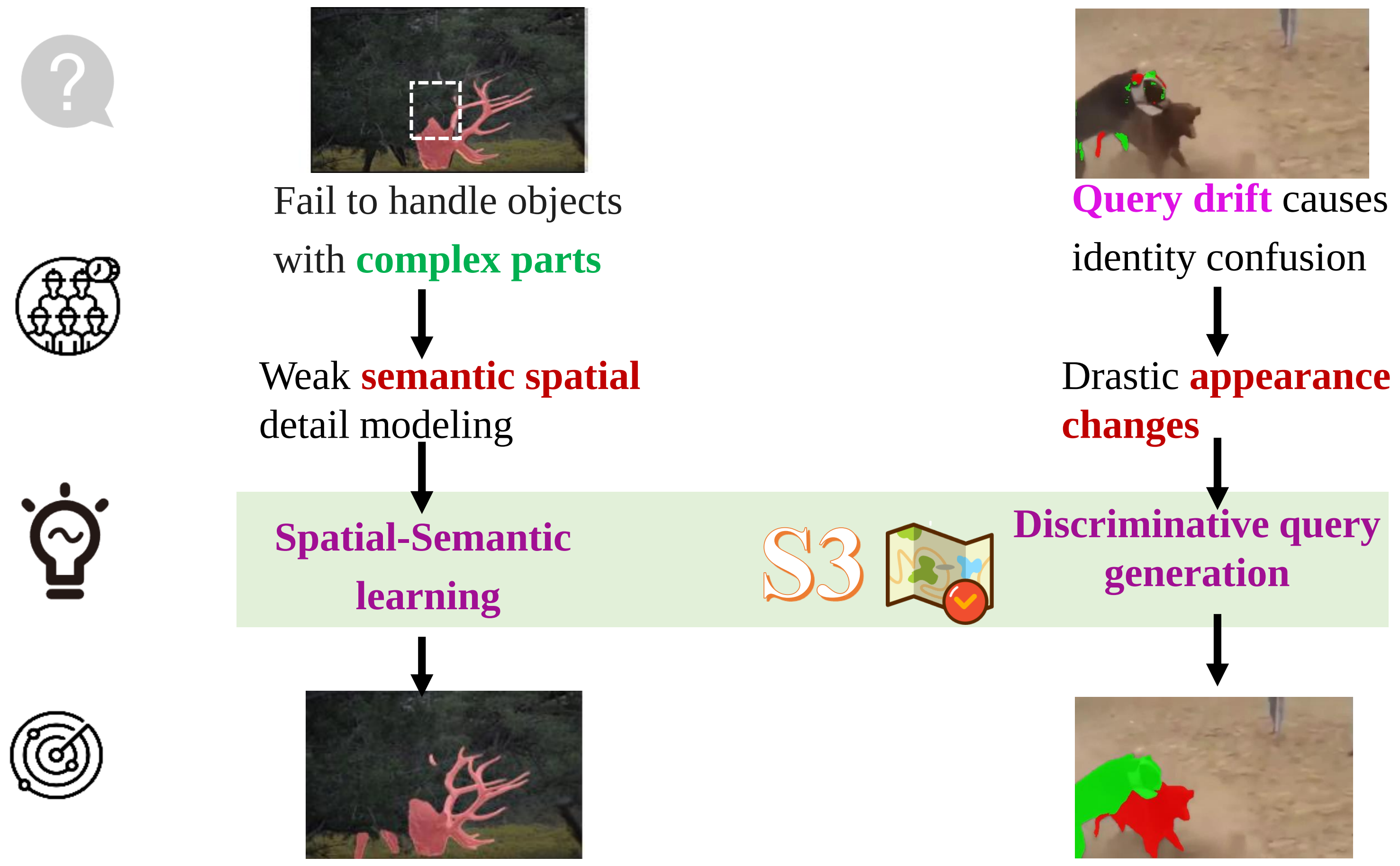
S3

Learning Spatial-Semantic Features for Robust Video Object Segmentation

Xin Li[123#], Deshui Miao[2#], Zhenyu He[21], Yaowei Wang[21*], Huchuan Lu[3], Ming−Hsuan Yang[56]*

[1]*Pengcheng Laboratory* [2] *Harbin Institute of Technology, Shenzhen* [3] *Pazhou Lab (Huangpu)*
[4] *Dalian University of Technology* [5] *University of California at Merced* [6] *Yonsei University*

Motivation



Fail to handle objects with **complex parts**

Weak **semantic spatial** detail modeling

Query drift causes identity confusion

Drastic **appearance changes**

Spatial-Semantic learning

S3

Discriminative query generation

Contributions

1. We propose a **spatial-semantic block** to incorporate semantic information with spatial information for VOS, which integrates global semantic information from the CLS token of a pre-trained ViT backbone into the base features of the input sample and then models spatial dependencies using a spatial dependency modeling module.

