Meta

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

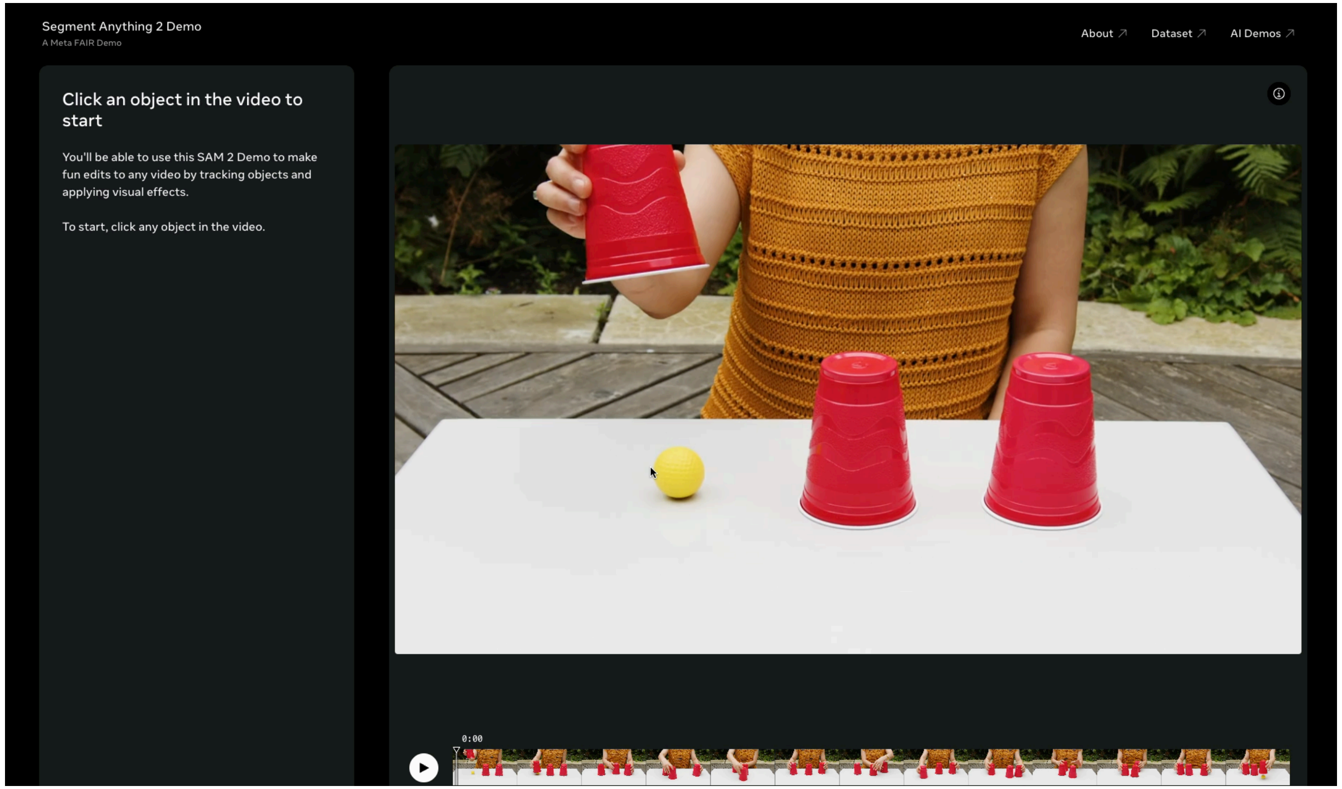
[*]core contributor, [†]project lead

Accepted by ICLR 2025 (Oral)
OpenReview Rating: 8, 8, 10, 10

Presenter: Pengyu Zhang
Date: May-07-2025

Contributions

- **New task**: Promptable Visual Segmentation (PVS) – expand SAM1 in *image and video* segmentation.
- **New dataset**: SA-V – a large-scale dataset for video object segmentation engined by SAM2.

