

The Fourth Visual Object Tracking and Segmentation Challenge

VOTSt and VOTSr Tracks



Outline

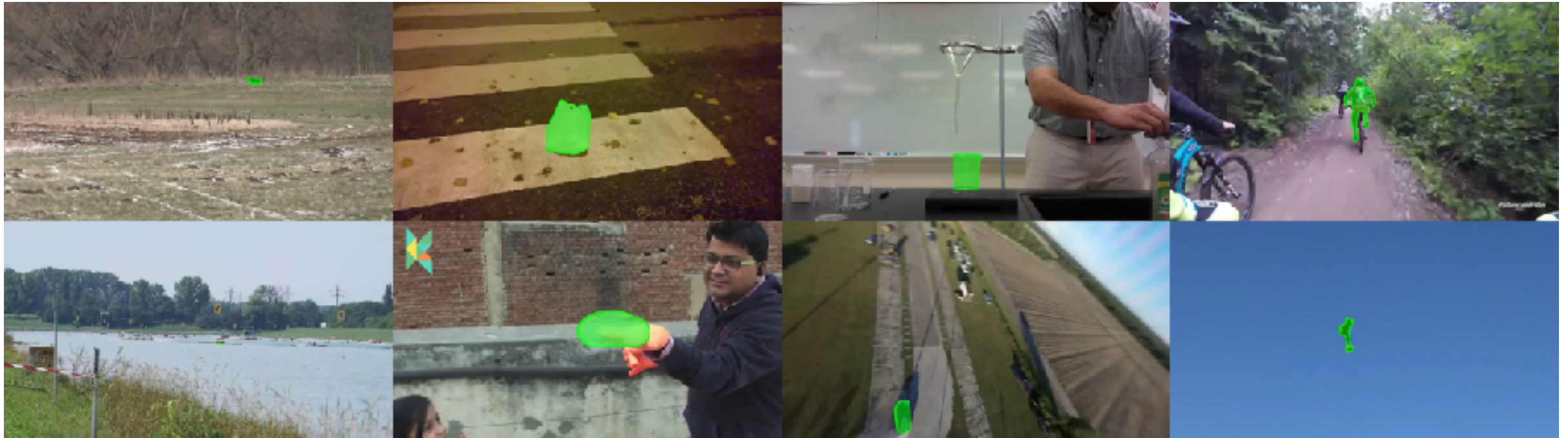
- VOTSt dataset and challenge
- VOTSt 2026 Challenge results
- VOTSr dataset and challenge
- VOTSr 2026 Challenge results
- Conclusions

Outline

- **VOTSt dataset and challenge**
- VOTSt 2026 Challenge results
- VOTSr dataset and challenge
- VOTSr 2026 Challenge results
- Conclusions

VOTS is the key tracking challenge

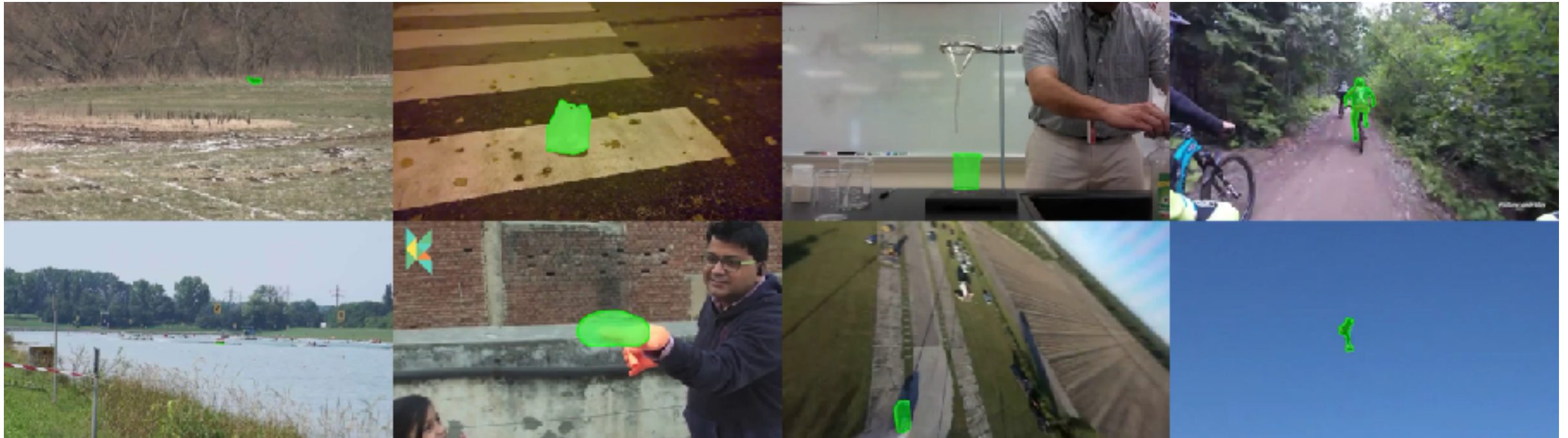
Focusing on core challenges (multi-object, long videos, target disappearance)



