AU24 CSE 6249 Presentation

Meta

SAM 2: Segment Anything in Images and Videos

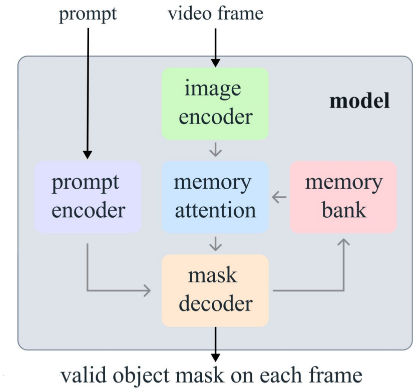
Nikhila Ravi et. al.

Demo: sam2.metademolab.com
Code: github.com/facebookresearch/segment-anything-2
Website: ai.meta.com/sam2

Presented by **Ram Sai Ganesh**

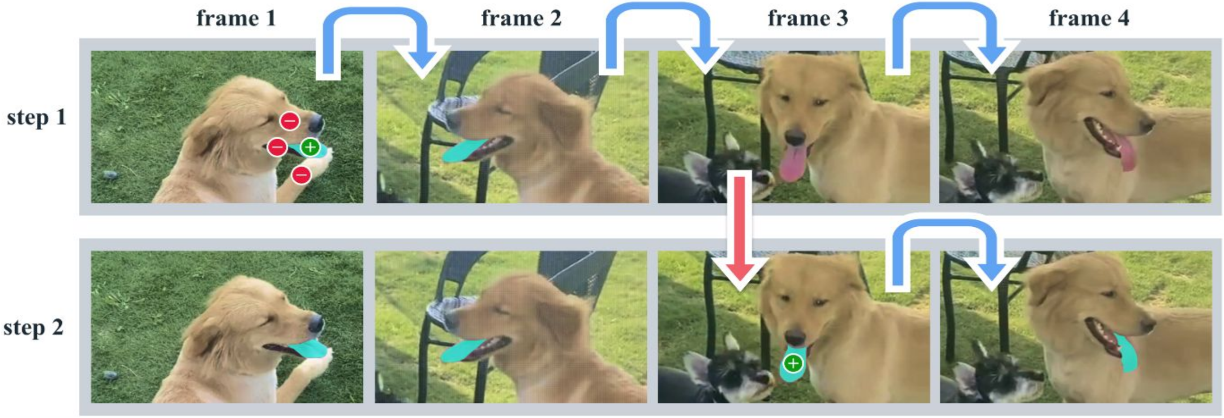
Introduction

- SAM (2023) – Foundation model for promptable semantic *image* segmentation.
  - Many applications have *temporal* dimension.
- Challenges –
  - Entities change drastically in appearance; fast motion; lower resolution.
- Solution –
  - Model which produces *masklets* by conditioning on *stored object memory*.
- Contributions – **SAM 2** & **SA-V** dataset



Task: **P**romptable **V**isual **S**egmentation (**PVS**)

- Input – points, bounding boxes, or masks – on *any* frame of a video
    - Define a segment of interest
- Output – *Masklets* (ie. one or more series of masks per-frame)
- SAM's single-frame image segmentation ⊂ PVS



