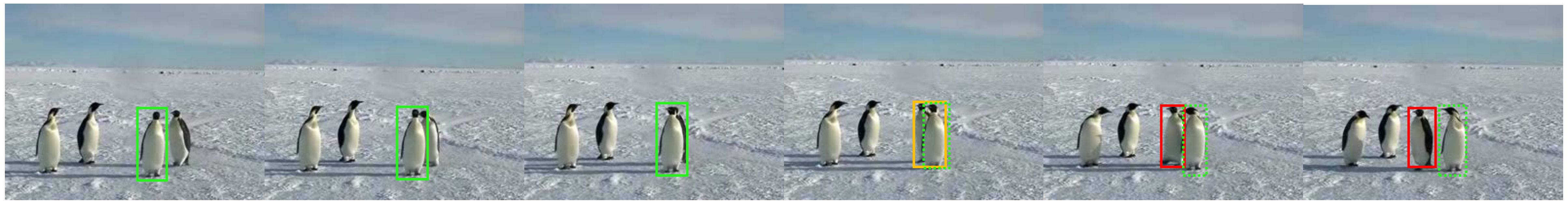
SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory

STUDENTS: CHENG-YEN YANG, HSIANG-WEI HUANG, WENHAO CHAI, ZHONGYU JIANG

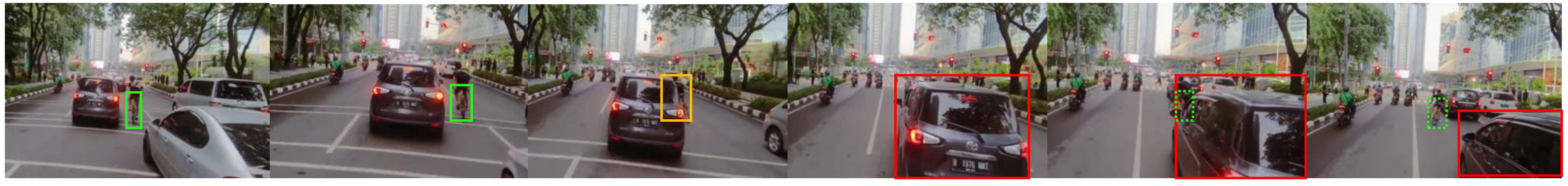
TL;DR: WE PROPOSED A MOTION-AWARE MEMORY ON TOP OF SAM2 FOR ZERO-SHOT VISUAL TRACKING!