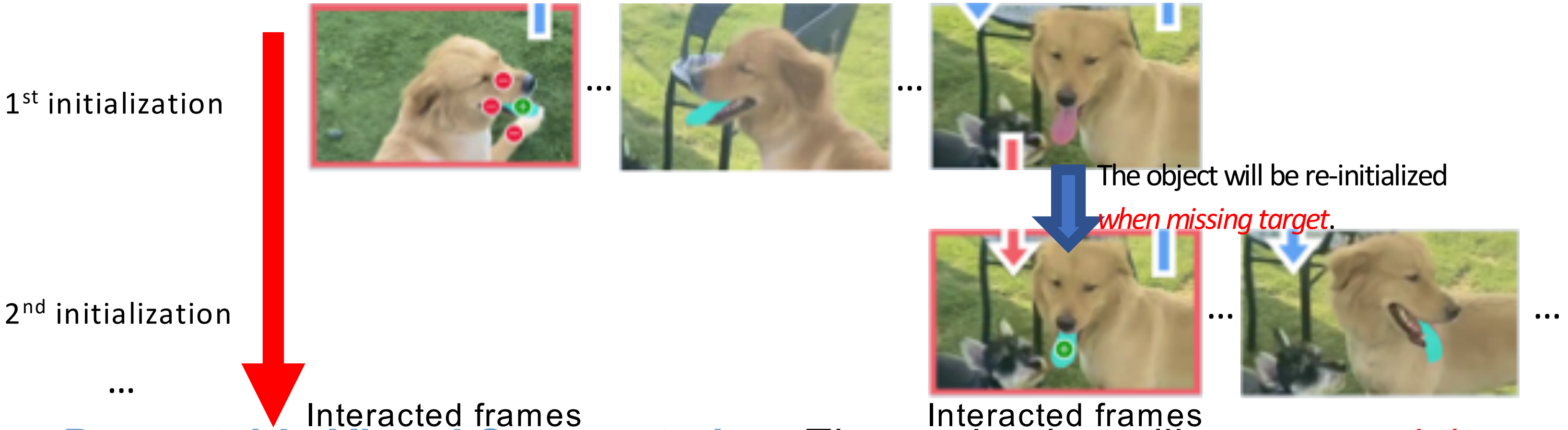


- Track arbitrary object
- Satisfying performance on occlusion and similar appearance
- Potential applications on video editing: object removal, pixelate and colorization, etc.

