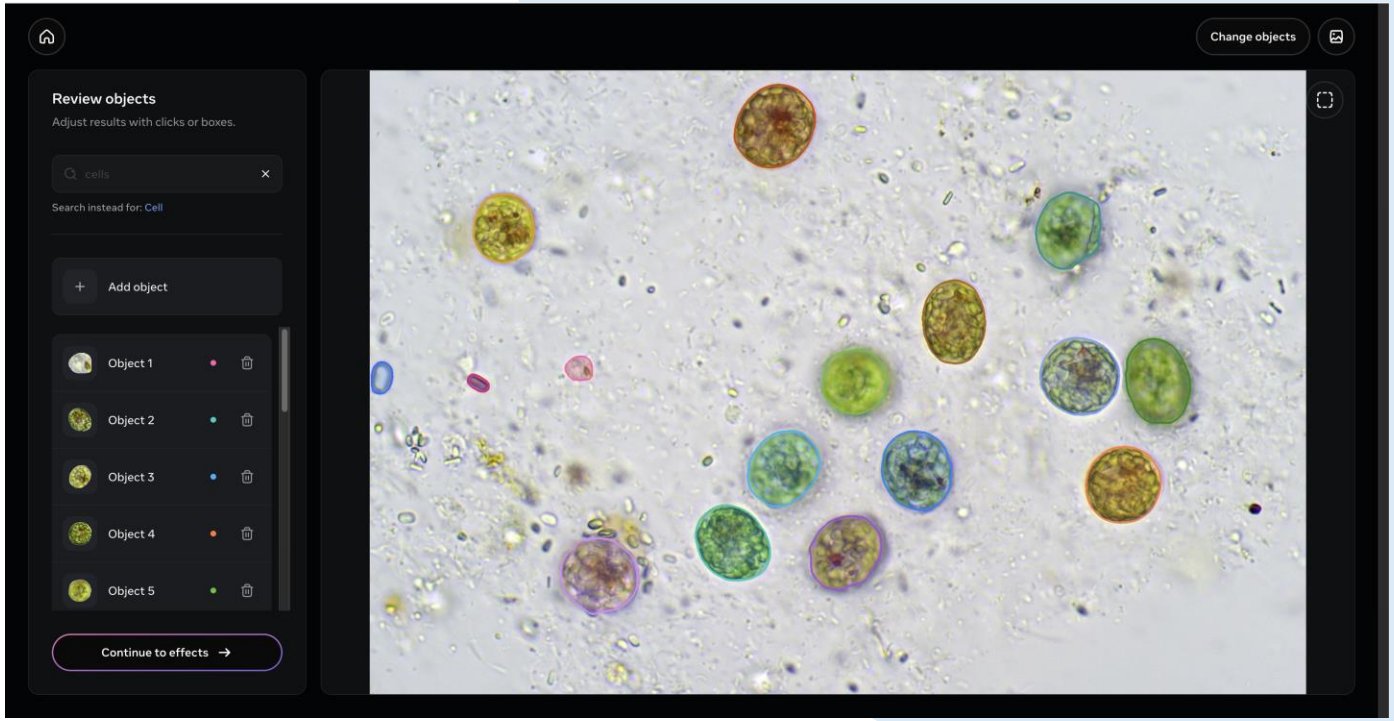
Demo video from https://sam2.metademolab.com/demo

Kirillov et al., 2023 (SAM)

SAM3

SAM3 moves from prompts towards concepts. It tries to understanding what an object is, not just where it is. Instead of clicking to segment all dogs it requires some vision-language alignment (similar to CLIP). SAM3 fully supports text prompting.

Core new task is Promptable Concept Segmentation (PCS). SAM3 was trained with 4M concept labels and can generalize to rare objects, long-tail categories and compositional prompts e.g. red striped shirt. No fixed class list anymore (unlike classic segmentation).



