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

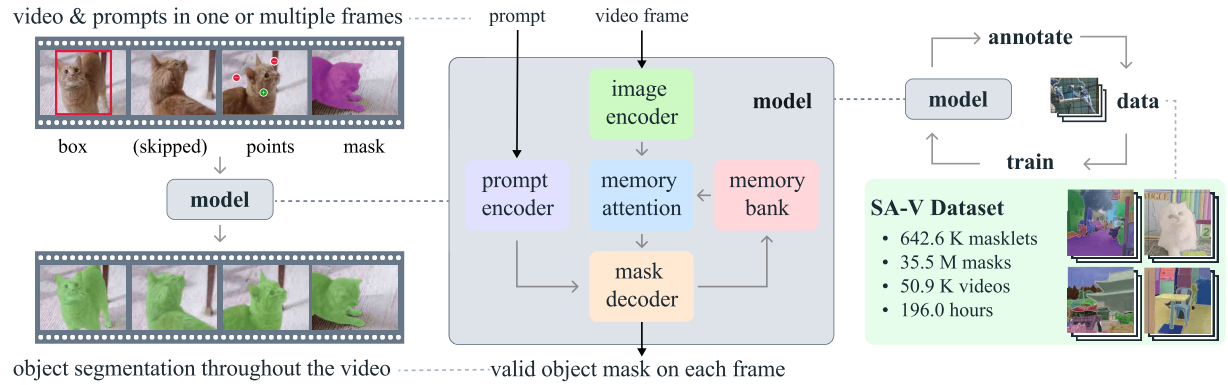
Segment Anything (SA) introduced a foundation model for promptable *segmentation* in *images* (Kirillov et al., 2023). However an image is only a static snapshot of the real world in which visual segments can exhibit complex motion, and with the rapid growth of multimedia content, a significant portion 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.



(a) Task: promptable visual segmentation **(b) Model:** Segment Anything Model 2 **(c) Data:** data engine and dataset

Figure 1 We introduce the Segment Anything Model 2 (SAM 2), towards solving the promptable visual segmentation task (a) with our foundation model (b), trained on our large-scale SA-V dataset collected through our data engine (c). SAM 2 is capable of interactively segmenting regions through prompts (clicks, boxes, or masks) on one or multiple video frames by utilizing a streaming memory that stores previous prompts and predictions.

We employ a *data engine* (§5) to generate training data by using our model in the loop with annotators to interactively annotate new and challenging data. Different from most existing video segmentation datasets, our data engine is not restricted to objects of specific categories, but instead targeted to provide training data for segmenting *any* object with a valid boundary, including parts and subparts. Compared to existing model-assisted approaches, our data engine with SAM 2 in the loop is 8.4× faster at comparable quality. Our final Segment Anything Video (SA-V) dataset (§5.2) consists of 35.5M masks across 50.9K videos, 53× more masks than any existing video segmentation dataset. SA-V is challenging with small objects and parts that get occluded and re-appear throughout the video. Our SA-V dataset is geographically diverse, and a fairness evaluation of SAM 2 indicates minimal performance discrepancy in video segmentation based on perceived gender, and little variance among the three perceived age groups we evaluated.

Our experiments (§6) show that SAM 2 delivers a step-change in the *video* segmentation experience. SAM 2 can produce *better* segmentation accuracy while using 3× *fewer* interactions than prior approaches. Further, SAM 2 outperforms prior work in established *video* object segmentation benchmarks, under multiple evaluation settings, *and* delivers better performance compared to SAM on *image* segmentation benchmarks, while being 6× faster. SAM 2 is shown to be effective across a variety of video and image distributions as observed through numerous zero-shot benchmarks including 17 for video segmentation and 37 for single-image segmentation.

We are releasing our work under permissive open licences, including the SA-V dataset (CC by 4.0), the SAM 2 model checkpoints[1], training code (Apache 2.0), and code for our interactive online demo (Apache 2.0).

2 Related work

Image segmentation. Segment Anything (Kirillov et al., 2023) introduces a promptable image segmentation task where the goal is to output a valid segmentation mask given an input prompt such as a bounding box or a point that refers to the object of interest. SAM trained on the SA-1B dataset allows for zero-shot segmentation which enabled its adoption to a wide range of applications. Recent work has extended SAM, e.g., by introducing a High-Quality output token to train on fine-grained masks (Ke et al., 2024), or improve SAM's efficiency (Xiong et al., 2023; Zhang et al., 2023a; Zhao et al., 2023). More broadly, SAM is used in a wide range of applications, including medical imaging (Ma et al., 2024; Deng et al., 2023; Mazurowski et al., 2023; Wu et al., 2023a), remote sensing (Chen et al., 2024; Ren et al., 2024), motion segmentation (Xie et al., 2024), and camouflaged object detection (Tang et al., 2023).

[1]All the results presented in this paper are based on a new version of SAM 2 (improved over our initial release; denoted as "SAM 2.1" in https://github.com/facebookresearch/sam2), which we will refer to as SAM 2 throughout for brevity.