Progress Report

- 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 and Christoph Feichtenhofer. **SAM 2: Segment Anything in Images, Videos**. arXiv:2408.00714, Aug 2024.

Introduction

- A foundation model for segmentation in both images and videos
- Expand the capabilities of the original Segment Anything Model (SAM)
- Designed to handle the dynamic challenges of object segmentation in videos
- Goal: To create a unified model capable of processing both images and videos

Related Work

- **Segment Anything**:
  - Laid the groundwork for promptable image segmentation, allowing users to input bounding boxes, points, or masks to identify objects in static images
- **Interactive Video Object Segmentation (iVOS)**:
  - Focused on obtaining segmentation through interactive inputs like scribbles and clicks, with approaches that optimize the segmentation based on these user inputs
  - Limitations:
    - Tracker may not work for all objects
    - There is no mechanism to interactively refine a model’s mistake

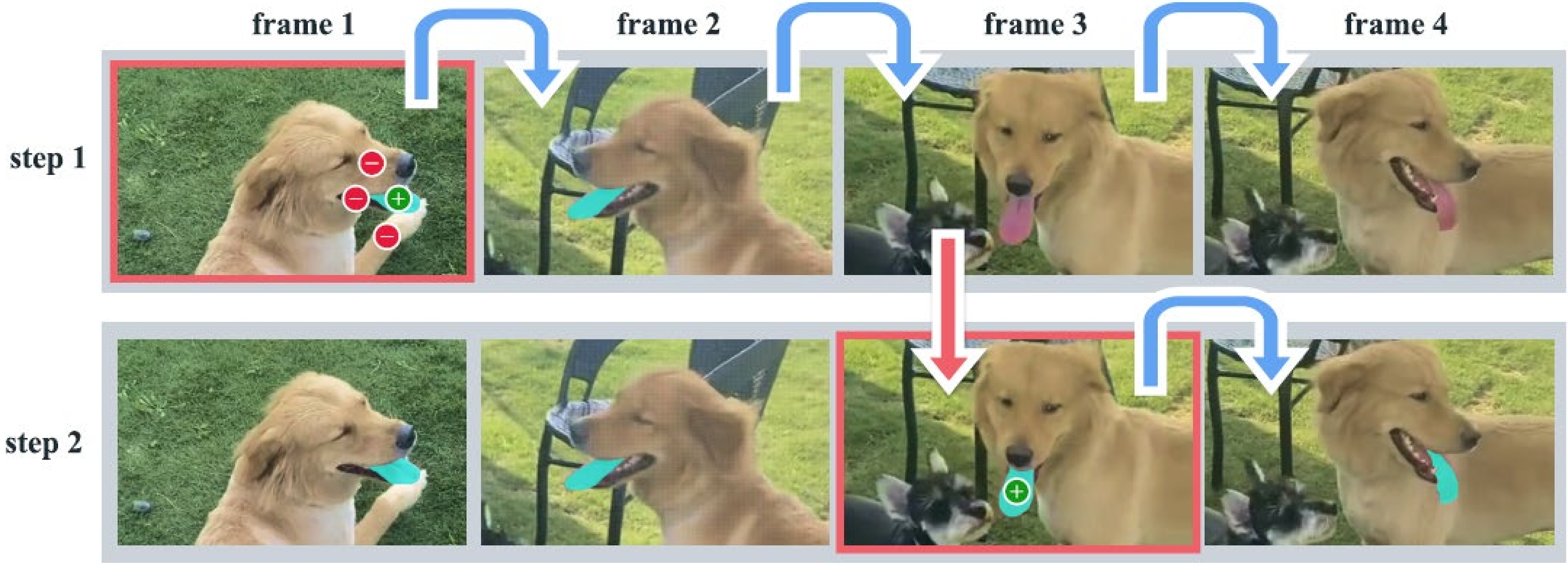
Related Work

- **Semi-supervised Video Object Segmentation (VOS)**:
    - Used a mask on the first frame to track objects through the remaining frames
    - Limitations:
        - Time-consuming in annotating the required high-quality object mask in the first frame
        - Lack sufficient coverage to achieve the capability of segmenting anything in video

Promptable Visual Segmentation (PVS)