2. We develop a **discriminative query mechanism** to capture the most representative region of the target for better target representation learning and updating.

3. We demonstrate that the proposed method achieves **state-of-the-art performance** on five diverse datasets and evaluate the contribution of each proposed component with comprehensive ablation studies.

Implement Details

◆ Base Training Setting

- ✓ Trained on YoutubeVOS and DAVIS.
- ✓ ViT-base trained from **DepthAnything** is adopted.

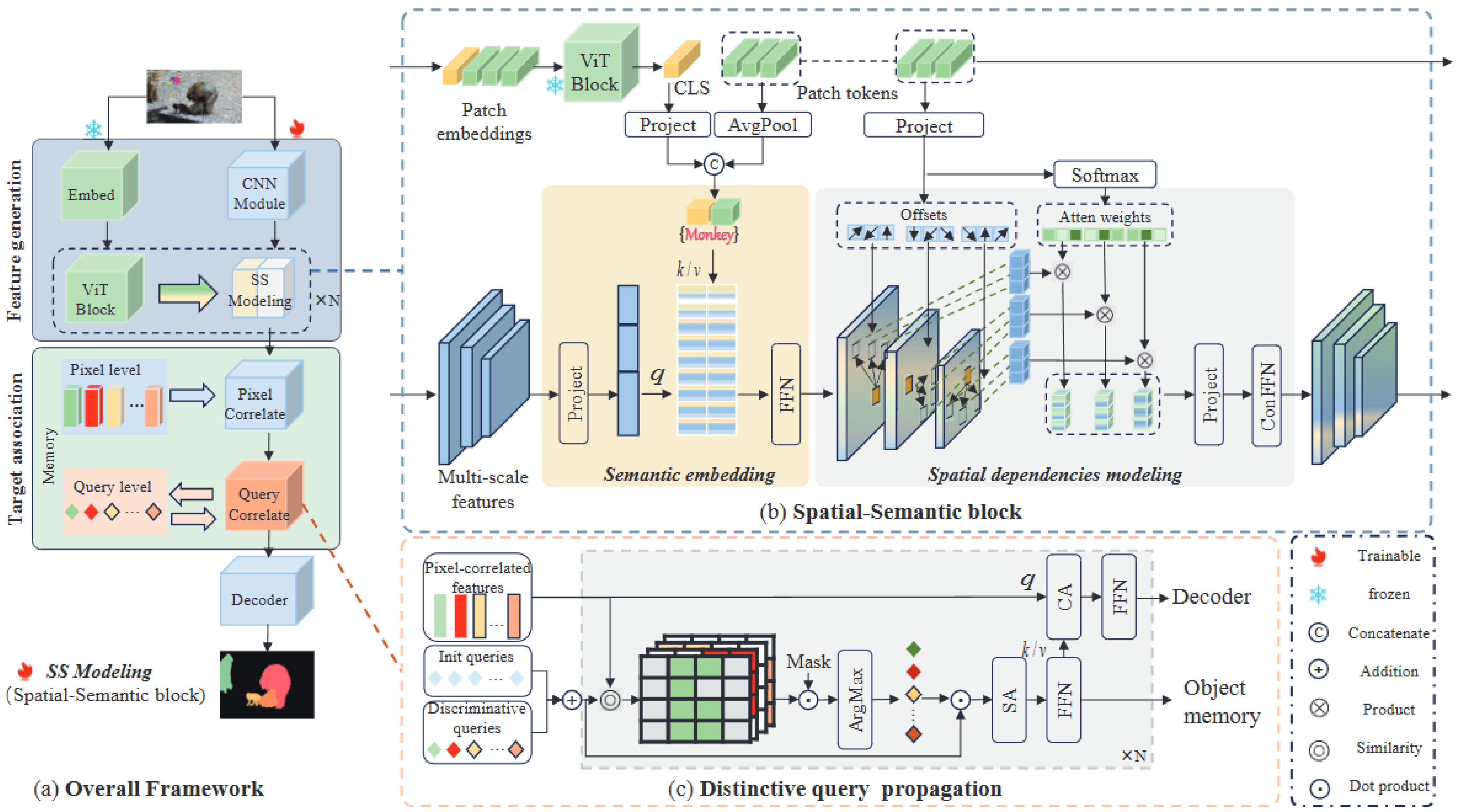
Config	DAVIS YTVOS	MEGA
optimizer	AdamW	AdamW
base learning rate	5e-5	5e-5
weight decay	0.05	0.05
droppath rate	0.15	0.15
batch size	16	16
num ref frames	3	3
num frames	8	8
max-skip	[5, 10, 15, 5]	[5, 10, 15, 5]
max-skip-itr	[0.1,0.3,0.8,1]	[0.1, 0.3, 0.8, 1]
Iterations	150,000	190,000
learning rate schedule	steplr	steplr