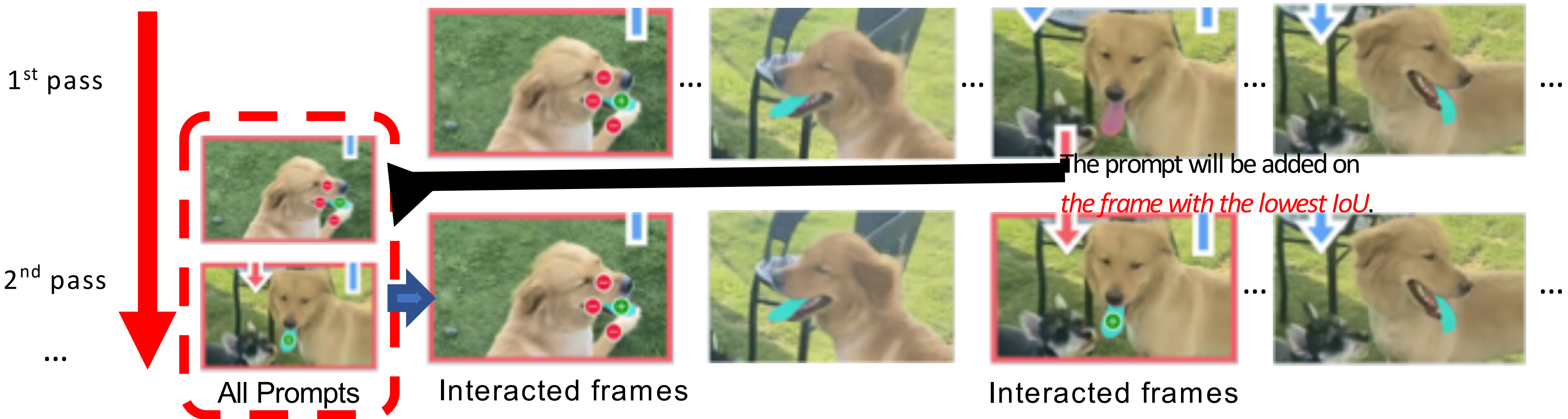
Task Description

Promptable Visual Segmentation (PVS) allows providing prompts (e.g. points, boxes and masks) to the model on any frame of a video.

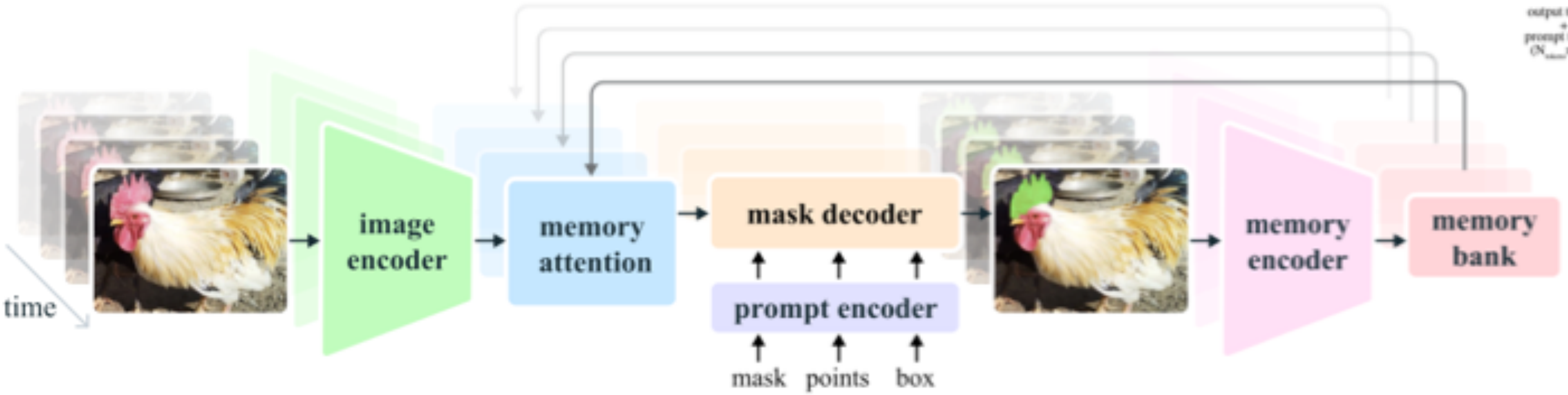
- **Online Promptable Visual Segmentation**: The evaluation follows *one-pass evaluation* manner.

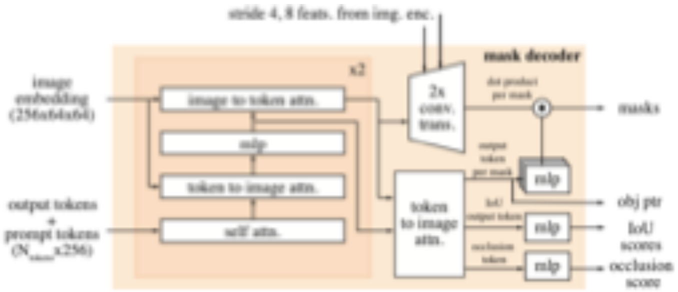


- **Offline Promptable Visual Segmentation**: The evaluation will run several times with all previous prompts.



SAM2 Model





Mask decoder

Image encoder: The *image-level features* are extracted by MAE Pretrained Hiera image encoder[1].

Memory attention: Several Transformer blocks, with self-attention and cross-attention with memories.

Prompt encoder: The point and box prompts are represented by positional encodings and masks are firstly embedded by convolutions and summed with the frame embedding.