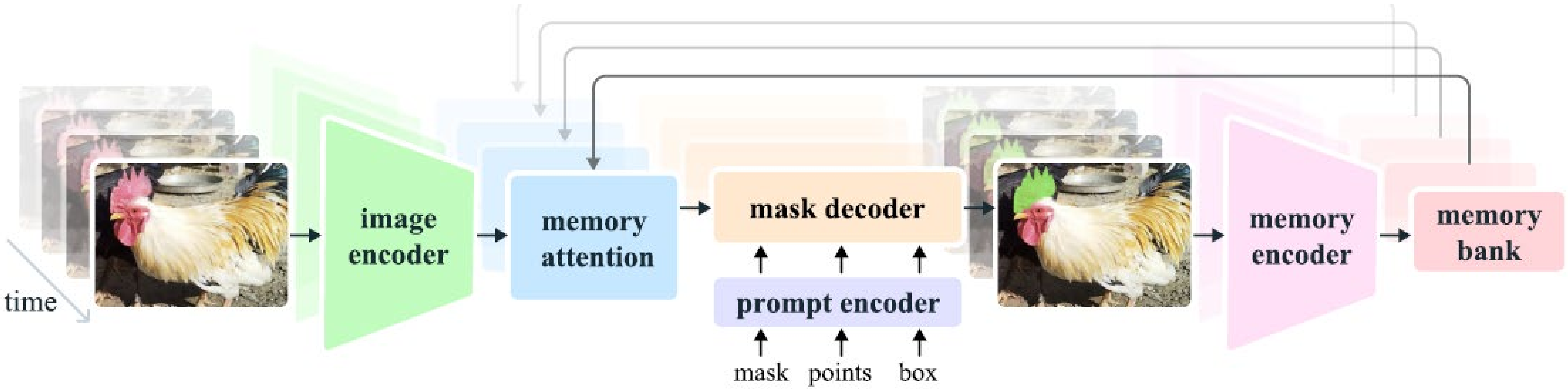
- The PVS task allows providing prompts to the model on any frame of a video
- Upon receiving a prompt on a specific frame, the model should immediately respond with a valid segmentation mask of the object on this frame
- After receiving initial prompts, the model should propagate these prompts to obtain the masklet of the object across the entire video

Promptable Visual Segmentation (PVS)



Model Architecture

- Image Encoder
- Memory Attention
- Prompt Encoder
- Mask Decoder
- Memory Encoder
- Memory Bank



Data Engine

- Data engine went through three phases, each categorized based on the level of model assistance provided to annotators
- Phase 1: **SAM per frame**
    - Annotators a re tasked with annotating the mask of a target object in every frame of the video at 6 frames per second (FPS) using SAM
    - Pixel-precise manual editing tools such as a brush and eraser
    - Collected 16K masklets across 1.4K videos in phase 1
- Phase 2: **SAM+SAM2 Mask**
    - Annotators used SAM and other tools as in Phase 1 to generate spatial masks in the first frame
    - Then use SAM 2 Mask to temporally propagate the annotated mask to other frames to get the full spatio-temporal masklets
    - Collected 63.5K masklets
    - Annotation time went down to 7.4 s/frame, a ~5.1x speed up over Phase 1

Data Engine

- Data engine went through three phases, each categorized based on the level of model assistance provided to annotators
- Phase 1: **SAM per frame**
  - Annotators are tasked with annotating the mask of a target object in every frame of the video at 6 frames per second (FPS) using SAM
  - Pixel-precise manual editing tools such as a brush and eraser
  - Collected 16K masklets across 1.4K videos in phase 1
- Phase 2: **SAM+SAM2 Mask**
  - Annotators used SAM and other tools as in Phase 1 to generate spatial masks in the first frame
  - Then use SAM 2 Mask to temporally propagate the annotated mask to other frames to get the full spatio-temporal masklets
  - Collected 63.5K masklets
  - Annotation time went down to 7.4 s/frame, a ~5.1x speed up over Phase 1