CHALLENGES FOR VISUAL TRACKING

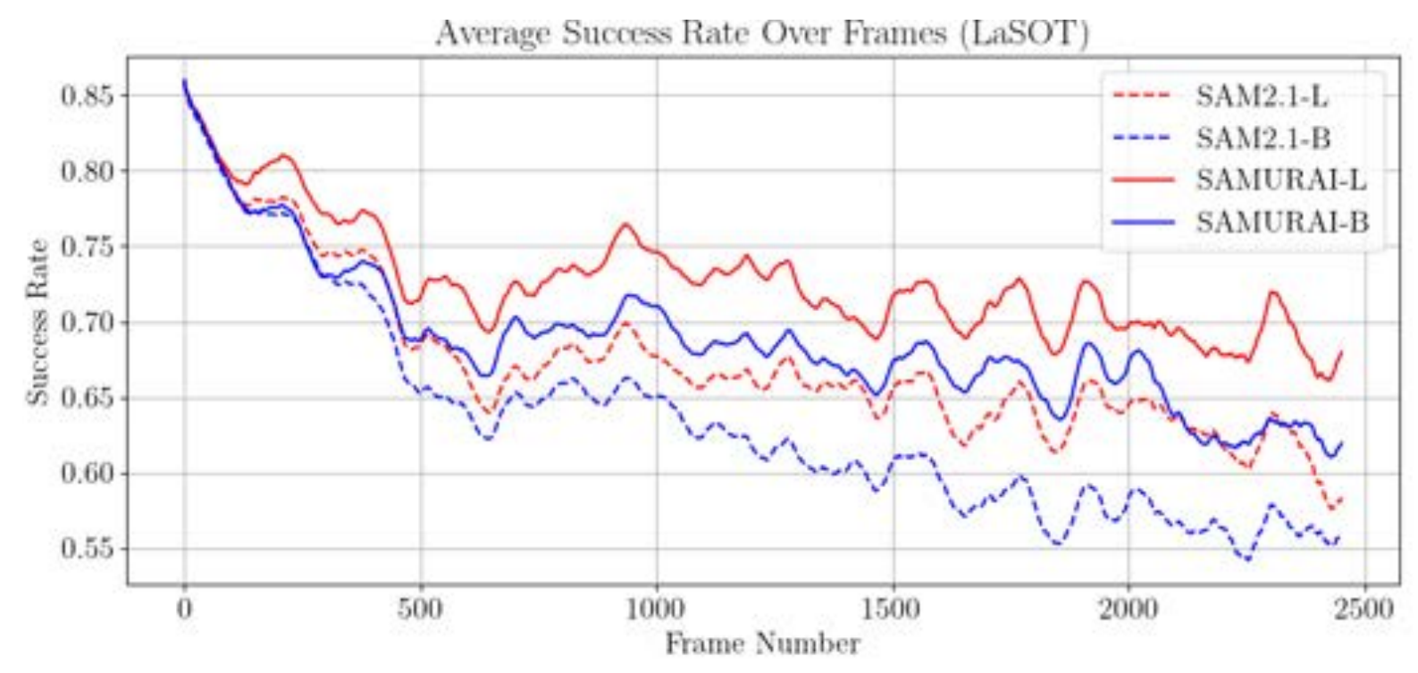
Case 1: Ambiguous prediction in crowded scene with similar appearance 💡 Consider motion during mask selection!