Memory encoder and memory bank: The memory generates a memory by downsampling the output mask using a convolutional module. The memory bank contains *N memories* (downsampled output mask), *M prompted frames* (prompts with frame embedding) and *a list of object pointers* (foreground object features)

Mask decoder: generates both mask and occlusion confidence to evaluate the quality of generated mask.

Experiments

Significant improvement against baseline method

Results on semi-supervised VOS

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	**64.7**	**75.3**	**77.6**	**74.4**	**79.3**

Table 4 Zero-shot accuracy across 17 video datasets using different prompts. We report average accuracy for each type of prompt (1, 3 or 5 clicks, bounding boxes, or ground-truth masks) in the first video frame (‡: this case directly uses masks as inputs into XMem++ or Cutie without SAM).

Method	$\mathcal{J}\&\mathcal{F}$					$\mathcal{G}$
	MOSE val	DAVIS 2017 val	LVOS val	SA-V val	SA-V test	YTVOS 2019 val
STCN (Cheng et al., 2021a)	52.5	85.4	-	61.0	62.5	82.7
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
XMem (Cheng & Schwing, 2022)	59.6	86.0	-	60.1	62.3	85.6
SimVOS-B (Wu et al., 2023b)	-	88.0	-	44.2	44.1	84.2
JointFormer (Zhang et al., 2023b)	-	90.1	-	-	-	87.4
ISVOS (Wang et al., 2022)	-	88.2	-	-	-	86.3
DEVA (Cheng et al., 2023b)	66.0	87.0	55.9	55.4	56.2	85.4
Cutie-base (Cheng et al., 2023a)	69.9	87.9	66.0	60.7	62.7	87.0
Cutie-base+ (Cheng et al., 2023a)	71.7	88.1	-	61.3	62.8	87.5
SAM 2 (Hiera-B+)	76.6	90.2	**78.0**	76.8	77.0	88.6
SAM 2 (Hiera-L)	**77.9**	**90.7**	**78.0**	**77.9**	**78.4**	**89.3**

SOTA performance on DAVIS and most VOS datasets.

Experiments

Results on Offline PVS

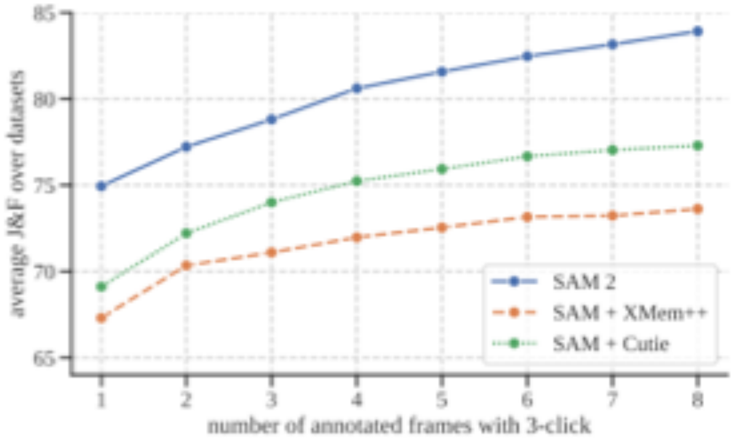
Method	EndoVis 2018	ESD	LVOSv2	LV-VIS	PUMaVOS	UVO	VIPSeg	Virtual KITTI 2	VOST	(average)
SAM + XMem++	68.9	88.2	72.1	86.4	60.2	74.5	84.2	63.8	46.6	71.7
SAM + Cutie	71.8	87.6	82.1	87.1	59.4	75.2	84.4	70.3	54.3	74.7
SAM 2	**77.0**	**90.2**	**87.9**	**90.3**	**68.5**	**79.2**	**88.3**	**74.1**	**67.5**	**80.3**

(b) average $\mathcal{J}\&\mathcal{F}$ on each dataset over 8 interacted frames (3-click)