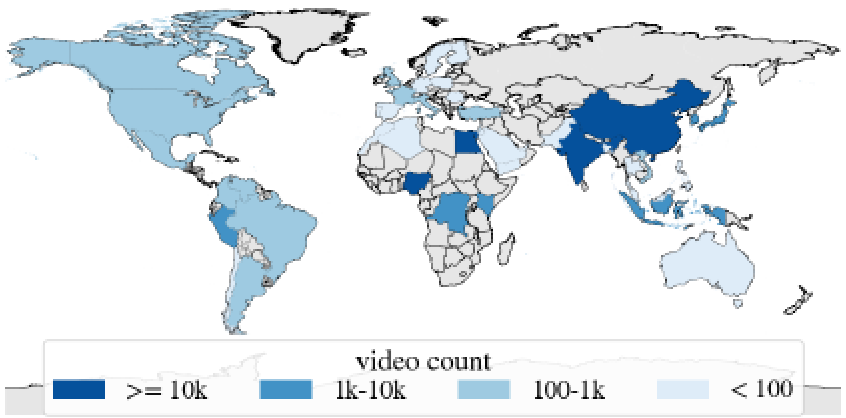
Data Engine

- Phase 3: **SAM 2**
  - Utilize the fully-featured SAM 2
  - Accepts various types of prompts
  - Annotators only need to provide occasional refinement clicks to SAM 2 to edit the predicted masklets in intermediate frames
  - Collected 197.0K masklets
  - Annotation time per frame went down to 4.5 seconds, a ~8.4x speed up over Phase 1

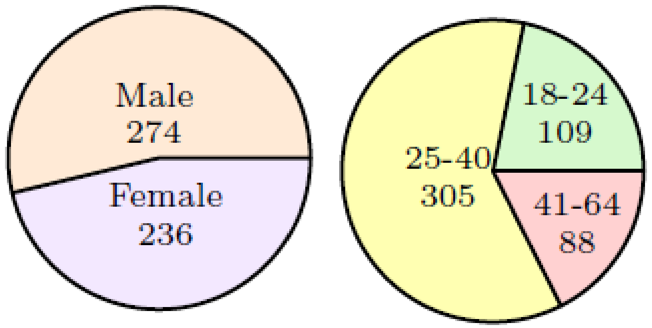
	Model in the Loop	Time per Frame	Edited Frames	Clicks per Clicked Frame	Phase 1 Mask Alignment Score (IoU>0.75)			
					All	Small	Medium	Large
Phase 1	SAM only	37.8 s	100.00 %	4.80	-	-	-	-
Phase 2	SAM + SAM 2 Mask	7.4 s	23.25 %	3.61	86.4 %	71.3 %	80.4 %	97.9 %
Phase 3	**SAM 2**	**4.5 s**	**19.04 %**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

Dataset

- Collected with dsta engine comprises 50.9K videos with 642.6K masklets
- Videos
  - Comprise 54% indoor and 46% outdoor scenes with an average duration of 14 seconds, spanning 47 countries and were capture by diverse participant