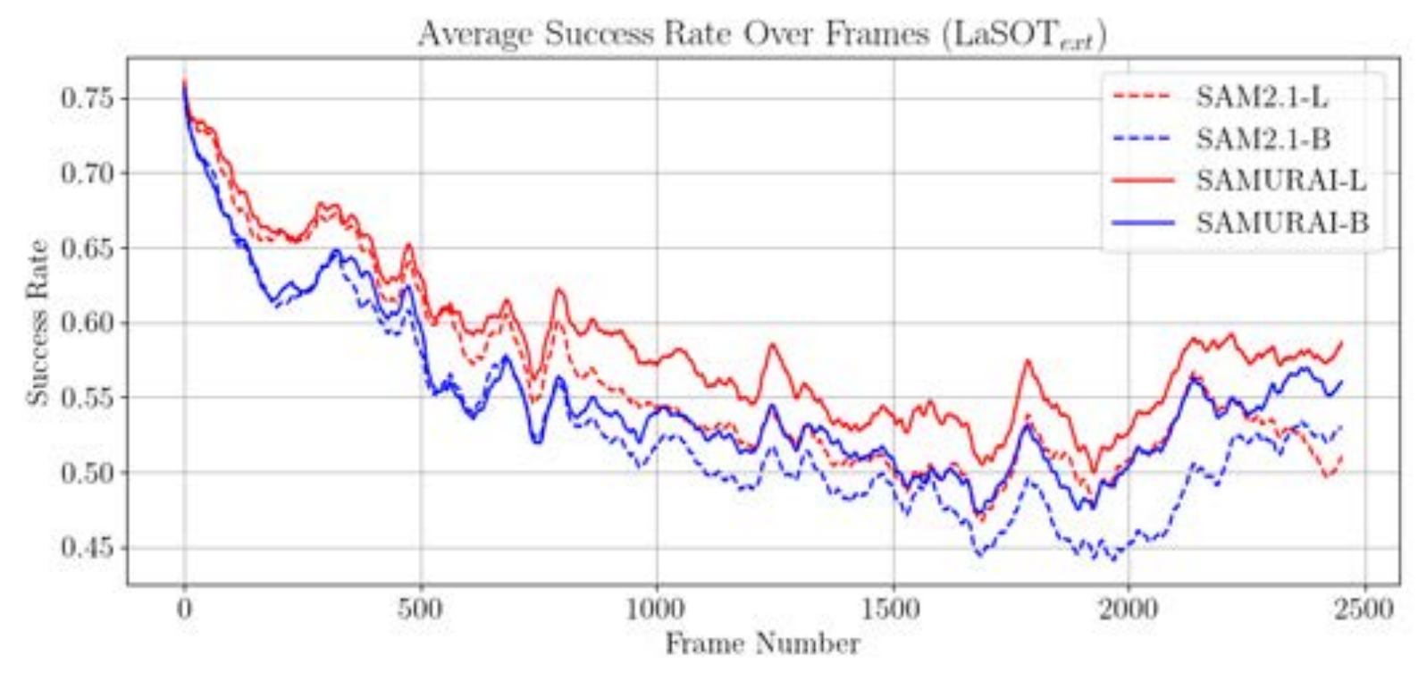


Case 2: Ambiguous prediction in occlusion resulting bad memory feature 💡 Motion-aware memory selection!

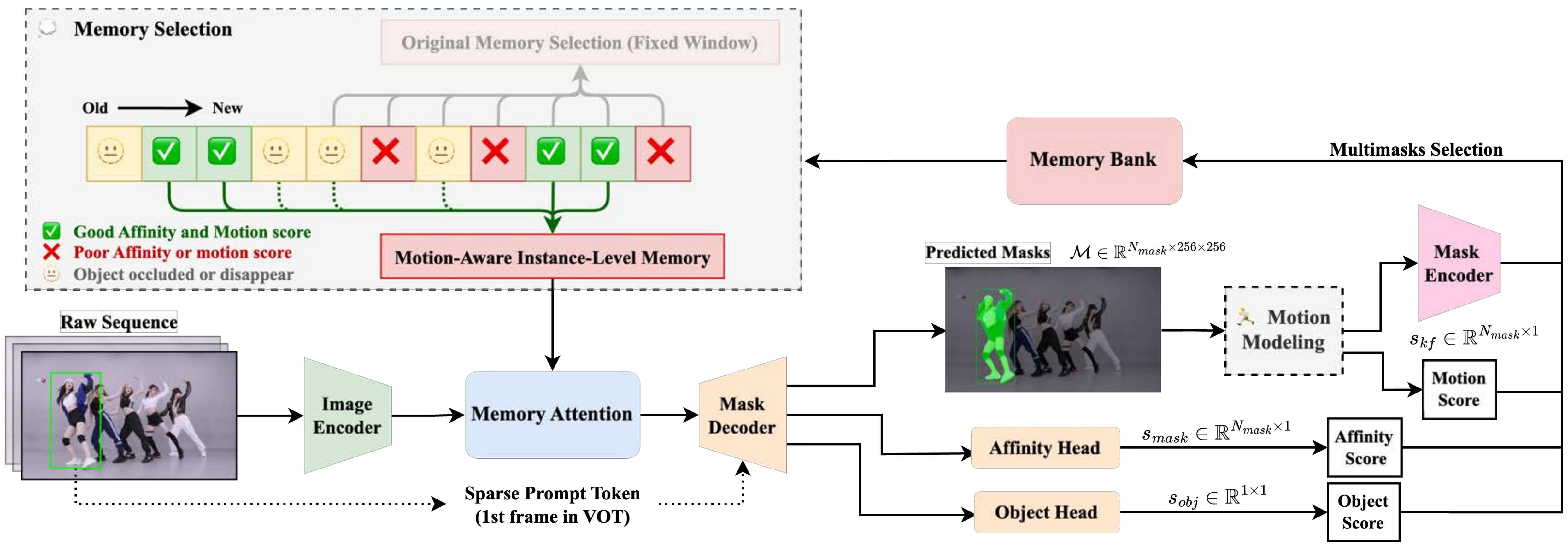


🟩 GT 🟨 Ambiguous Prediction 🟥 Incorrect Prediction





OVERALL FRAMEWORK: SAMURAI



Memory Selection

Original Memory Selection (Fixed Window)