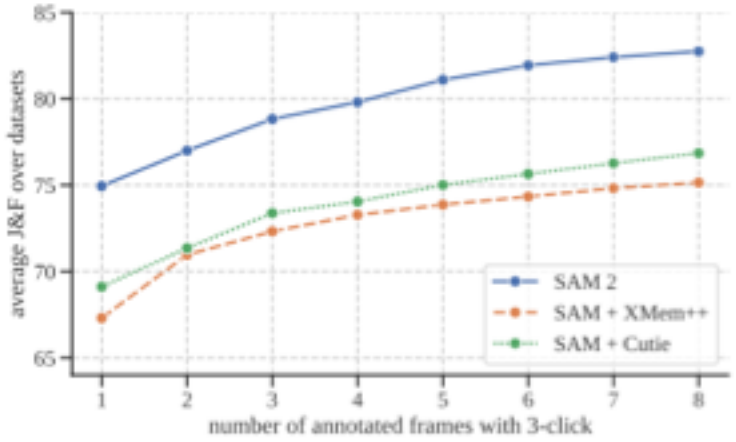
Results on Online PVS

Method	EndoVis 2018	ESD	LVOSv2	LV-VIS	PUMaVOS	UVO	VIPSeg	Virtual KITTI 2	VOST	(average)
SAM + XMem++	71.4	87.8	72.9	85.2	63.7	74.7	82.5	63.9	52.7	72.8
SAM + Cutie	70.5	87.3	80.6	86.0	58.9	75.2	82.1	70.4	54.6	74.0
SAM 2	**77.5**	**88.9**	**87.8**	**88.7**	**72.7**	**78.6**	**85.5**	**74.0**	**65.0**	**79.8**

(b) average $\mathcal{J}\&\mathcal{F}$ on each dataset over 8 interacted frames (3-click)