◆ Training with MEGA

- ✓ Trained on MEGA datasets, including YouTubeVOS, DAVIS, OVIS, MOSE and BURST.
- ✓ Test with base input size (480p) and larger input size (720p or 600p).

S3 Framework

◆ Overall framework:



(a) Overall Framework

(b) Spatial-Semantic block

(c) Distinctive query propagation

Visualization

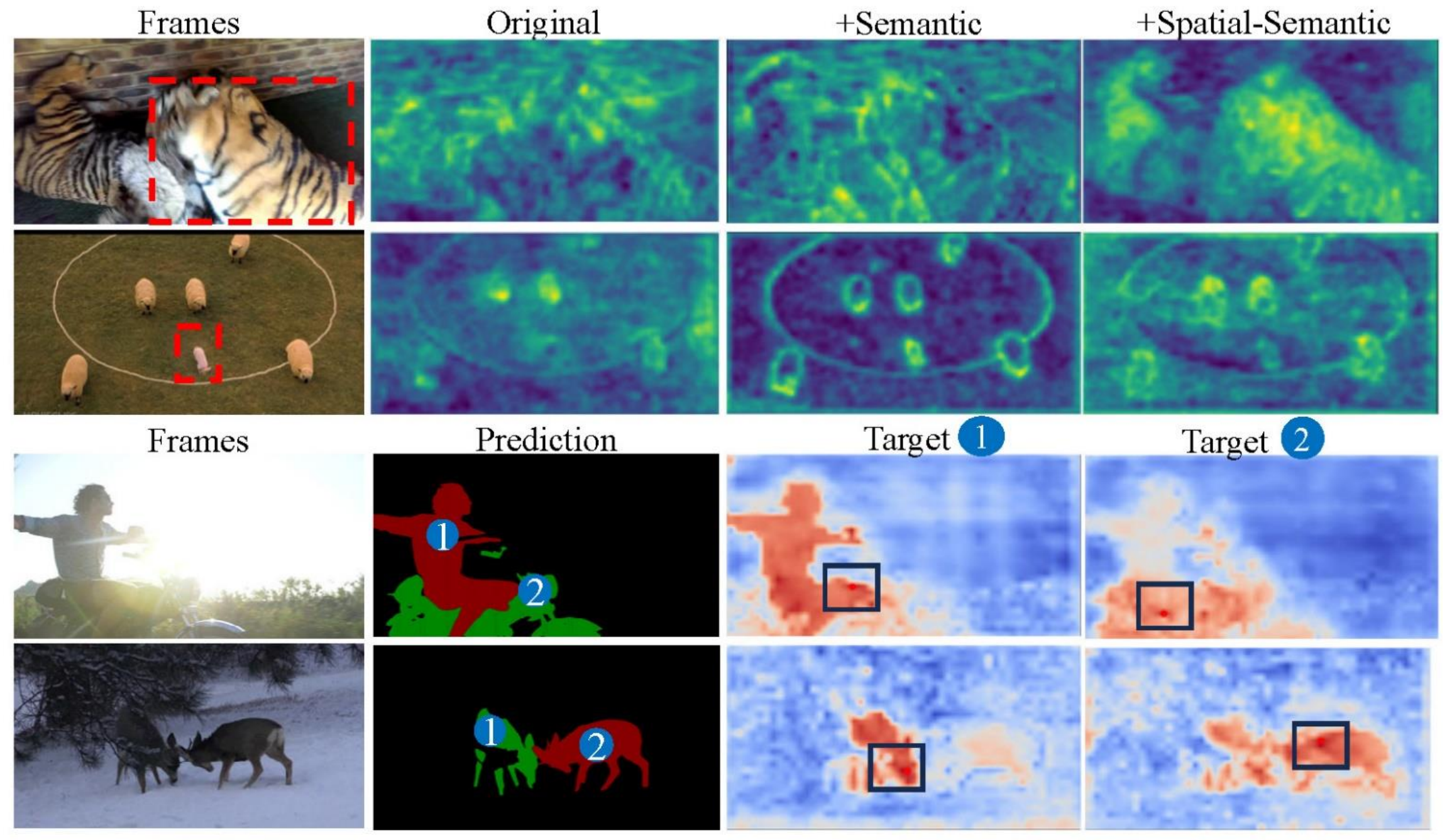
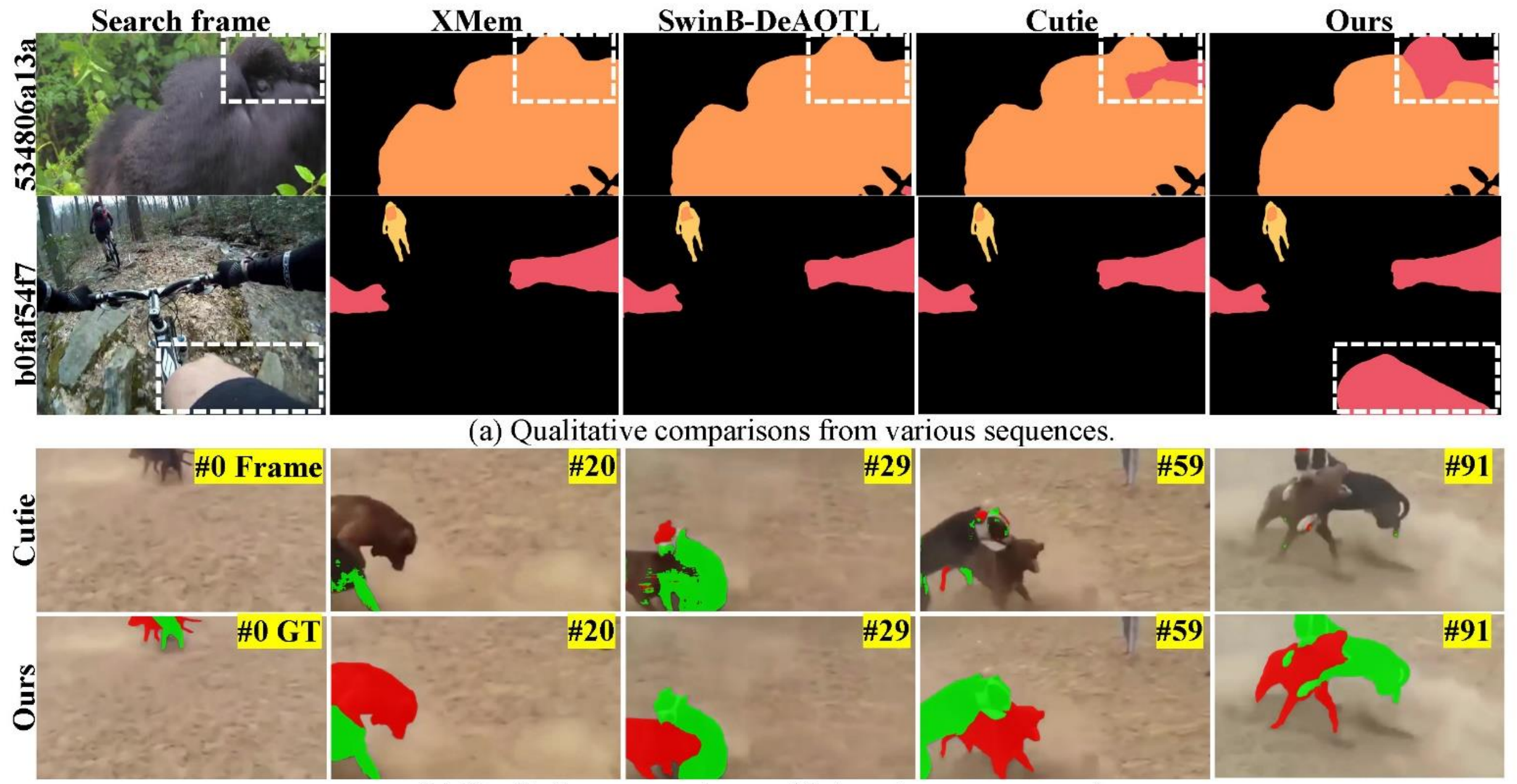