Old → New

☑ Good Affinity and Motion score

☒ Poor Affinity or motion score

Object occluded or disappear

Motion-Aware Instance-Level Memory

Raw Sequence

Image Encoder

Memory Attention

Mask Decoder

Sparse Prompt Token (1st frame in VOT)

Memory Bank

Multimasks Selection

Predicted Masks $\mathcal{M} \in \mathbb{R}^{N_{mask} \times 256 \times 256}$

Motion Modeling

Mask Encoder

$s_{kf} \in \mathbb{R}^{N_{mask} \times 1}$

Motion Score

Affinity Head

$s_{mask} \in \mathbb{R}^{N_{mask} \times 1}$

Affinity Score

Object Head

$s_{obj} \in \mathbb{R}^{1 \times 1}$

Object Score

SAM-based **U**nified and **R**obust zero-shot visual tracker with motion-**A**ware **I**nstance-level memory

MOTION-AWARE MEMORY UPDATE

Algorithm 1 Motion-Aware Memory Bank Update

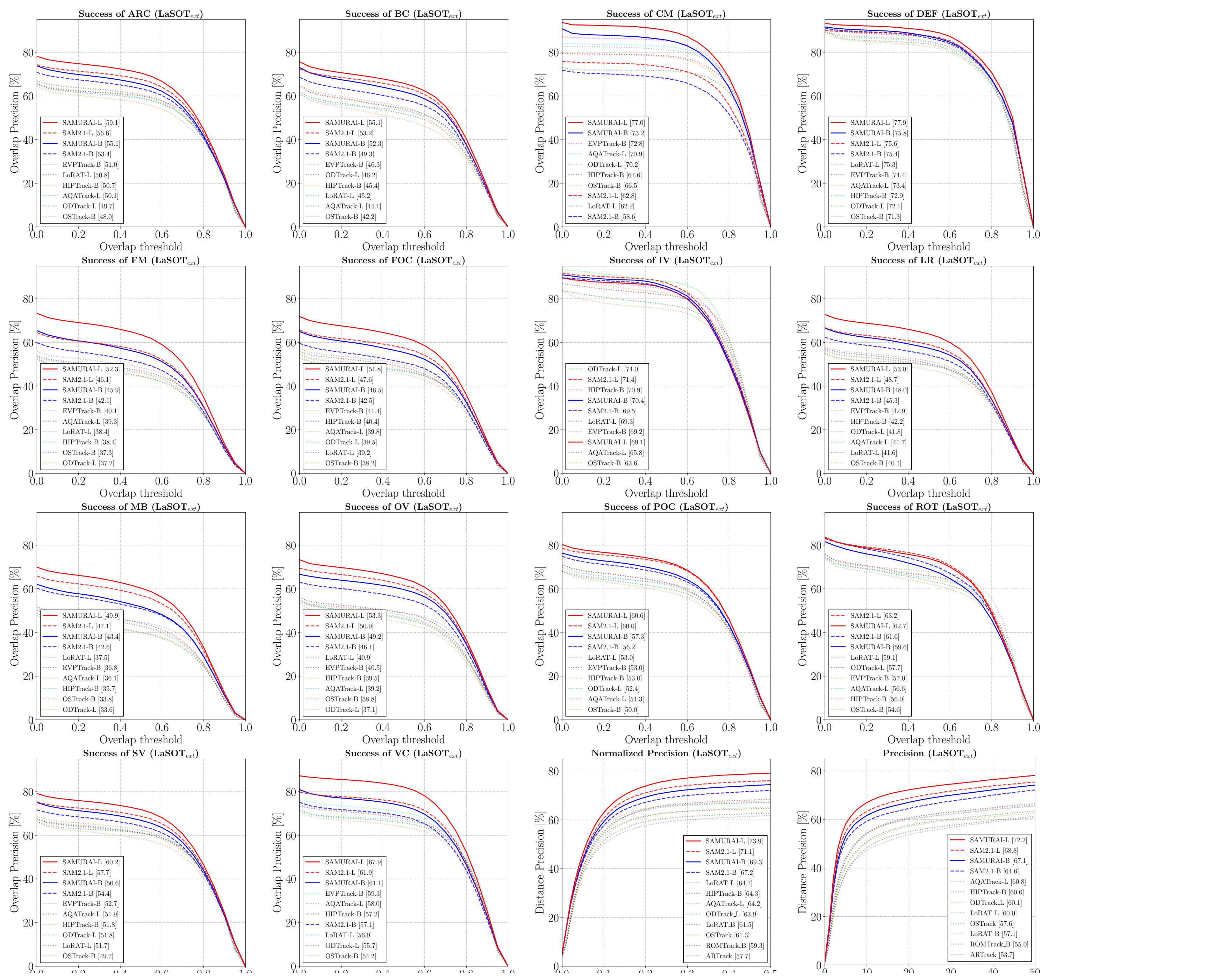
```
Input: Video frames $V$, Memory Bank $\mathcal{B}$, Kalman Filter State $\mathcal{K}$, Thresholds $\tau_{mask}, \tau_{obj}, \tau_{kf}$, Trajectory $\mathcal{R}$, Weight $w_{kf}$
for $f = 0$ to $|V| - 1$ do
    $I_{\text{emb}} \leftarrow \text{MemoryAttention}(\text{ImageEncoder}(V_f), \mathcal{B})$
