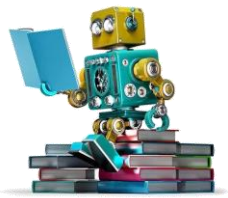
SDML
ML Paper Review

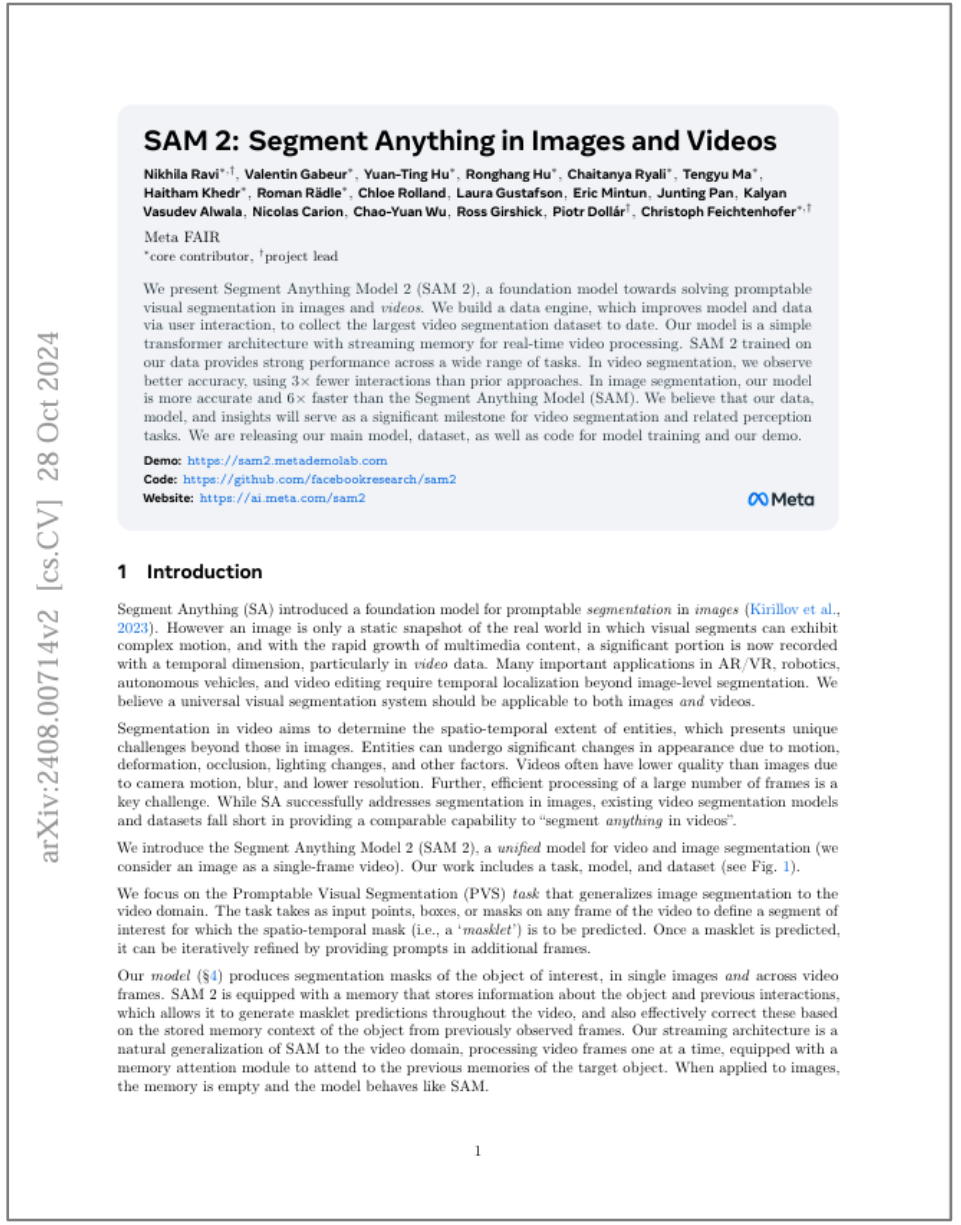
December 2024



SAM 2: Segment Anything in Images and Videos

Nikhila Ravi et al., Meta FAIR

https://arxiv.org/abs/2408.00714



SAM 2: Segment Anything in Images and Videos

Nikhila Ravi*,†, Valentin Gabeur*, Yuan-Ting Hu*, Ronghang Hu*, Chaitanya Ryali*, Tengyu Ma*, Haitham Khedr*, Roman Rädle*, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár†, Christoph Feichtenhofer*,†