(b) Geography

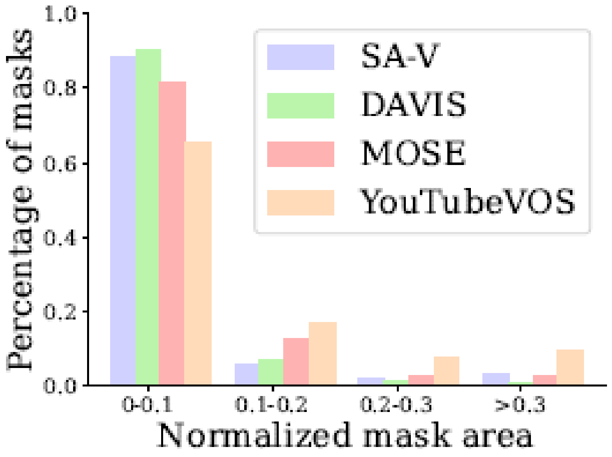


(i) Gender (ii) Age

(c) Crowdworker Demographics

Dataset

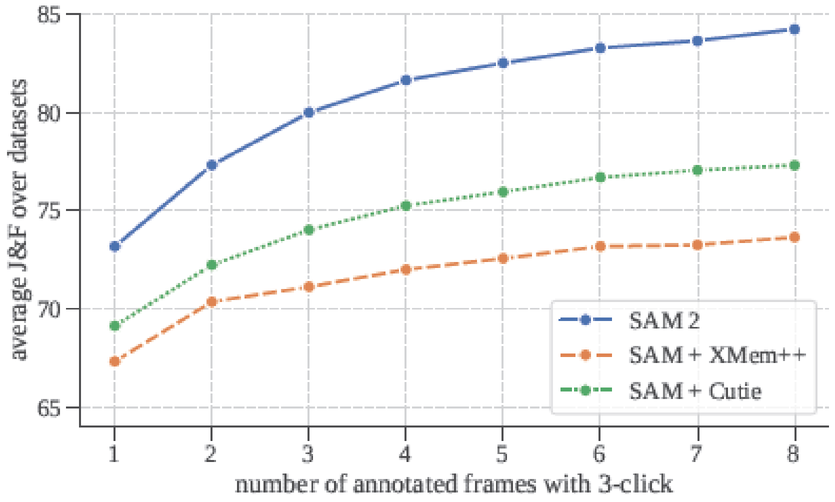
- Masklets
  - The annotations comprise 190.9K manual masklet annotations and 451.7K automatic masklets collected using the data engine
  - SA-V has 53× (15× without auto annotations) more masks than the largest VOS dataset



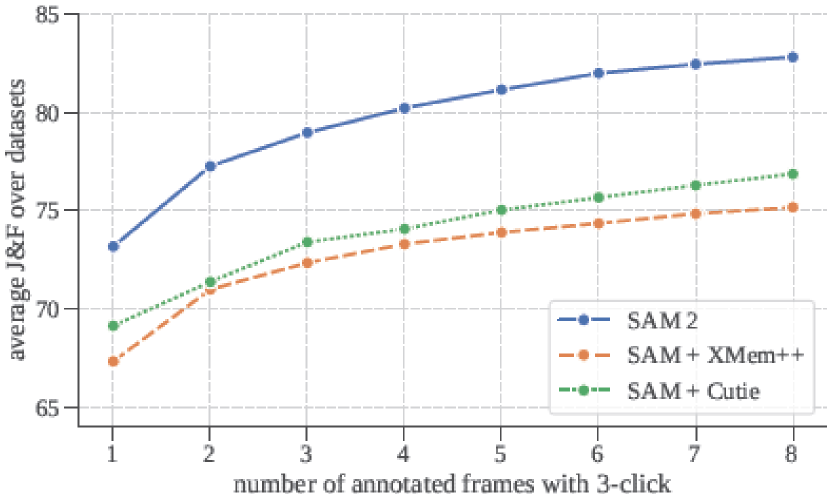
(a) Size

Zero-Shot Experiment

- Compare SAM 2 with previous work on zero-shot video tasks and image tasks
- Evaluate promptable video segmentation with two baseline
  - SAM+XMem++
  - SAM+Cutie



(a) *offline* average $\mathcal{J}\&\mathcal{F}$ across datasets (3-click)



(b) *online* average $\mathcal{J}\&\mathcal{F}$ across datasets (3-click)

Zero-Shot Experiment

- Evaluate the VOS setting with click, box, or mask prompts only on the first frame of the video in semi-supervised video object segmentation

Method	1-click	3-click	5-click	bounding box	ground-truth mask‡
SAM+XMem++	56.9	68.4	70.6	67.6	72.7
SAM+Cutie	56.7	70.1	72.2	69.4	74.1
SAM 2	**64.3**	**73.2**	**75.4**	**72.9**	**77.6**

Zero-Shot Experiment

- Evaluate SAM 2 on the Segment Anything task across 37 zero-shot datasets, including 23 datasets previously used by SAM for evaluation

		1 (5) click mIoU				
Model	Data	SA-23 All	SA-23 Image	SA-23 Video	14 new Video	FPS
SAM	SA-1B	58.1 (81.3)	60.8 (82.1)	54.5 (80.3)	59.1 (83.4)	21.7
SAM 2	SA-1B	58.9 (81.7)	60.8 (82.1)	56.4 (81.2)	56.6 (83.7)	**130.1**
SAM 2	our mix	**61.4 (83.7)**	**63.1 (83.9)**	**59.1 (83.3)**	**69.6 (86.0)**	**130.1**

The average mIoU of 1-click and 5-click by dataset domain and model speed in frames per second (FPS) on a single A100 GPU.