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



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

>5% higher than the baseline

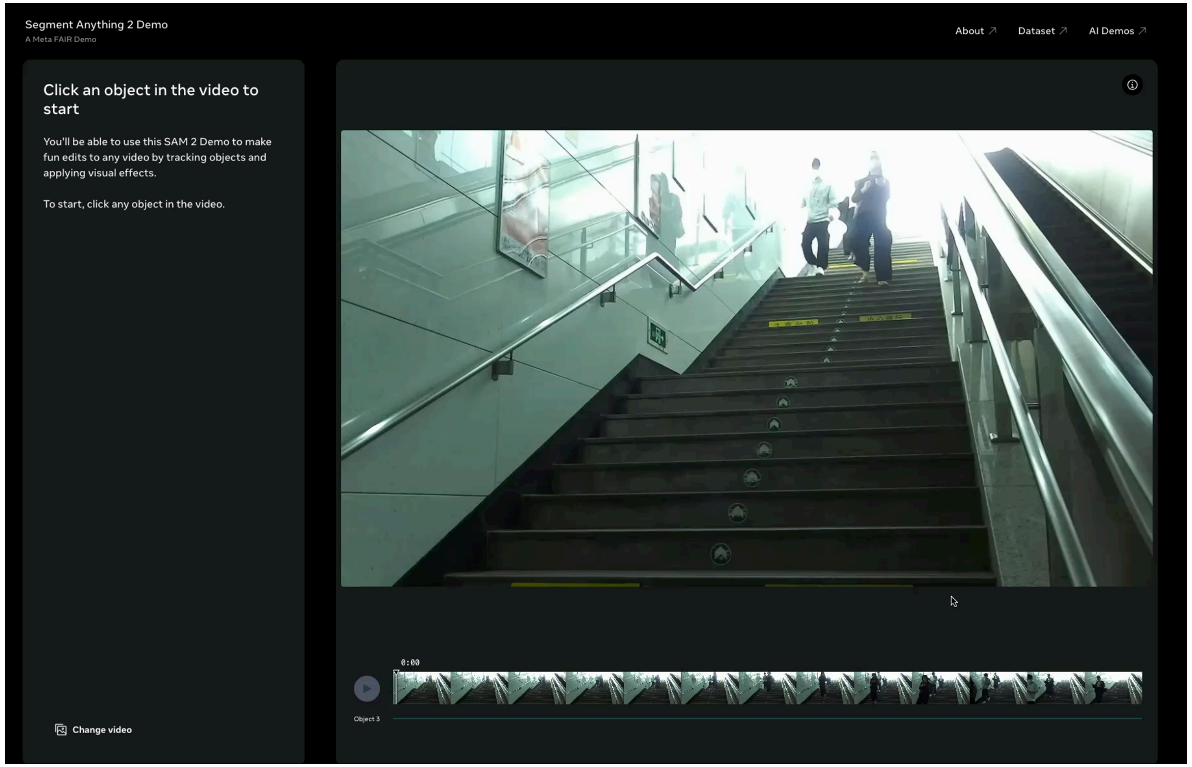
SA-V Dataset

	#Videos	Duration	#Masklets	#Masks	#Frames	Disapp. Rate
DAVIS 2017 (Pont-Tuset et al., 2017)	0.2K	0.1 hr	0.4K	27.1K	10.7K	16.1 %
YouTube-VOS (Xu et al., 2018b)	4.5K	5.6 hr	8.6K	197.3K	123.3K	13.0 %
UVO-dense (Wang et al., 2021b)	1.0K	0.9 hr	10.2K	667.1K	68.3K	9.2 %
VOST (Tokmakov et al., 2022)	0.7K	4.2 hr	1.5K	175.0K	75.5K	41.7 %
BURST (Athar et al., 2022)	2.9K	28.9 hr	16.1K	600.2K	195.7K	37.7 %
MOSE (Ding et al., 2023)	2.1K	7.4 hr	5.2K	431.7K	638.8K	41.5 %
Internal	62.9K	281.8 hr	69.6K	5.4M	6.0M	36.4 %
SA-V Manual	50.9K	196.0 hr	190.9K	10.0M	4.2M	42.5 %
SA-V Manual+Auto	50.9K	196.0 hr	642.6K	35.5M	4.2M	27.7 %

- Very large-scale video dataset for object segmentation
- General object categories
- 190.9K manual masklets and 451.7K automatic masklets
- Semi-supervised annotation
  - Step1: Image-level annotation using SAM
  - Step2: Video-level annotation using SAM and SAM2
  - Step3: Video-level annotation using fully-featured SAM2
- Auto masklet generation using SAM2