Figure 1. Visualization of feature maps from different blocks.



(a) Qualitative comparisons from various sequences.

(b) Qualitative comparisons with target appearance variance.

Experimental Results

Table 1. Comparison of S3 (Ours) and Current SOTA methods in terms of J&F in different VOS datasets. S3 achieves a now SOTA.

Dataset	MOSE-val			LVOS test			DAVIS 2017 test			YouTube-VOS 2018 val					YouTube-VOS 2019 val					FPS
Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	
Trained only on the YouTube VOS, and DAVIS datasets																				
MiVOS (Cheng et al., 2021b)*	-	-	-	-	-	-	78.6	74.9	82.2	82.6	81.1	85.6	77.7	86.2	82.4	80.6	84.7	78.1	86.4	-
STCN (Cheng et al., 2021a) *	52.5	48.5	56.6	45.8	41.1	50.5	77.8	74.3	81.3	84.3	83.2	87.9	79.0	87.3	84.2	82.6	87.0	79.4	87.7	13.2
Swin-B-AOT-L (Yang et al., 2021) *	59.4	53.6	65.2	54.4	49.3	59.4	81.2	77.3	85.1	85.1	85.1	90.1	78.4	86.9	85.3	84.6	89.5	79.3	87.7	12.1
DeAOT-R50 (Yang & Yang, 2022)	59.0	54.6	63.4	-	-	-	80.7	76.9	84.5	86.0	84.9	89.9	80.4	88.7	85.6	84.2	89.2	80.2	88.8	11.7
XMem (Cheng & Schwing, 2022)	-	-	-	-	-	-	79.8	76.3	83.4	84.3	83.9	88.8	77.7	86.7	84.2	83.8	88.3	78.1	86.7	
XMem (Cheng & Schwing, 2022) *	53.3	62.0	57.6	50.0	45.5	54.4	81.0	77.4	84.5	85.7	84.6	89.3	80.2	88.7	85.5	84.3	88.6	80.3	88.6	22.6
ISVOS (Wang et al., 2023) *	-	-	-	-	-	-	82.8	79.3	86.2	86.3	85.5	90.2	80.5	88.8	86.1	85.2	89.7	80.7	88.9	5.8
SimVOS-B (Wu et al., 2023)	61.6	57.9	65.3	-	-	-	82.3	78.7	85.8	-	-	-	-	-	84.2	83.1	87.5	79.1	87.2	3.3
Cutie (Cheng et al., 2023a)*	64.0	60.0	67.9	56.2	51.8	60.5	84.2	80.6	87.7	86.1	85.5	90.0	80.6	88.3	86.1	85.8	90.5	80.0	88.0	36.4
JointFormer (Zhang et al., 2023)	-	-	-	-	-	-	87.0	83.4	90.6	86.0	86.0	91.0	79.5	87.5	86.2	85.7	90.5	80.4	88.2	3.0
JointFormer (Zhang et al., 2023)*	-	-	-	-	-	-	87.6	84.2	91.1	87.0	86.2	91.0	81.4	89.3	87.0	86.1	90.6	82.0	89.5	3.0
S3 (Ours)	68.5	64.5	72.6	66.5	62.1	70.8	86.7	82.7	90.8	87.4	87.0	92.0	80.9	89.7	87.5	86.8	91.8	81.3	89.9	13.1
Trained on the MEGA dataset																				
DEVA (Cheng et al., 2023b)	66.5	62.3	70.8	54.0	49.0	59.0	83.2	79.6	86.8	86.2	85.4	89.9	80.5	89.1	85.8	84.8	89.2	80.3	88.8	25.3
Cutie (Cheng et al., 2023a) *	69.9	65.8	74.1	66.7	62.4	71.0	86.1	82.4	89.9	87.0	86.4	91.1	81.4	89.2	87.0	86.0	90.5	82.0	89.6	36.4
S3 (Ours)	74.0	69.8	78.3	73.0	68.3	77.8	87.8	84.0	91.7	88.0	87.0	91.8	82.5	90.7	88.1	87.4	92.5	81.9	90.7	13.1

Table 2. Comparison of S3 and current SOTA methods with different training and testing settings.