VOTS Challenge (2023)

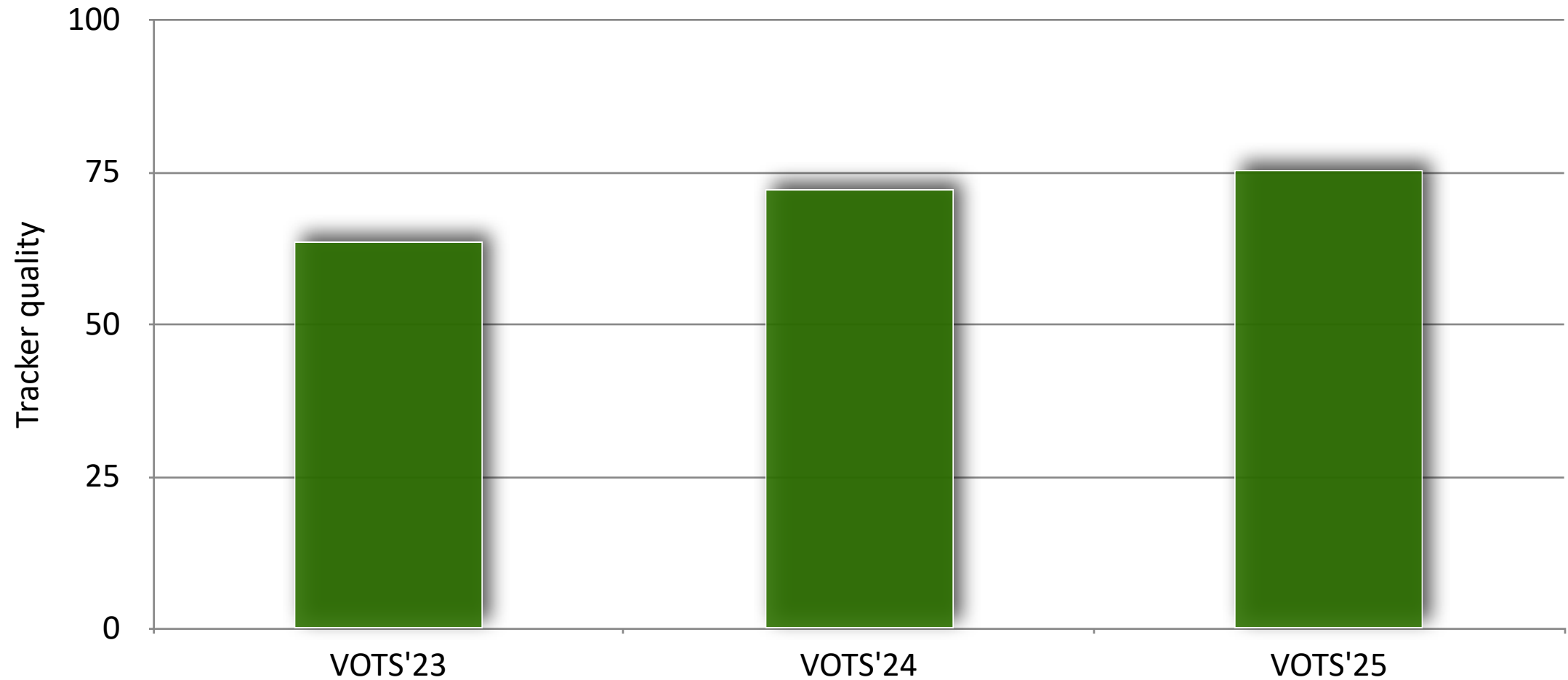
VOTS is the key tracking challenge

Focusing on core challenges (multi-object, long videos, target disappearance)