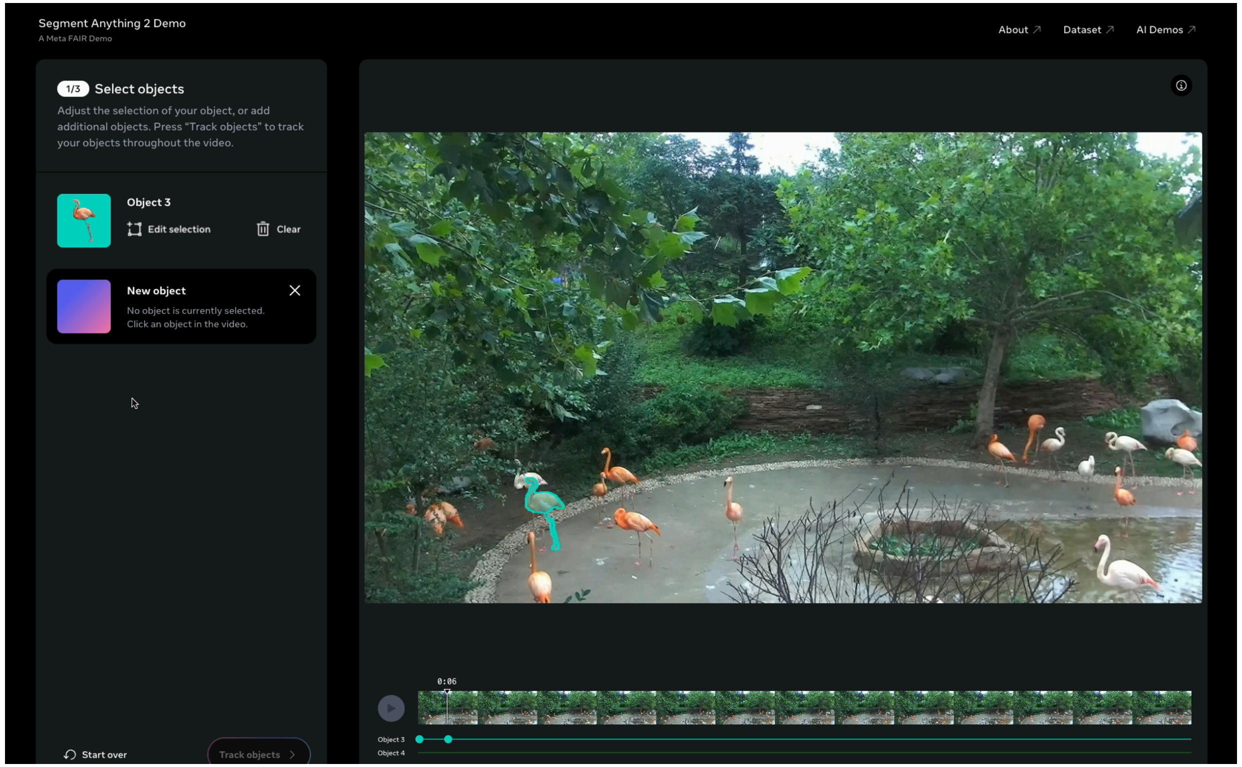
Try the demo

Extreme Illumination



Hard to segment
Missing segmentation when scale variation

Occlusion and Similar Appearance



Inferior performance when occlusion

A Distractor-Aware Memory for Visual Object Tracking with SAM2

Jovana Videnovic*, Alan Lukezic*, Matej Kristan
Faculty of Computer and Information Science, University of Ljubljana, Slovenia
jv8043@student.uni-lj.si, {alan.lukezic, matej.kristan}@fri.uni-lj.si

Accepted by CVPR2025 Poster

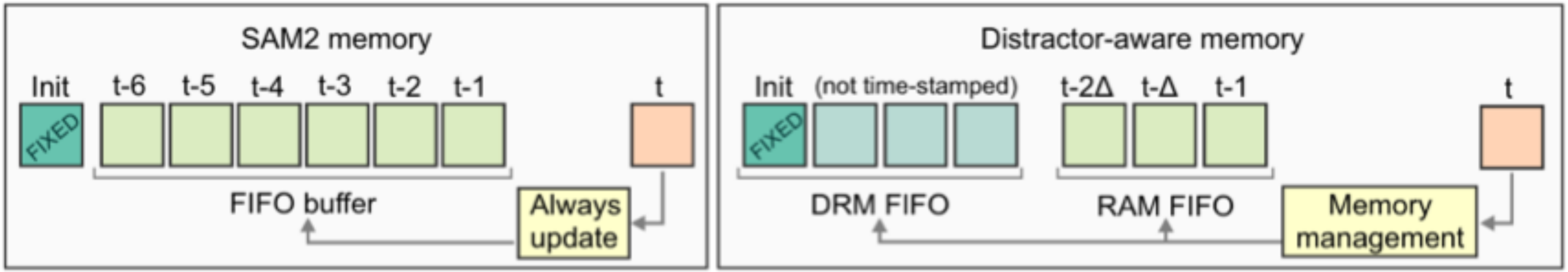
Motivation and Contributions

Visual trackers struggle in the scenes with distractors, which indicates the *importance of the memory model*.

- A *distractor-aware memory model* are proposed to stress the importance of memory model in visual tracking, which contains Recent Appearance Memory (RAM) and Distractor Resolving Memory (DRM).

- A new *distractor-distilled (DiDi) dataset* is proposed to study the distractor problem.