Dataset	MOSE-val			DAVIS 2017 test			YouTube-VOS 2018 val					YouTube-VOS 2019 val				
Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
Cutie-base (Cheng et al., 2023a)*	64.0	60.0	67.9	84.2	80.6	87.7	86.1	85.8	90.5	80.0	88.0	86.1	85.5	90.0	80.6	88.3
ISVOS (Wang et al., 2023)*+BL30K (Cheng et al., 2021b)	-	-	-	84.0	80.1	87.8	86.7	86.1	90.8	81.0	89.0	86.3	85.2	89.7	81.0	89.1
JointFormer (Zhang et al., 2023)*+BL30K (Cheng et al., 2021b)	-	-	-	88.1	84.7	91.6	87.6	86.4	91.0	82.2	90.7	87.4	86.5	90.9	82.0	90.3
Ours	68.5	64.5	72.6	87.1	83.1	91.1	87.4	87.0	92.0	80.9	89.7	87.5	86.8	91.8	81.3	89.9
Cutie-base (Cheng et al., 2023a)*+	66.2	62.3	70.1	85.9	82.6	89.2	-	-	-	-	-	86.9	86.2	90.7	81.6	89.2
Ours+	70.5	66.5	74.6	87.9	84.6	91.3	87.6	86.9	91.7	81.5	90.1	87.8	86.8	91.6	82.2	90.8
Cutie-base* (Cheng et al., 2023a) w/ MEGA	69.9	65.8	74.1	86.1	82.4	89.9	87.0	86.4	91.1	81.4	89.2	87.0	86.0	90.5	82.0	89.6
Ours w/MEGA	73.2	68.8	77.5	88.2	84.3	92.1	88.1	87.4	92.5	81.9	90.7	88.0	88.0	91.8	82.5	90.8
Cutie-base (Cheng et al., 2023a)*+ w/ MEGA	71.7	67.6	75.8	88.1	84.7	91.4	-	-	-	-	-	87.5	86.3	90.6	82.7	90.5
Ours+ w/MEGA	75.1	71.0	79.2	89.1	85.8	92.4	88.5	87.6	92.6	82.7	91.3	88.5	87.3	92.0	83.1	91.4

Table 3. Detailed ablation study about the proposed components, training and testing settings.