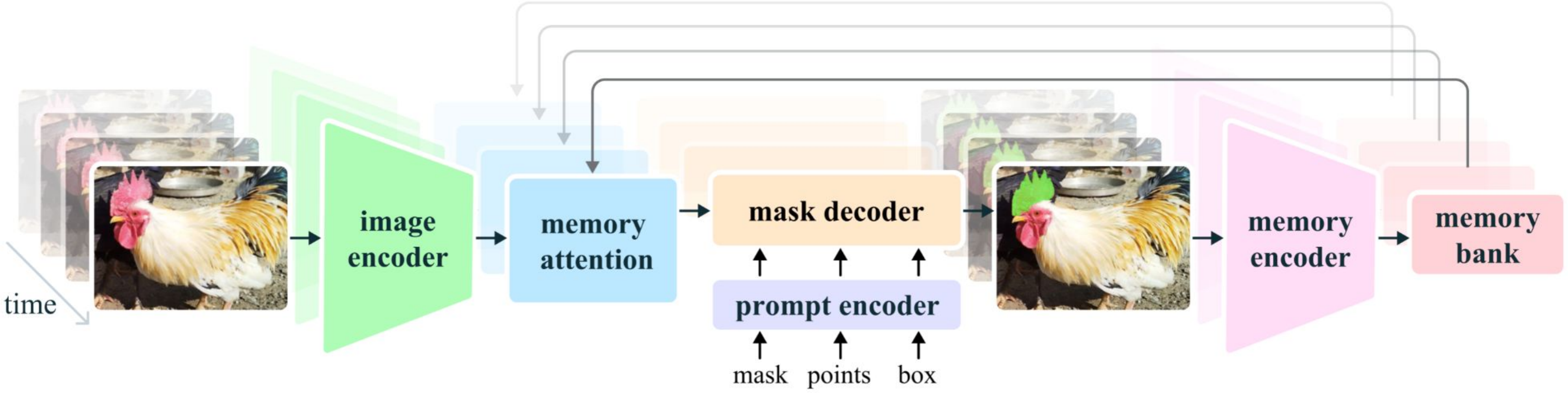
Related Work

- Image Segmentation
  - SAM trained on SA-1B & its adaptations for specific tasks.
- Interactive Video Object Segmentation (iVOS)
  - VOS supervisory signal (clicks/scribbles) – notably the DAVIS benchmark.
- Semi-supervised VOS
  - Automatically propagate initial supervised mask through entire video.
- Video Segmentation Datasets
  - Quality annotations at scale – scarce until recently.
  - Many challenge-specific datasets.

SAM 2 Model

Architecture & Overview

- Frames processed one-at-a-time; "segmentation prediction is conditioned on the current prompt…and cross-attended to memories of the target object."
- Mask decoder predicts frame's segmentation mask(s).
- Memory encoder saves prediction + image embeddings for use later.



SAM 2 Model

Input & Prompts

- Image Encoder –
  - Pre-trained hierarchical MAE.
  - Run once per frame; provides image embeddings for masks & memory.
  - 4 sizes – T, S, B+ & L.
- Prompt Encoder –
  - Identical to SAM
  - Prompted by +/- clicks, bounding boxes, or masks.
  - Can support text-prompting with CLIP.