    $(m, b, s_{mask}, s_{obj}) \leftarrow \text{MaskDecoder}(x_{\text{prompt}}, I_{\text{emb}})$
    // Predict object location using Kalman filter
    $b_{kf} \leftarrow \mathcal{K}.\text{predict}()$
    // Calculate KF-IoU scores
    $s_{kf} \leftarrow \text{IoU}(b_{kf}, b)$
    // Select best mask and bounding box
    $(m_s, b_s) \leftarrow \text{argmax}(\alpha_{kf} \cdot s_{kf}(\mathcal{M}_i) + (1 - \alpha_{kf}) \cdot s_{mask,i})$
    // Update Kalman filter with selected box
    $\mathcal{K}.\text{update}(b_s)$
    // Update memory bank
    $\mathcal{R}.\text{append}(m_s, s_{mask}[m_s], s_{obj}[m_s], s_{kf}[m_s])$
    // Construct memory features
    $\mathcal{B} \leftarrow [], fid \leftarrow f$
    while $|\mathcal{B}| < N_{mem}$ and $fid \geq 0$ do
        $(\_, s_{mask}, s_{obj}, s_{kf}) \leftarrow \mathcal{R}[fid]$
        if $s_{mask} > \tau_{mask}$ and $s_{obj} > \tau_{obj}$ and $s_{kf} > \tau_{kf}$ then
            $\mathcal{B}.\text{append}(\mathcal{M}_{fid})$
        end if
        $fid \leftarrow fid - 1$
    end while
end for
```