Distractor-Aware Memory (DAM)



The memory model in SAM2 only utilizes *the most recent appearances to model the target*, suffering model draft when the distractors occur.

In DAM, the memory can be separated into *Recent Appearance Model (RAM)* and *Distractor Resolving Memory (DRM).*

- RAM stores the most previous appearance and updates within a fixed interval (Δ = 5)
- DRM stores the critical information for resolving distractors, *where the distractor is detected by the SAM2 model within high confidence.* (The overlap between alternative and selected masks is less than a threshold and the IoU score from SAM is larger than a threshold.)

Provide a training-free method to enhance the ability in handling distractors.

Distractor-Distilled (DiDi) Dataset

A subset of popular tracking benchmarks, which contains non-negligible distractors. – The feature similarity between regions outside and inside the bounding box area.

- Select 180 sequences from 808 sequence.
- Selected from GOT10k, LaSOT, UTB180, VOT-ST2020, VOT-LT2020, VOT-ST2022, VOT-LT2022.
- Most sequences are from VOT challenges

Experiments

Table 1. SAM2.1++ architecture justification on DiDi dataset.

	Quality	Accuracy	Robustness
SAM2.1	0.649	0.720	0.887
$\text{SAM2.1}_{\text{PRES}}$	0.665	0.723	0.903
$\text{SAM2.1}_{\Delta=5}$	0.667	0.718	0.914
$\text{SAM2.1}_{\text{DRM1}}$	0.672	0.710	0.932
$\text{SAM2.1}_{\text{DRM2}}$	0.644	0.691	0.913
SAM2.1++	0.694	0.727	0.944

$\text{SAM2.1}_{\text{PRES}}$: Suspend the memory update when the target is absent.
$\text{SAM2.1}_{\Delta=5}$: Update the memory within the interval of 5 frames.
$\text{SAM2.1}_{\text{DRM1}}$: Update DRM only when the IoU score is larger than a threshold.
$\text{SAM2.1}_{\text{DRM2}}$: Update DRM when the distractor is detected.
SAM2.1++: The proposed method.

Frequent memory update will influence the robustness of appearance model due to the appearance redundancy.

DRM module highly depends on the segmentation accuracy.

A simple modification on memory module can increase performance significantly !

Table 2. State-of-the-art comparison on DiDi dataset.

	Quality	Accuracy	Robustness
SAMURAI [45]	0.680 ②	0.722 ③	0.930 ②
SAM2.1Long [14]	0.646	0.719	0.883
ODTrack [50]	0.608	**0.740** ①	0.809
Cutie [9]	0.575	0.704	0.776
AOT [47]	0.541	0.622	0.852
AQATrack [40]	0.535	0.693	0.753
SeqTrack [6]	0.529	0.714	0.718
KeepTrack [32]	0.502	0.646	0.748
TransT [5]	0.465	0.669	0.678
SAM2.1 [36]	0.649 ③	0.720	0.887 ③
SAM2.1++	**0.694** ①	0.727 ②	**0.944** ①

Table 6. State-of-the-art comparison on three standard bounding-box benchmarks.

	LaSoT (AUC)	$\text{LaSoT}_{\text{ext}}$ (AUC)	GoT10k (AO)
MixViT [11]	72.4	-	75.7
LORAT [27]	**75.1** ①	56.6 ③	78.2 ③
ODTrack [50]	74.0 ②	53.9	78.2 ③
DiffusionTrack [30]	72.3	-	74.7
DropTrack [38]	71.8	52.7	75.9
SeqTrack [6]	72.5 ③	50.7	74.8
MixFormer [10]	70.1	-	71.2
GRM-256 [17]	69.9	-	73.4
ROMTrack [4]	71.4	51.3	74.2
OSTrack [48]	71.1	50.5	73.7
KeepTrack [32]	67.1	48.2	-
TOMP [33]	68.5	-	-
SAM2.1 [36]	70.0	56.9 ②	80.7 ②
SAM2.1++	**75.1** ①	**60.9** ①	**81.1** ①

Thanks for Listening!

Q&A