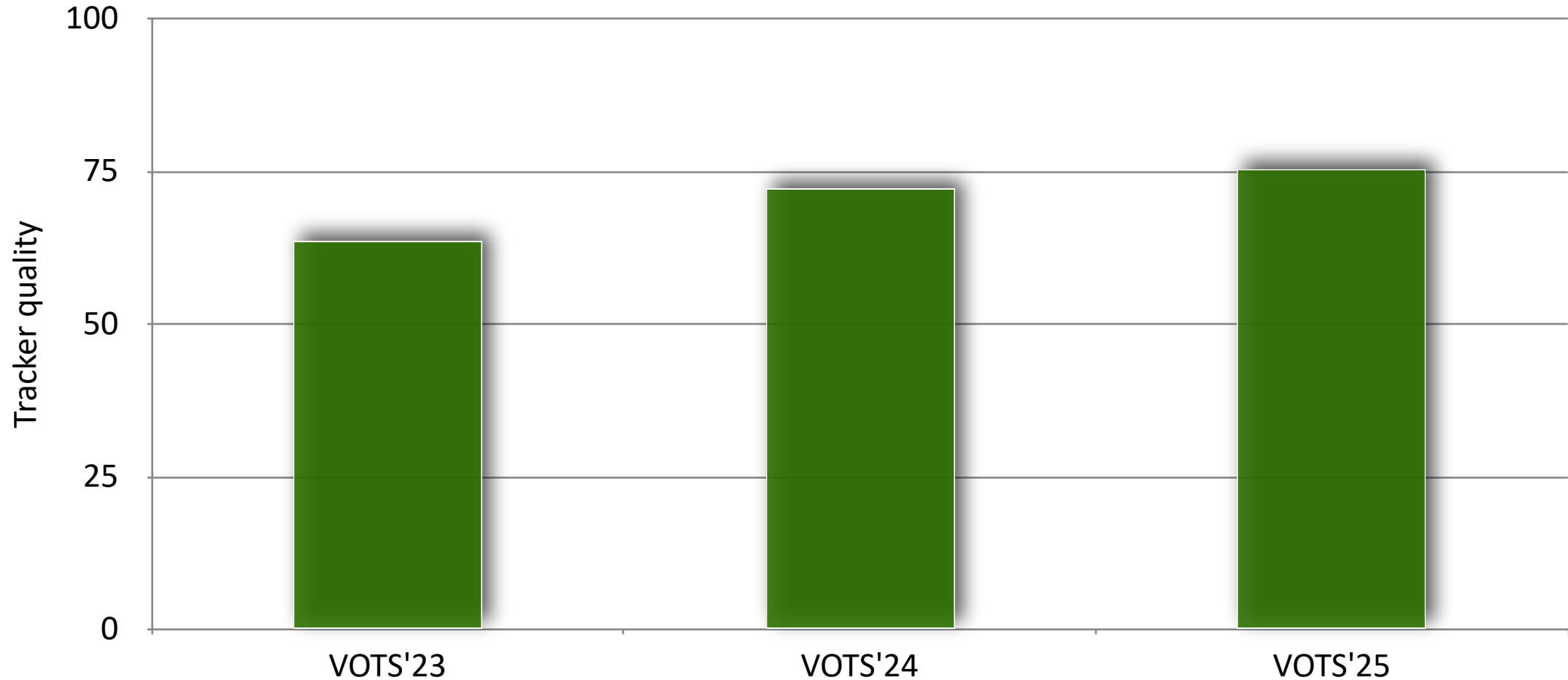
VOTS Challenge (2023)

VOTS winners



VOTS winners

Steady progress year after year!