EXPERIMENT RESULTS

Trackers	Source	LaSOT			LaSOT$_{ext}$			GOT-10k		
		AUC(%)	P$_{norm}$(%)	P(%)	AUC(%)	P$_{norm}$(%)	P(%)	AO(%)	OP$_{0.5}$(%)	OP$_{0.75}$(%)
SiamRPN++	CVPR'19	49.6	56.9	49.1	34.0	41.6	39.6	51.7	61.6	32.5
DiMP$_{288}$	CVPR'20	56.3	64.1	56.0	-	-	-	61.1	71.7	49.2
TransT$_{256}$	CVPR'21	64.9	73.8	69.0	-	-	-	67.1	76.8	60.9
AutoMatch$_{255}$	ICCV'21	58.2	67.5	59.9	-	-	-	65.2	76.6	54.3
STARK$_{320}$	ICCV'21	67.1	76.9	72.2	-	-	-	68.8	78.1	64.1
SwinTrack-B$_{384}$	NeurIPS'22	71.4	79.4	76.5	-	-	-	72.4	80.5	67.8
MixFormer$_{288}$	CVPR'22	69.2	78.7	74.7	-	-	-	70.7	80.0	67.8
OSTrack$_{384}$	ECCV'22	71.1	81.1	77.6	50.5	61.3	57.6	73.7	83.2	70.8
ARTrack-B$_{256}$	CVPR'23	70.8	79.5	76.2	48.4	57.7	53.7	73.5	82.2	70.9
SeqTrack-B$_{384}$	CVPR'23	71.5	81.1	77.8	50.5	61.6	57.5	74.5	84.3	71.4
GRM-B$_{256}$	CVPR'23	69.9	79.3	75.8	-	-	-	73.4	82.9	70.4
NCSiam-L	TIP'23	63.9	72.4	67.0	-	-	-	67.8	78.0	61.3
ROMTrack-B$_{256}$	ICCV'23	69.3	78.8	75.6	47.2	53.5	52.9	72.9	82.9	70.2
TaMOs-B$_{384}$	WACV'24	70.2	79.3	77.8	-	-	-	-	-	-
EVPTrack-B$_{384}$	AAAI'24	72.7	82.9	80.3	53.7	65.5	61.9	76.6	86.7	73.9
ODTrack-L$_{384}$	AAAI'24	74.0	**84.2**	**82.3**	53.9	65.4	61.7	78.2	87.2	**77.3**
HIPTrack-B$_{384}$	CVPR'24	72.7	82.9	79.5	53.0	64.3	60.6	77.4	88.0	74.5
AQATrack-L$_{384}$	CVPR'24	72.7	82.9	80.2	52.7	64.2	60.8	76.0	85.2	74.9
MCTrack-B$_{384}$	TIP'24	72.2	81.6	77.7	51.1	61.8	58.8	76.5	87.1	75.4
LoRAT-L$_{224}$	ECCV'24	**74.2**	83.6	80.9	52.8	64.7	60.0	75.7	84.9	75.0
SAMURAI-T	Ours	69.3	76.4	73.8	55.1	65.6	63.7	79.0	89.6	72.3
SAMURAI-S	Ours	70.0	77.6	75.2	58.0	69.6	67.7	78.8	88.7	72.9
SAMURAI-B	Ours	70.7	78.7	76.2	57.5	69.3	67.1	79.6	90.8	72.9
SAMURAI-L	Ours	**74.2**	82.7	80.2	**61.0**	**73.9**	**72.2**	**81.7**	**92.2**	76.9

Trackers	LaSOT			LaSOT$_{ext}$		
	AUC(%)	P$_{norm}$(%)	P(%)	AUC(%)	P$_{norm}$(%)	P(%)
SAM2.1-T	66.70	73.70	71.22	52.25	62.03	60.30
SAMURAI-T	**69.28** (+2.58)	**76.39** (+2.69)	**73.78** (+2.56)	**55.13** (+2.88)	**65.60** (+2.57)	**63.72** (+3.42)
SAM2.1-S	66.47	73.67	71.25	56.11	67.57	65.81
SAMURAI-S	**70.04** (+3.57)	**77.55** (+3.88)	**75.23** (+3.98)	**57.99** (+1.88)	**69.60** (+2.03)	**67.73** (+1.92)
SAM2.1-B	65.97	73.54	70.96	55.51	67.17	64.55
SAMURAI-B	**70.65** (+4.68)	**78.69** (+4.15)	**76.21** (+5.25)	**57.48** (+1.97)	**69.28** (+2.11)	**67.09** (+2.54)
SAM2.1-L	68.54	76.16	73.59	58.55	71.10	68.83
SAMURAI-L	**74.23** (+5.69)	**82.69** (+6.53)	**80.21** (+6.62)	**61.03** (+2.48)	**73.86** (+2.76)	**72.24** (+3.41)

State-of-the-art on multiple benchmarks: LaSOT$_{ext}$, GOT-10k, VOT2020, VOT2022, TrackngNet, NFS!



ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISOR: JENQ-NENG HWANG

Stars 6.6k

FOR CODE, DEMO, AND MORE!