SAM3 demonstration from aidemos.meta.com

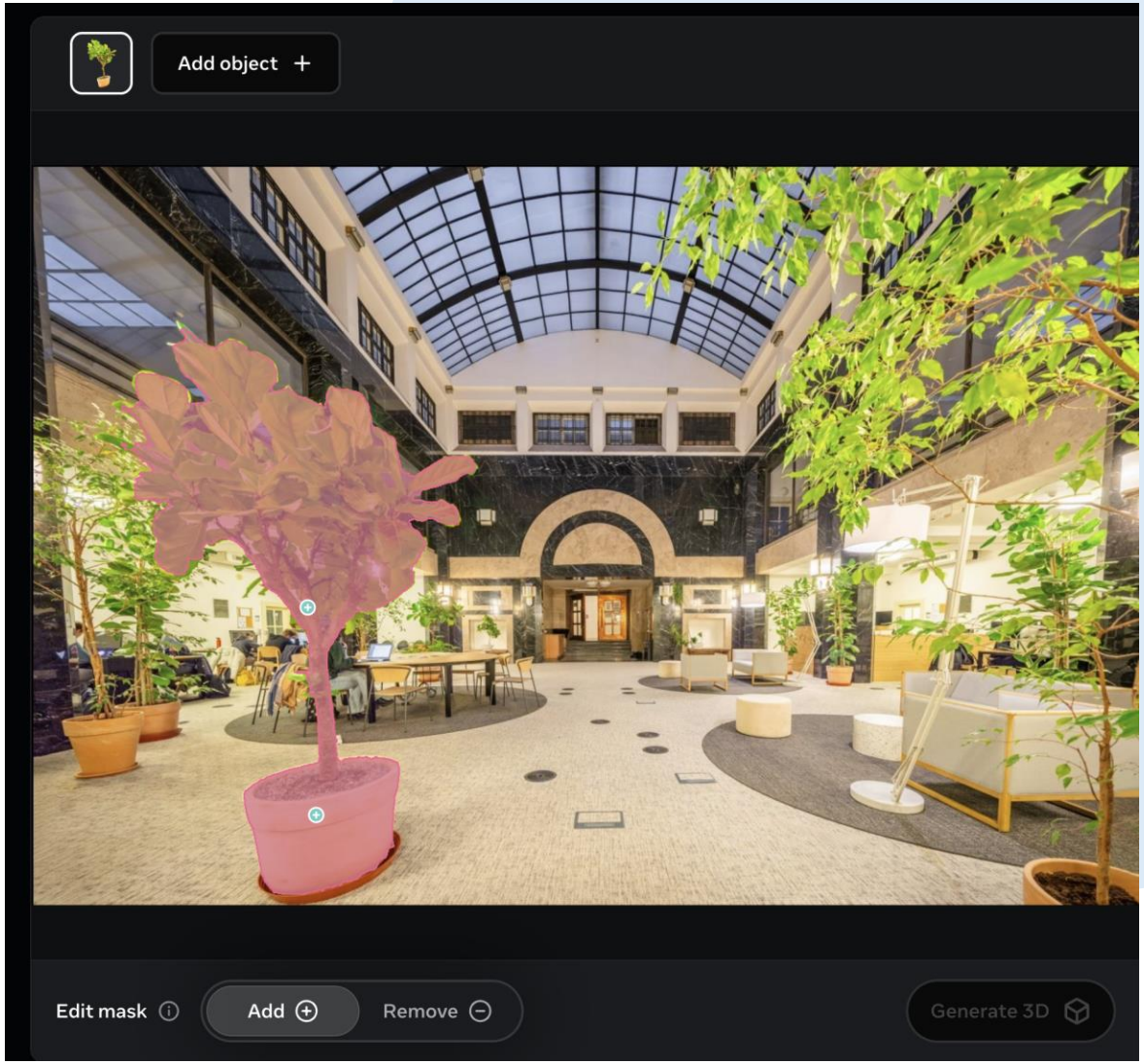
SAM 3 is Detector + Tracker + Multimodal encoder.
It has Vision + Text encoder (joint embedding)
It has DETR-style detector
It has Memory-based tracker (basically taken from SAM 2)

Kirillov et al., 2023 (SAM)

SAM3D

Extends SAM to 3D reconstruction. It takes single image as input (or video) and outputs either 3D scene or object. Works for both human and random objects.

SAM3D is capable of monocular 3D reconstruction given geometry + segmentation combined.



SAM3D demonstration from aidemos.meta.com. MFF UK Rotunda as example.

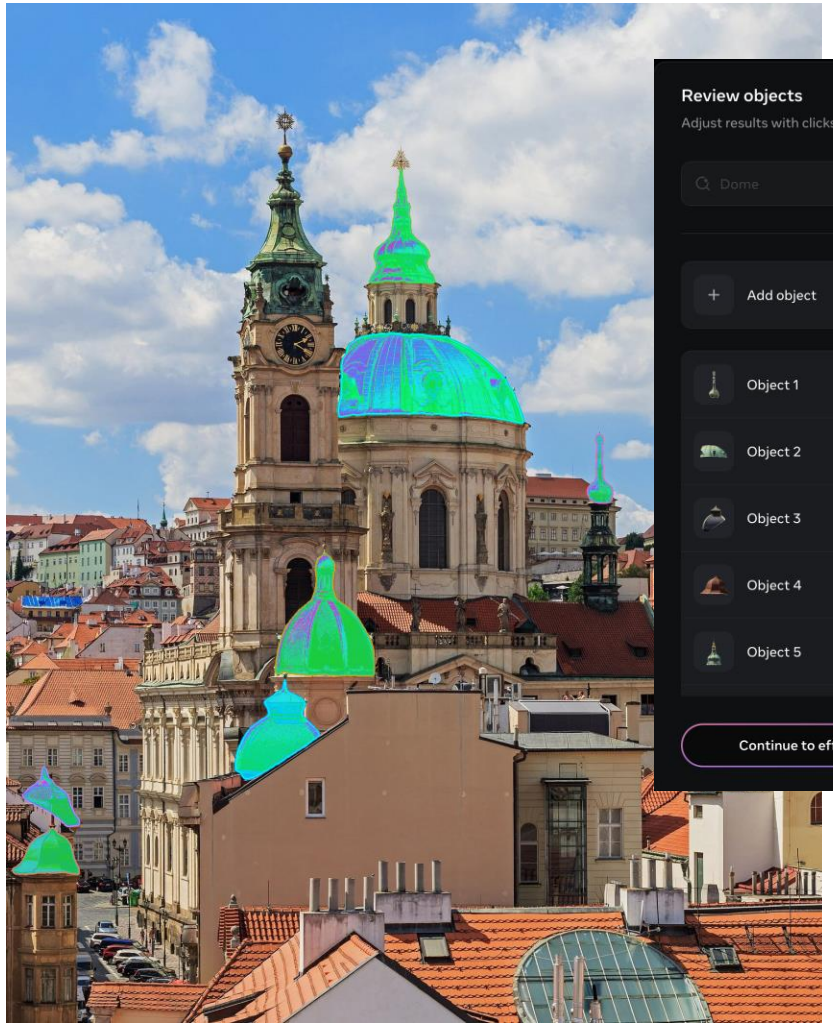
Kirillov et al., 2023 (SAM)