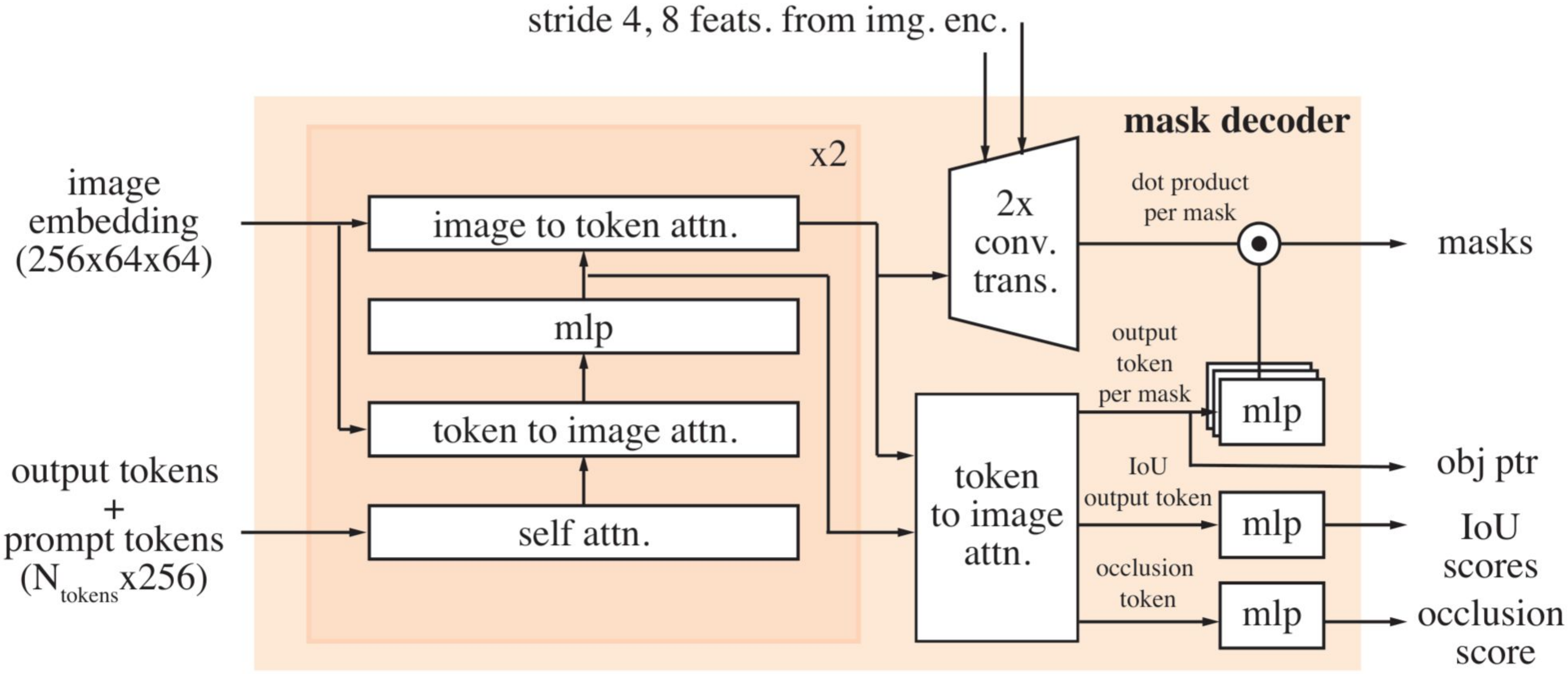
SAM 2 Model

- Largely follows SAM.
- “We stack two-way transformer blocks that update prompt and frame embeddings.”
- Handles ambiguity by predicting multiple masks.
  - Video ambiguity can be across frames.
  - MLP predicts IoU score; retain only highest scoring mask for subsequent frames.
- Handles occlusion in video –
  - Another MLP predicts object presence.

Mask Decoder

SAM 2 Model

Mask Decoder

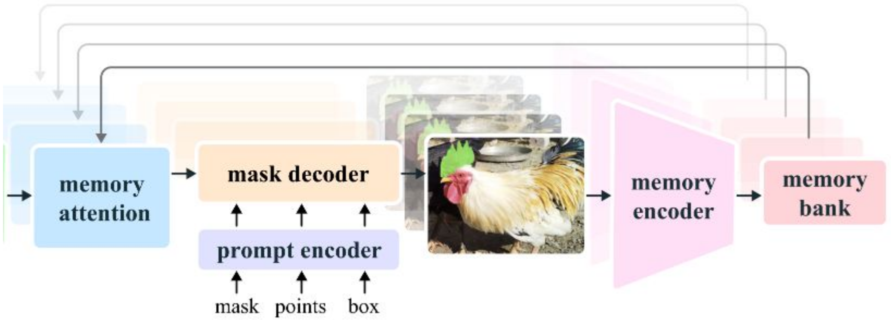


SAM 2 Model

Temporal Memory

- **Memory Encoder** given a frame + mask, generates per-frame 'memory' –
  - Downsamples output mask & fuses with *unconditioned* frame embedding.

- **Memory Bank** maintains a queue –
  - Info about mask predictions for past *N* frames.
  - Info about prompts for past *M* frames, as a *spatial* feature map.



- Bank also stores semantic information about prompted object.
  - Used for cross-attention along with spatial memory features.

SAM 2 Model

- Pre-trained on SA-1B
- Jointly trained on image and video data:
    1. Sample 8-frame sequences.
    2. Randomly select 2 frames for ‘corrective’ prompts using ground-truth.
    3. Task: sequentially and ‘interactively’ predict masklets.
- Randomly:
    - Reverse temporal order (50%)
    - Additional corrective clicks (10%)

Training

Data

Data Engine

- 4.2M frames, 196 hours of video @ 240p-4K resolution.
- Data engine:
  - *Phase 1:* Per-frame SAM –
    - Manually pixel-refining SAM annotations (37.8 s/frame.)
  - *Phase 2:* SAM + SAM 2 Mask –
    - SAM generates prompted masks that SAM 2 propagates (7.4 s/frame.)
  - *Phase 3:* SAM 2–
    - SAM 2 accepts any prompt, annotators may refine masks (4.5s/frame.)
- Overall – 8.4x speedup from Phase 1; 276K+ masks collected in 3 phases.

Data

Quality Evaluation & Analysis

- Data Engine efficacy – # of edited frames is a proxy for the “challengingness” of an object’s segmentation.
- Quality verification – independent set of annotators tasked with labeling masklets un/satisfactory; poor data re-annotated with the data engine.