Meta FAIR

*core contributor, †project lead

We present Segment Anything Model 2 (SAM 2), a foundation model towards solving promptable visual segmentation in images and *videos*. We build a data engine, which improves model and data via user interaction, to collect the largest video segmentation dataset to date. Our model is a simple transformer architecture with streaming memory for real-time video processing. SAM 2 trained on our data provides strong performance across a wide range of tasks. In video segmentation, we observe better accuracy, using 3× fewer interactions than prior approaches. In image segmentation, our model is more accurate and 6× faster than the Segment Anything Model (SAM). We believe that our data, model, and insights will serve as a significant milestone for video segmentation and related perception tasks. We are releasing our main model, dataset, as well as code for model training and our demo.

Demo: https://sam2.metademolab.com
Code: https://github.com/facebookresearch/sam2
Website: https://ai.meta.com/sam2

Meta

1 Introduction

Segment Anything (SA) introduced a foundation model for promptable *segmentation* in *images* (Kirillov et al., 2023). However an image is only a static snapshot of the real world in which visual segments can exhibit complex motion, and with the rapid growth of multimedia content, a significant portion is now recorded with a temporal dimension, particularly in *video* data. Many important applications in AR/VR, robotics, autonomous vehicles, and video editing require temporal localization beyond image-level segmentation. We believe a universal visual segmentation system should be applicable to both images *and* videos.

Segmentation in video aims to determine the spatio-temporal extent of entities, which presents unique challenges beyond those in images. Entities can undergo significant changes in appearance due to motion, deformation, occlusion, lighting changes, and other factors. Videos often have lower quality than images due to camera motion, blur, and lower resolution. Further, efficient processing of a large number of frames is a key challenge. While SA successfully addresses segmentation in images, existing video segmentation models and datasets fall short in providing a comparable capability to "segment *anything* in videos".

We introduce the Segment Anything Model 2 (SAM 2), a *unified* model for video and image segmentation (we consider an image as a single-frame video). Our work includes a task, model, and dataset (see Fig. 1).

We focus on the Promptable Visual Segmentation (PVS) *task* that generalizes image segmentation to the video domain. The task takes as input points, boxes, or masks on any frame of the video to define a segment of interest for which the spatio-temporal mask (i.e., a '*masklet*') is to be predicted. Once a masklet is predicted, it can be iteratively refined by providing prompts in additional frames.

Our *model* (§4) produces segmentation masks of the object of interest, in single images *and* across video frames. SAM 2 is equipped with a memory that stores information about the object and previous interactions, which allows it to generate masklet predictions throughout the video, and also effectively correct these based on the stored memory context of the object from previously observed frames. Our streaming architecture is a natural generalization of SAM to the video domain, processing video frames one at a time, equipped with a memory attention module to attend to the previous memories of the target object. When applied to images, the memory is empty and the model behaves like SAM.

SAM2 overview

- Image segmentation traditionally involves defining classes of interest
  - Goal is to identify all pixels associated with these classes
  - Every other pixel is considered background
- Segment Anything (2023) goal was to create a foundation model that was taught how to segment everything in images into its parts
  - None of the parts are classified or named
  - Teaching the model generalized patterns for object boundaries
- SAM2 keeps most of SAM's architecture, runs faster, and now can segment objects in videos
  - Introduces a "memory" of other frames that the current frame can attend to