Dataset	MOSE-val			DAVIS 2017 test			LVOS test			YouTube-VOS 2019 val				
Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
Trained on the YouTubeVOS, and DAVIS datasets														
XMem (Cheng & Schwing, 2022) (Baseline)	53.3	62.0	57.6	81.0	77.4	84.5	50.0	45.5	54.4	85.5	84.3	88.6	80.3	88.6
Cutie (Cheng et al., 2023a)	64.0	60.0	67.9	84.2	80.6	87.7	56.2	51.8	60.5	86.1	85.8	90.5	80.0	88.0
+Discriminative Query	64.2	60.3	68.1	85.2	81.8	88.5	57.4	53.3	61.5	86.5	86.2	90.7	80.6	88.8
+ViT	64.2	60.2	68.3	85.6	82.0	89.2	58.3	53.7	62.8	86.7	86.4	90.3	81.0	88.7
+Spatial	68.2	64.0	72.4	86.2	82.4	90.1	67.4	62.9	71.9	87.3	86.4	91.3	81.5	90.3
+Semantic (Full)	68.5	64.5	72.6	86.7	82.7	90.8	66.5	62.1	70.8	87.5	86.8	91.8	81.3	89.9
Improved test size (600/720)														
Cutie (Cheng et al., 2023a)	66.2	62.3	70.1	85.9	82.6	89.2	56.2	51.8	60.5	86.9	86.2	90.7	81.6	89.2
+Discriminative Query	66.4	62.4	70.1	87.9	84.6	91.2	57.4	53.3	61.5	87.1	86.3	90.6	82.0	89.5
+Spatial	69.9	67.5	74.1	87.0	83.7	90.2	67.4	62.9	71.9	87.5	86.8	91.6	81.6	90.2
+Semantic (full)	70.5	66.5	74.6	87.8	84.6	91.3	66.5	62.1	70.8	87.7	86.8	91.6	82.2	90.8
Trained on the MEGA datasets														
Cutie (Cheng et al., 2023a)	69.9	65.8	74.1	86.1	82.4	89.9	66.7	62.4	71.0	87.0	86.0	90.5	82.0	90.7
+Discriminative query	70.6	66.5	74.6	86.6	82.7	90.5	66.5	62.1	70.8	87.5	86.0	90.6	82.8	90.6
+Spatial	73.5	69.1	77.7	87.6	83.8	91.5	68.8	64.4	73.1	87.9	86.9	91.8	82.3	90.4
+Semantic (Full)	74.0	69.8	78.3	87.8	84.0	91.7	73.0	68.3	77.8	88.1	87.4	92.5	81.9	90.7
Improved test size (600/720)														
Cutie (Cheng et al., 2023a)	71.7	67.6	75.8	88.1	84.7	91.4	66.7	62.4	71.0	87.5	86.3	90.6	82.7	90.5
+Discriminative query	71.6	67.7	75.5	88.1	84.6	91.4	66.5	62.1	70.8	88.0	86.2	90.6	83.6	91.5
+Spatial	75.3	71.3	79.2	89.0	85.8	92.3	68.8	64.4	73.1	88.3	87.0	91.9	83.0	91.1
+Semantic(Full)	75.1	71.0	79.2	89.1	85.8	92.4	73.0	68.3	77.8	88.5	87.3	92.0	83.1	91.4