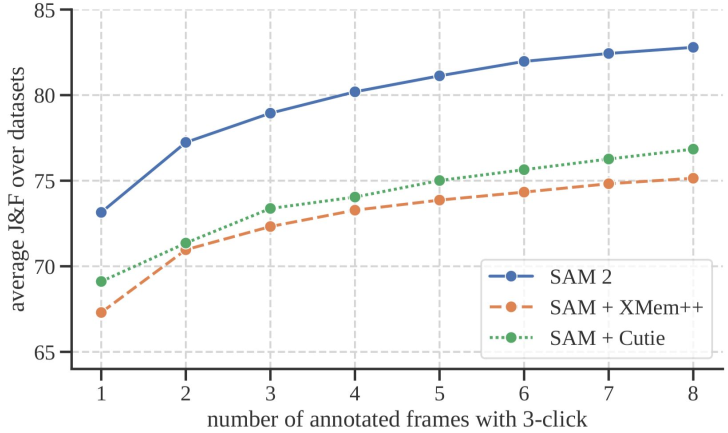
Training data	SA-V val	9 zero-shot
VOS + SA-1B	50.0	62.5
+ Phase 1	53.0	66.9
+ Phase 2	58.8	70.9
+ Phase 3	62.5	71.2
+ Auto	**63.2**	**71.5**

	Model in the Loop	Time per Frame	Edited Frames	Clicks per Clicked Frame	Phase 1 Mask Alignment Score (IoU>0.75)			
					All	Small	Medium	Large
Phase 1	SAM only	37.8 s	100.00 %	4.80	-	-	-	-
Phase 2	SAM + SAM 2 Mask	7.4 s	23.25 %	3.61	86.4 %	71.3 %	80.4 %	97.9 %
Phase 3	**SAM 2**	**4.5 s**	**19.04 %**	**2.68**	**89.1 %**	**72.8 %**	**81.8 %**	**100.0 %**

Zero-shot Experiments

PVS, VOS, Images

- Promptable Video Segmentation (PVS) –
  - SAM 2 *J&F Metric* outperforms competing baselines (XMem++ & Cutie), with >3x fewer prompts
- Semi-supervised Video Object Segmentation –
  - SAM 2 outperforms baselines on 17 datasets.
  - Excels at the conventional non-interactive VOS task.



Method	3-click	bounding box	ground-truth mask‡
SAM+XMem++	68.4	67.6	72.7
SAM+Cutie	70.1	69.4	74.1
SAM 2	**73.2**	**72.9**	**77.6**

Experiments

vs. VOS SoTA

- Primary focus is PVS, but SAM-2 achieves SoTA in VOS as well:

Method	$\mathcal{J}\&\mathcal{F}$ MOSE val	$\mathcal{J}\&\mathcal{F}$ DAVIS 2017 val	$\mathcal{J}\&\mathcal{F}$ LVOS val	$\mathcal{J}\&\mathcal{F}$ SA-V val	$\mathcal{J}\&\mathcal{F}$ SA-V test	$\mathcal{G}$ YTVOS 2019 val	FPS
STCN (Cheng et al., 2021a)	52.5	85.4	-	61.0	62.5	82.7	13.2
SwinB-AOT (Yang et al., 2021b)	59.4	85.4	-	51.1	50.3	84.5	-
SwinB-DeAOT (Yang & Yang, 2022)	59.9	86.2	-	61.4	61.8	86.1	-
RDE (Li et al., 2022a)	46.8	84.2	-	51.8	53.9	81.9	24.4
XMem (Cheng & Schwing, 2022)	59.6	86.0	-	60.1	62.3	85.6	22.6
SimVOS-B (Wu et al., 2023b)	-	88.0	-	44.2	44.1	84.2	3.3
JointFormer (Zhang et al., 2023b)	-	90.1	-	-	-	87.4	3.0
ISVOS (Wang et al., 2022)	-	88.2	-	-	-	86.3	5.8
DEVA (Cheng et al., 2023b)	66.0	87.0	55.9	55.4	56.2	85.4	25.3
Cutie-base (Cheng et al., 2023a)	69.9	87.9	66.0	60.7	62.7	87.0	36.4
Cutie-base+ (Cheng et al., 2023a)	71.7	88.1	-	61.3	62.8	87.5	17.9
SAM 2 (Hiera-B+)	75.8	90.9	74.9	73.6	74.1	88.4	**43.8**
SAM 2 (Hiera-L)	**77.2**	**91.6**	**76.1**	**75.6**	**77.6**	**89.1**	30.2

Ablations

Dataset Mix

- Fixed hyperparameters, varying only training data –
  - Observe that a pure-VOS model generalizes poorly.