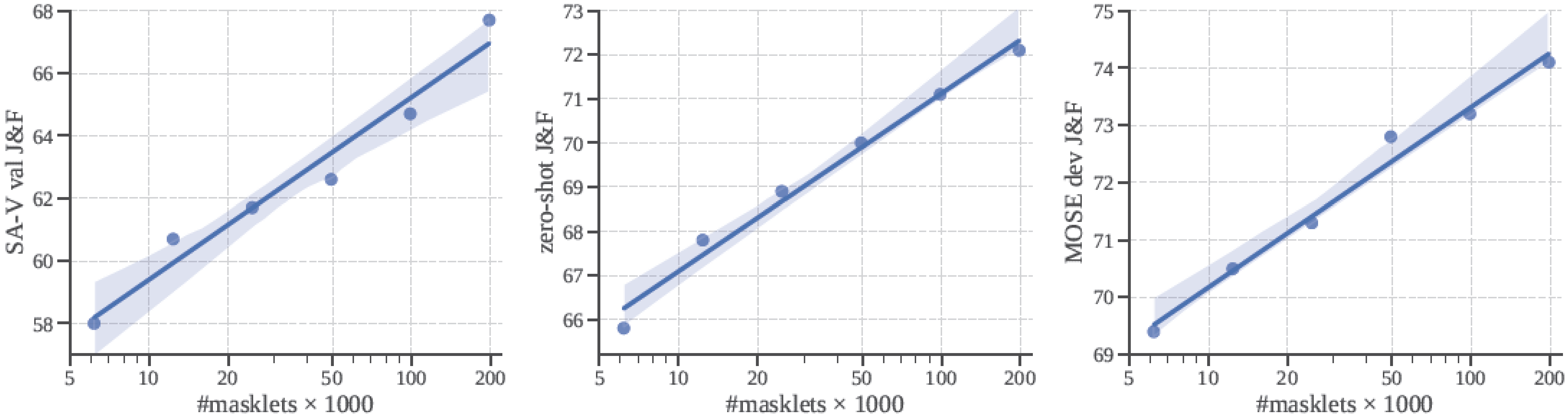
Data Ablations

- Data mix ablation
  - Compare the accuracy of SAM-2 when trained on different data mixtures
  - Best result when mixing all datasets

	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	58.6
5		✓	✓		63.0	73.2	73.3	70.9	55.8
6		✓	✓	✓	**63.6**	**75.0**	74.4	71.6	58.6
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	78.5	70.3	57.3
11	✓	✓	✓		61.8	74.4	78.5	**71.8**	55.7
12	✓	✓	✓	✓	63.1	73.7	**79.0**	71.6	**58.9**

Data Ablations

- Data quantity ablation
  - Report J&F accuracy for 3-click prompts in the first frame on SA-V val, 9 zero-shot datasets, and MOSE dev
  - Shows a consistent power law relationship between the quantity of training data and the video segmentation accuracy on all benchmarks



Data Ablations

- Data quality ablation
  - Experiment with filtering strategies for quality
  - However, it is worse than using all 190k SA-V masklets

	$\mathcal{J}\&\mathcal{F}$				mIoU
Setting	SA-V val	Intern-test	MOSE dev	9 zero-shot	SA-23
SA-1B + SA-V 50k random	63.7	70.3	72.3	68.7	<u>59.1</u>
SA-1B + SA-V 50k most edited	<u>66.2</u>	<u>73.0</u>	<u>72.5</u>	<u>69.2</u>	58.6
SA-1B + SA-V	**69.9**	**73.8**	**73.9**	**70.8**	**59.8**

Model Ablations

- Input size
  - Sample sequences of frames of fixed resolution and fixed length
  - A higher resolution leads to significant improvements across image and video tasks
    - Use an input resolution of 1024 in final model
  - Increasing the number of frames brings notable gains on video benchmarks
    - Use a default of 8 to balance speed and accuracy

	$\mathcal{J}\&\mathcal{F}$				mIoU
res.	MOSE dev	SA-V val	9 zero-shot	speed	SA-23
512	73.0	68.3	70.7	**1.00×**	59.7
768	76.1	**71.1**	**72.5**	0.43×	61.0
1024	**77.0**	70.1	72.3	0.22×	**61.5**

(a) Resolution.