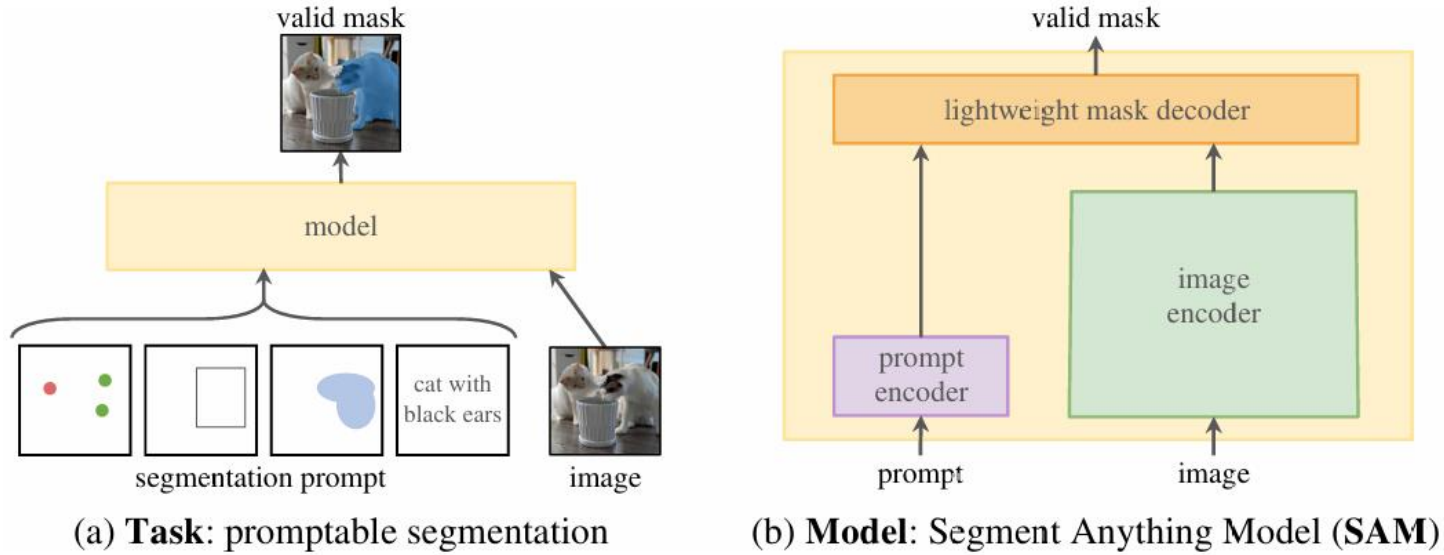
Classic segmentation vs. SAM

Instance segmentation for person and dog classes has found 2 objects

SAM segmenting image without any instructions about classes can find hundreds of objects

Segment Anything approach

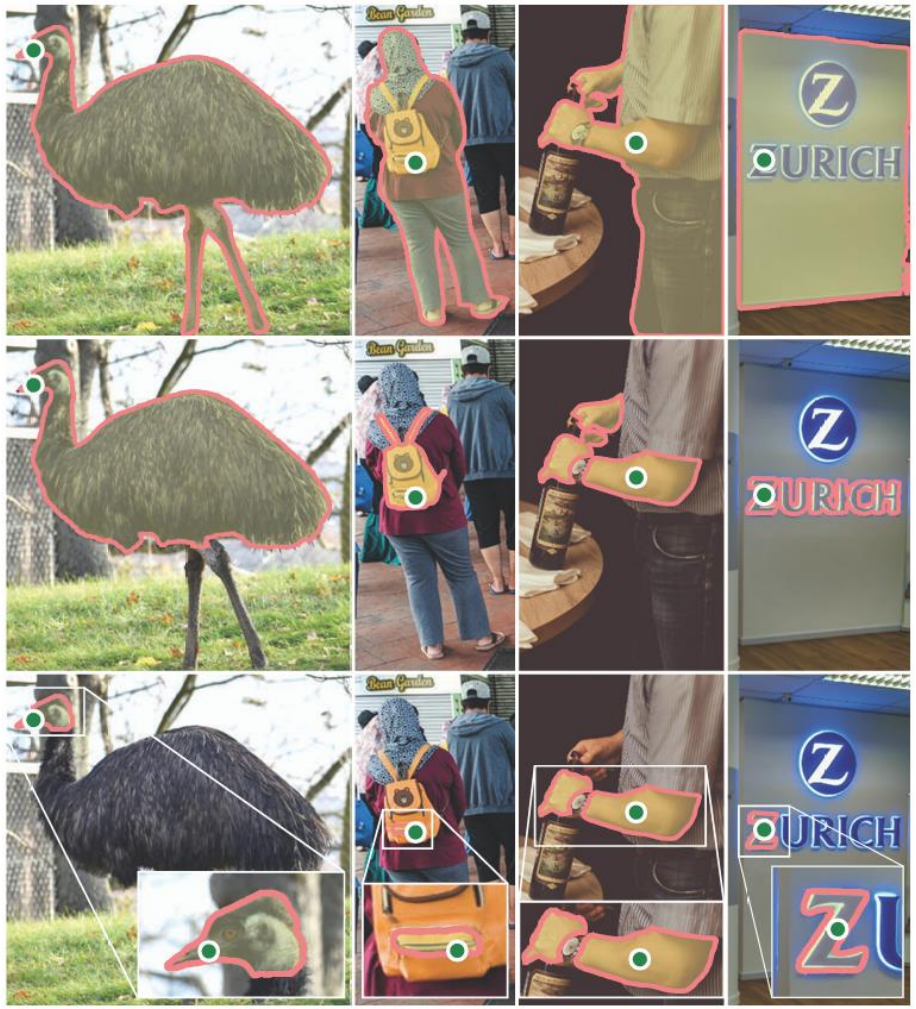
- Segment Anything had concept of prompts and images
- Prompt encoder would accept points, rectangles, masks, and text
- Image encoder would create an image embedding
- Mask decoder would output pixel masks based on image and prompt