	Training data				$\mathcal{J}\&\mathcal{F}$				mIoU
	VOS	Internal	SA-V	SA-1B	SA-V val	Internal-test	MOSE dev	9 zero-shot	SA-23
1	✓				48.1	60.2	76.9	59.7	45.4
2		✓			57.0	72.2	70.6	70.0	54.4
3			✓		63.0	72.6	72.8	69.7	53.0
4			✓	✓	62.9	73.2	73.6	69.7	<u>58.6</u>
5		✓	✓		63.0	73.2	73.3	70.9	55.8
6		✓	✓	✓	**63.6**	**75.0**	74.4	<u>71.6</u>	<u>58.6</u>
7	✓			✓	50.0	63.2	77.6	62.5	54.8
8	✓	✓			54.9	71.5	77.9	70.6	55.1
9	✓		✓		61.6	72.8	78.3	69.9	51.0
10	✓		✓	✓	62.2	74.1	<u>78.5</u>	70.3	57.3
11	✓	✓	✓		61.8	<u>74.4</u>	<u>78.5</u>	**71.8**	55.7
12	✓	✓	✓	✓	<u>63.1</u>	73.7	**79.0**	<u>71.6</u>	**58.9**

Ablations

SAM 2 Capacity

res.	$\mathcal{J}\&\mathcal{F}$ MOSE dev	SA-V val	9 zero-shot	FPS	mIoU SA-23
512	73.0	68.3	70.7	**77.3**	59.7
768	76.1	**71.1**	**72.5**	62.5	61.0
1024	**77.0**	70.1	72.3	44.6	**61.5**

(a) Resolution.

#frames	$\mathcal{J}\&\mathcal{F}$ MOSE dev	SA-V val	9 zero-shot	FPS	mIoU SA-23
4	71.1	60.0	67.7	**77.3**	**60.1**
8	73.0	**68.3**	70.7	**77.3**	59.7
10	**74.5**	68.1	**71.1**	**77.3**	59.9

(b) #Frames.

#mem.	$\mathcal{J}\&\mathcal{F}$ MOSE dev	SA-V val	9 zero-shot	FPS	mIoU SA-23
4	**73.5**	68.6	70.5	**77.4**	**59.9**
6	73.0	68.3	**70.7**	77.3	59.7
8	73.2	**69.0**	**70.7**	67.7	**59.9**

(c) #Memories.

chan. dim.	$\mathcal{J}\&\mathcal{F}$ MOSE dev	SA-V val	9 zero-shot	FPS	mIoU SA-23
64	73.0	**68.3**	**70.7**	**77.3**	59.7
256	**73.4**	66.4	70.0	77.0	**60.0**

(d) Memory channels.

(#sa, #ca)	$\mathcal{J}\&\mathcal{F}$ MOSE dev	SA-V val	9 zero-shot	FPS	mIoU SA-23
(2, 2)	**73.3**	67.3	70.2	**85.8**	59.9
(3, 2)	72.7	64.1	69.5	84.2	**60.0**
(4, 4)	73.0	**68.3**	**70.7**	77.3	59.7

(e) Memory attention.

img. enc.	$\mathcal{J}\&\mathcal{F}$ MOSE dev	SA-V val	9 zero-shot	FPS	mIoU SA-23
S	70.9	65.5	69.4	**78.3**	57.8
B+	73.0	**68.3**	70.7	77.3	59.7
L	**75.0**	66.3	**71.9**	62.6	**61.1**

(f) Image encoder size.

Conclusion

- Paper highlights:
    - SAM 2 extends the PVS task to video.
    - SAM 2 works by augmenting SAM architecture to use memory.
    - SA-V dataset for training and benchmarking video segmentation.
- Appendices – significantly more detail about SAM 2 training, architecture & experiments, and the SA-V data engine.
- Connections with current projects –
    - Back-calculating more point prompts could mitigate ambiguity in masks, solves issue evaluating quantization efficacy in SAM.
    - CLIP compatibility useful — applications using segments for KGD/KGI.

Thank you!

Questions?

SAM 2: Segment Anything in Images and Videos

Nikhila Ravi et. al.

Demo:	sam2.metademolab.com
Code:	github.com/facebookresearch/segment-anything-2
Website:	ai.meta.com/sam2