SAM3D

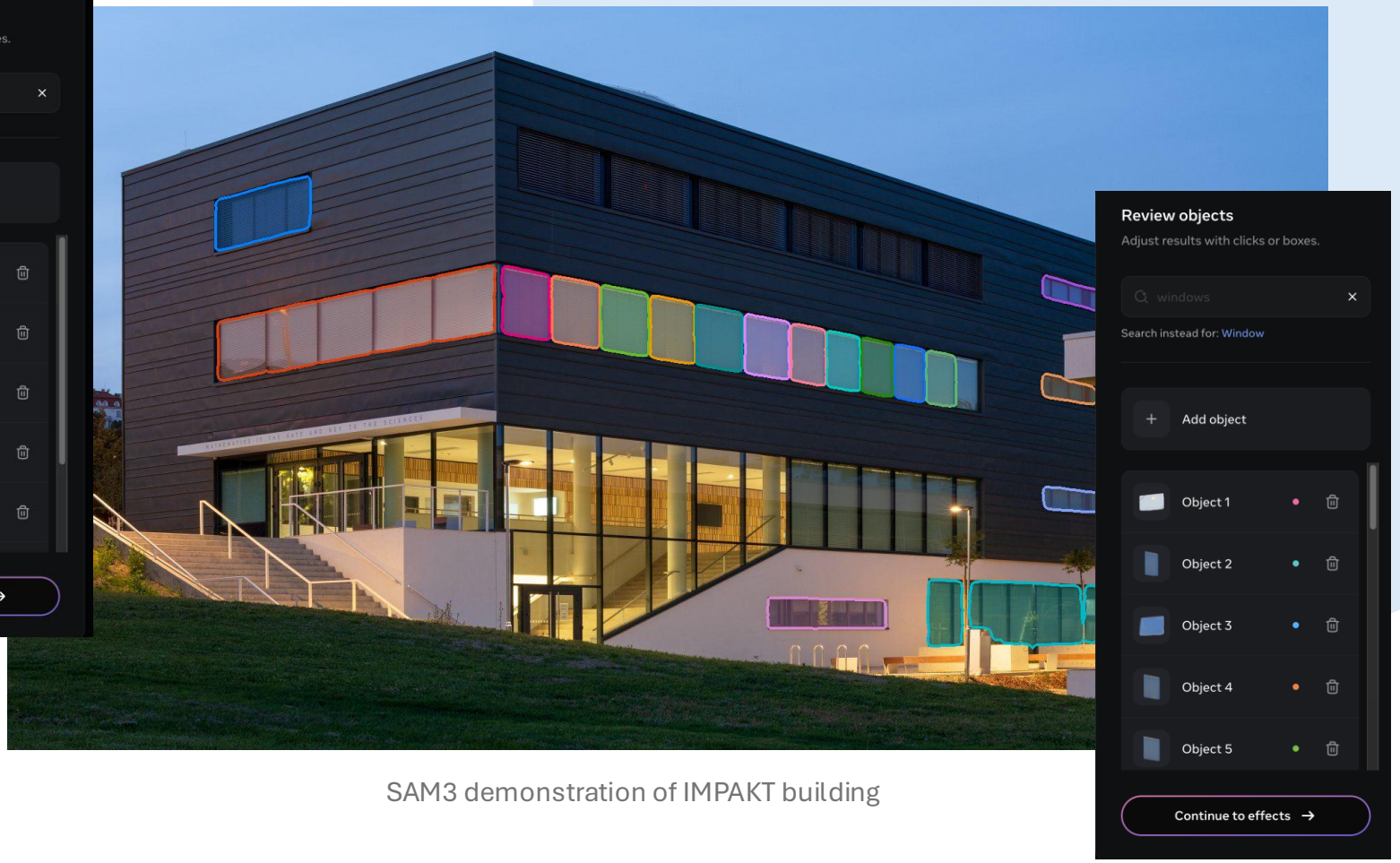
SAM3D demonstration from aidemos.meta.com

Kirillov et al., 2023 (SAM)

THANK YOU!



SAM3 demonstration of Kostel svatého Mikuláše na Malé Straně v Praze



SAM3 demonstration of IMPAKT building

Kirillov et al., 2023 (SAM)