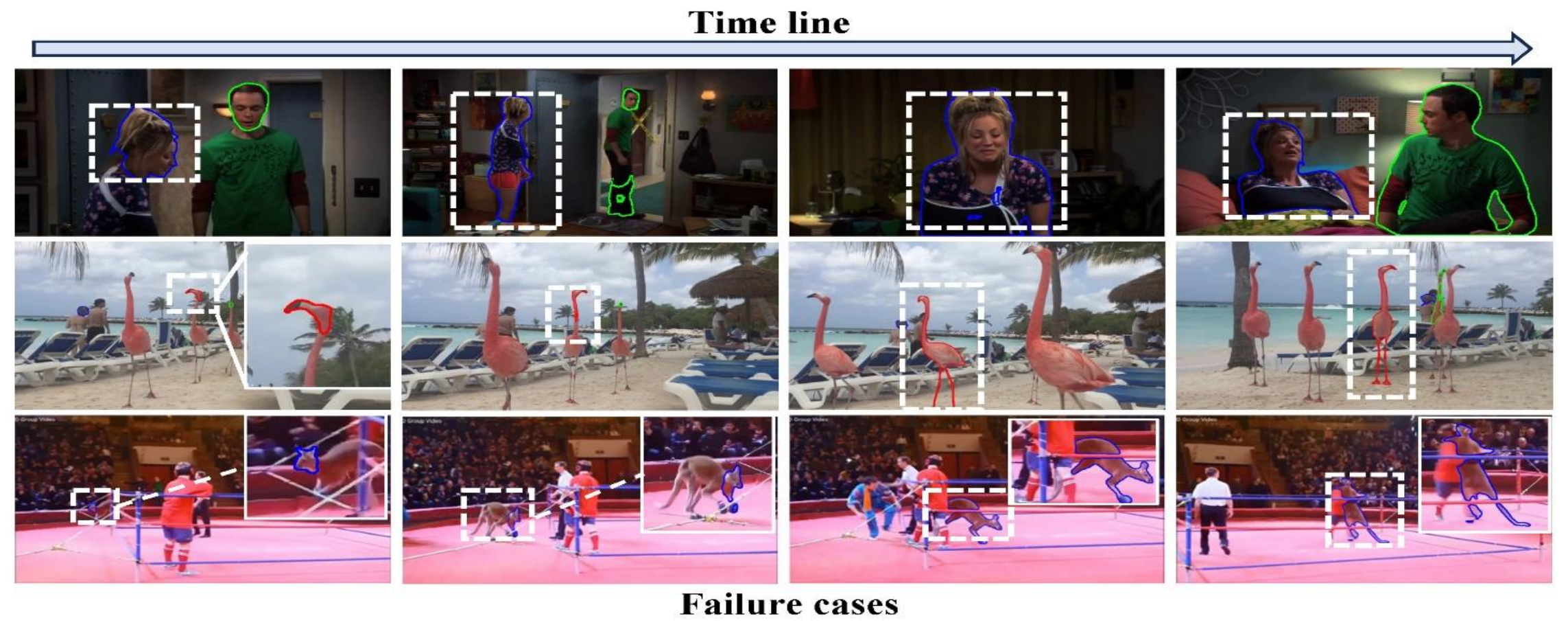


Figure 2. Results trained with base setting

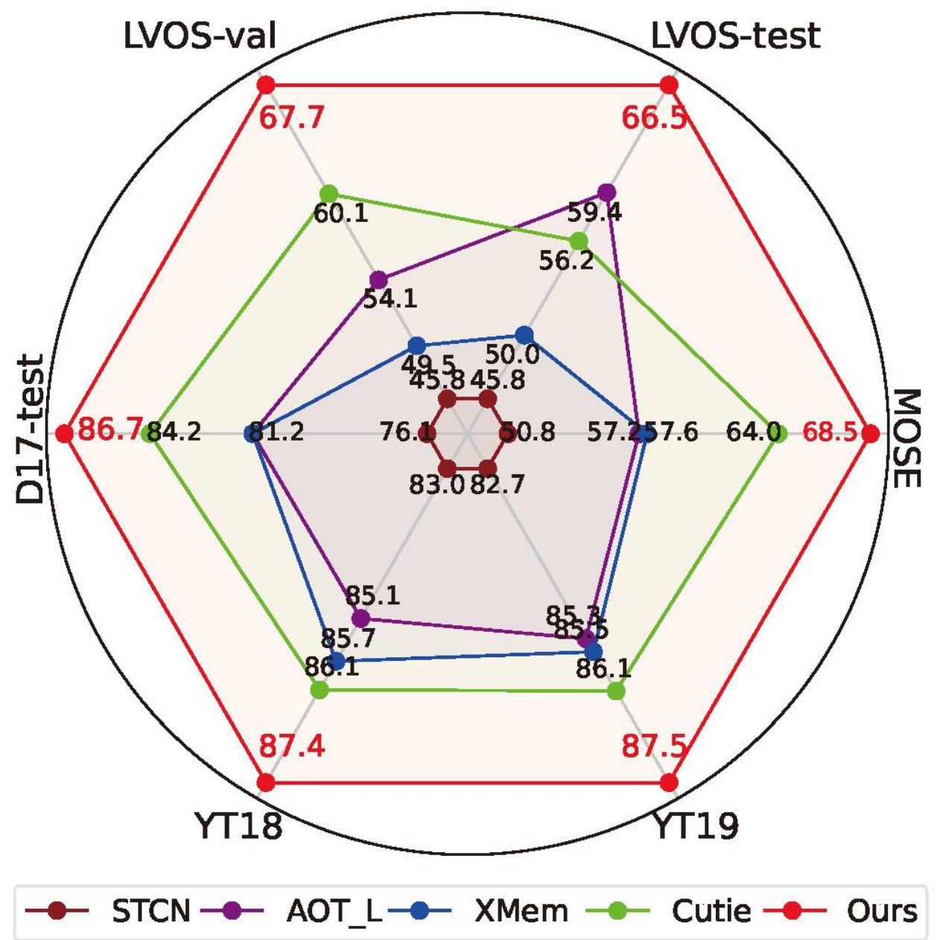


Table 4. Results trained only on MOSE.

Dataset	MOSE-val		
Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
RDE (Li et al., 2022a)	48.8	44.6	52.9
STCN (Cheng et al., 2021a)	50.8	46.6	55.0
AOT (Yang et al., 2021)	57.2	53.1	61.3
XMem (Cheng & Schwing, 2022)	57.6	53.3	62.0
DeAOT (Yang & Yang, 2022)	59.4	55.1	63.8
ResNet+Discriminative query	69.9	65.8	73.9
+Spatial	72.7	68.3	77.0
+Semantic	72.9	68.4	77.3
Improved test size (720)			
ResNet+Discriminative query	71.6	67.7	75.5
+Spatial	74.0	69.9	78.1
+Semantic	74.5	70.2	78.5