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**)

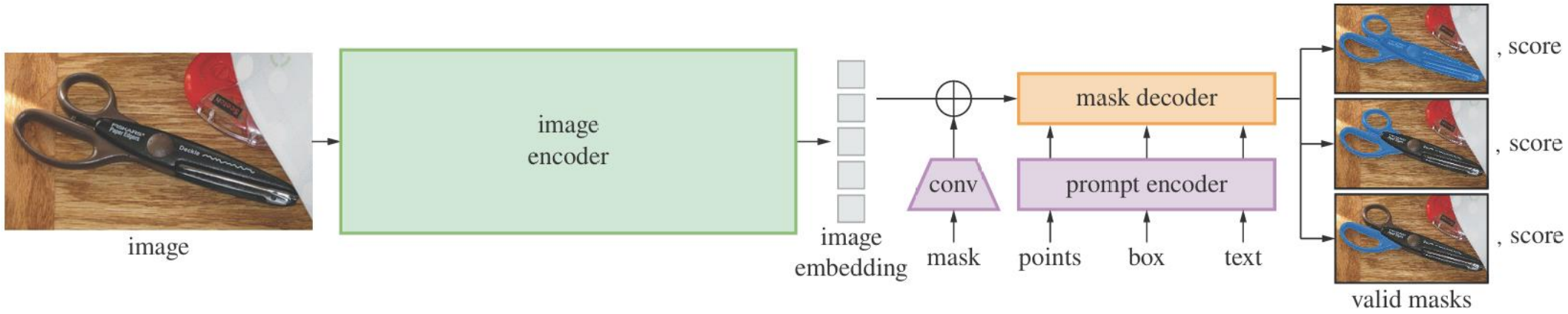
Addressing ambiguity

- With a point prompt, there is ambiguity what size object the user wants
- Decided that three levels were enough
  - Represents the whole, part, and subpart
- Decoder outputs 3 masks per prompt
  - Model predicts estimated IoU for each mask
  - Only the minimum loss (of the three) is backpropagated to update weights



Segment Anything architecture

- Image encoder is an MAE pre-trained ViT with minor changes
  - This is the most expensive component and runs only once per image
- Prompt encoder learns an embedding for prompt type and uses positional encodings for “sparse” prompts. Uses CNN for masks.
- Mask decoder is 2 transformer layers, an MLP, and prediction head
  - Layers have prompt self-attn, prompt-image, and image-prompt cross-attn