	$\mathcal{J}\&\mathcal{F}$				mIoU
#frames	MOSE dev	SA-V val	9 zero-shot	speed	SA-23
4	71.1	60.0	67.7	1.00×	**60.1**
8	73.0	**68.3**	70.7	1.00×	59.7
10	**74.5**	68.1	**71.1**	1.00×	59.9

(b) #Frames.

Model Ablations

- Memory size
  - Increasing the (maximum) number of memories, N, generally helps the performance
    - Use a default value of 6 past frames to strike a balance between temporal context length and computational cost
  - Using fewer channels for memories does not cause much performance regression

	$\mathcal{J}\&\mathcal{F}$				mIoU
#mem.	MOSE dev	SA-V val	9 zero-shot	speed	SA-23
4	**73.5**	68.6	70.5	**1.01×**	**59.9**
6	73.0	68.3	**70.7**	1.00×	59.7
8	73.2	**69.0**	**70.7**	0.93×	**59.9**

(c) #Memories.

	$\mathcal{J}\&\mathcal{F}$				mIoU
chan. dim.	MOSE dev	SA-V val	9 zero-shot	speed	SA-23
64	73.0	**68.3**	**70.7**	**1.00×**	59.7
256	**73.4**	66.4	70.0	0.92×	**60.0**

(d) Memory channels.

Model Ablations

- Model size
  - More capacity in the image encoder or memory-attention (#self-/#cross-attention blocks) generally leads to improved results
    - Scaling the image encoder brings gains on both image and video metrics
    - Scaling the memory-attention only improves video metrics
    - Using a B+ image encoder → balance between speed and accuracy

	$\mathcal{J}\&\mathcal{F}$				mIoU
(#sa, #ca)	MOSE dev	SA-V val	9 zero-shot	speed	SA-23
(2, 2)	**73.3**	67.3	70.2	**1.13**×	59.9
(3, 2)	72.7	64.1	69.5	1.08×	**60.0**
(4, 4)	73.0	**68.3**	**70.7**	1.00×	59.7

(e) Memory attention.

	$\mathcal{J}\&\mathcal{F}$				mIoU
img. enc.	MOSE dev	SA-V val	9 zero-shot	speed	SA-23
S	70.9	65.5	69.4	**1.33**×	57.8
B+	73.0	**68.3**	70.7	1.00×	59.7
L	**75.0**	66.3	**71.9**	0.60×	**61.1**

(f) Image encoder size.

Model Ablations

- Relative positional encoding
  - Use 2d-RoPE in memory attention while removing RPB from the image encoder
  - Removing RPB also allows us to enable FlashAttention-2
    - Gives a significant speed boost at 1024 resolution
    - The higher resolution of 1024, the speed gap between 2d-RoPE (1st row) and the no RoPE baseline (3rd row) becomes much smaller.

Model Ablations

- Relative positional encoding
  - Removing all RPB from the image encoder, with no performance regression on SA-23 and minimal regression on video benchmarks while giving a significant speed boost at 1024 resolution.
  - Find it is beneficial to use 2d-RoPE in the memory attention.

		$\mathcal{J}\&\mathcal{F}$					mIoU
RPB in img. enc.	2d-RoPE in mem. attn.	MOSE dev	SA-V val	LVOSv2 val	9 zero-shot	speed	SA-23
	✓	73.0	**68.3**	**71.6**	70.7	1.00×	59.7
✓	✓	**73.6**	67.9	71.0	**71.5**	0.93×	**60.0**
		72.8	67.1	70.3	70.3	**1.04**×	59.9

Conclusion

- Three key aspects
  - Extending the promptable segmentation task to video
  - Equipping the SAM architecture to use memory when applied to video
  - The diverse SA-V dataset for training and benchmarking video segmentation

Planned Tasks for This Week

- 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 and Christoph Feichtenhofer. **SAM 2: Segment Anything in Images, Videos**. arXiv:2408.00714, Aug 2024.
- Read paper about Video Grounding DINO.