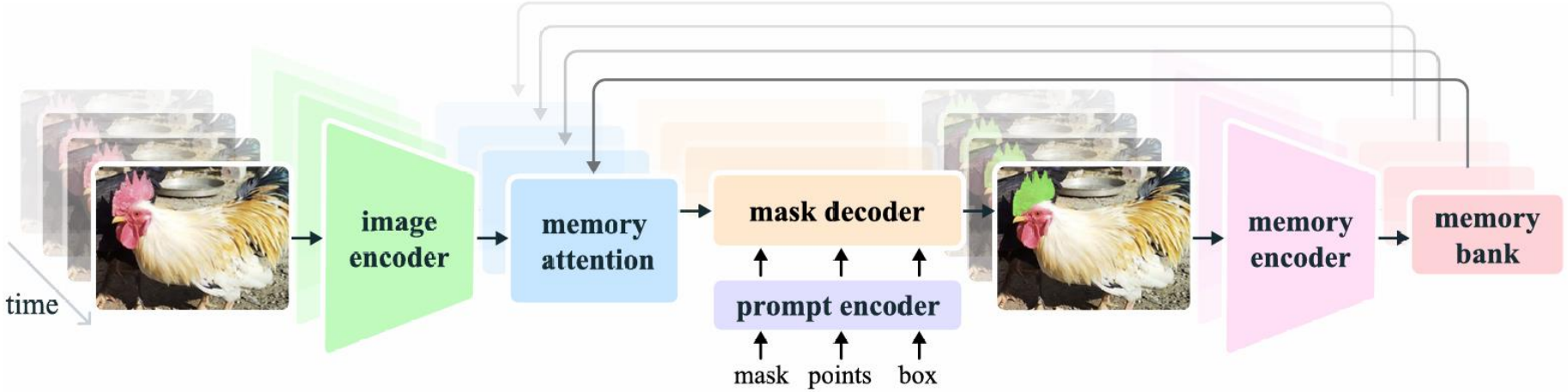


SAM2 architecture changes [1]

- Image encoder is an MAE pre-trained Hiera instead of ViT
  - Gives better multiscale features
- New *memory attention* starts with image self-attn and has multiple layers of cross-attn with memory bank information (see below)
- Prompt encoder is the same except eliminated support for text
- Mask decoder is largely the same design
  - Gathers input directly from image encoder as well as from memory attention
  - Additional binary head predicts whether object of interest is visible in the current frame (it may be occluded)

SAM2 architecture changes [2]

- New memory encoder downsamples the output mask, sums it with the image embedding, and runs “light-weight” convolutional layers
- New memory bank stores other frames, prompts, and predictions
  - A queue is for *N* recent frames of spatial info & temporal position embedding
  - Another queue is for *M* prompted frames of spatial info
  - Also stores a list of *object pointer* vectors based on frame predictions



Data and training

- Both SAM and SAM2 broke new ground, so training sets didn’t exist
- Both papers discuss an iterative, active learning process was used to build data and train the model
- First phase was small and slow with lots of manual annotation, and the initial model was trained
- Next phase would use the model from the previous training to pseudo-label data, making the annotation faster
  - The higher volume and higher quality of data made the next model better
- This repeated until the full-featured model was highly accurate

Additional details

- In order to segment all objects, SAM2 lays down a grid of points, segments each one, then deduplicates results
- You can specify multiple points per object, both positive and negative
- Masks do not need to be contiguous
- In SAM2, video frames are processed in consecutive order
  - Both forward and backwards, so an object detected in frame 30 informs frame 29 and earlier frames
- Much research has examined/extended SAM/SAM2 capabilities
  - SAMURAI does better tracking of fast-moving objects with a Kalman filter and modification to the memory bank

SAM2 conclusion

- SAM and SAM2 have dramatically altered image segmentation
- You can do good zero-shot segmentation of everyday images without any fine-tuning, but have identified various limitations
  - It hasn't been successful identifying healthy versus diseased tissue
  - It pays attention to color, but it's not just splitting regions by color
  - There are parameters which allow you to tune whether you tend to get big masks or lots of little masks
- The sizes range from 38.9M to 224.4M parameters
- It's an important model to know about
  - Unsurprisingly, it's pretty much just some attention layers and a lot of data

References

- Segment Anything
  Alexander Kirillov et al. (2023)
  https://arxiv.org/abs/2304.02643
- SAM 2 webpage – https://ai.meta.com/sam2/
- SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory
  Cheng-Yen Yang et al. (2024)
  https://arxiv.org/abs/2411.11922