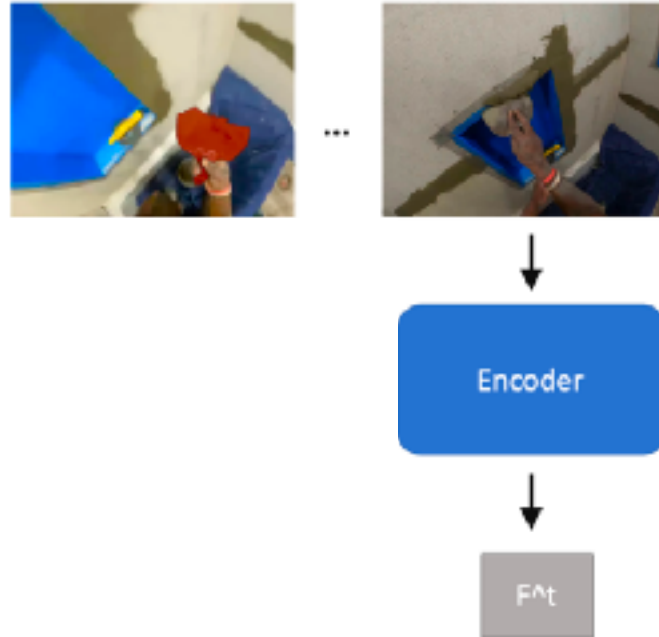


How do VOTS methods work?



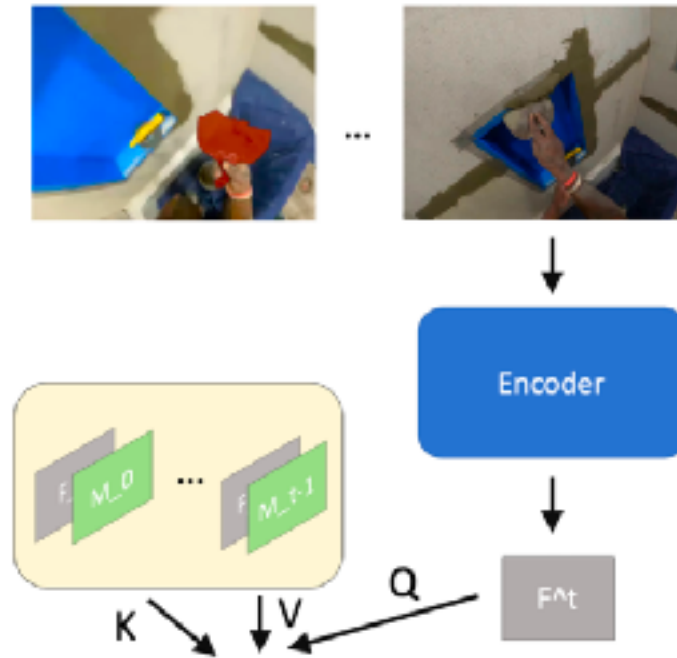
Memory-based VOS paradigm

How do VOTS methods work?



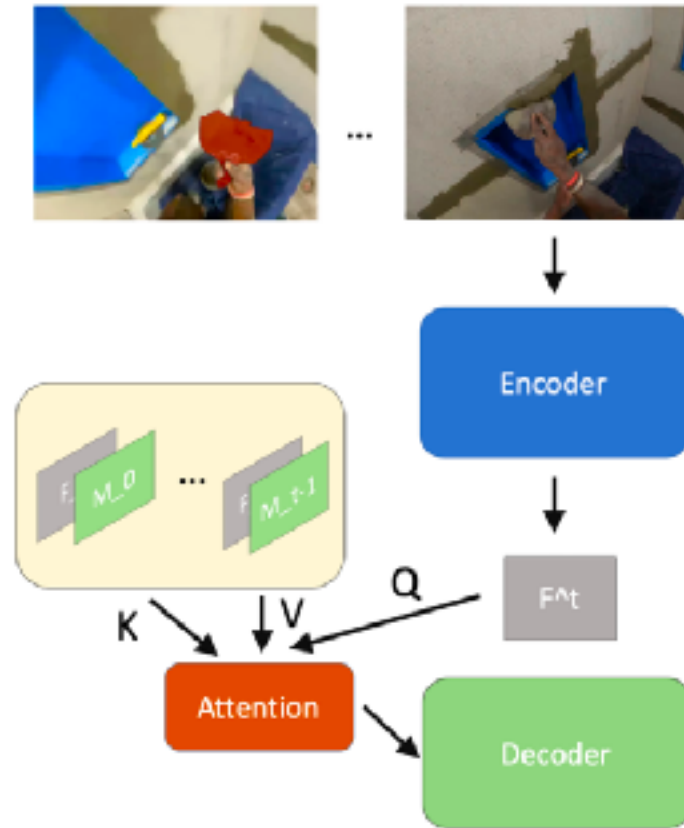
Memory-based VOS paradigm

How do VOTS methods work?



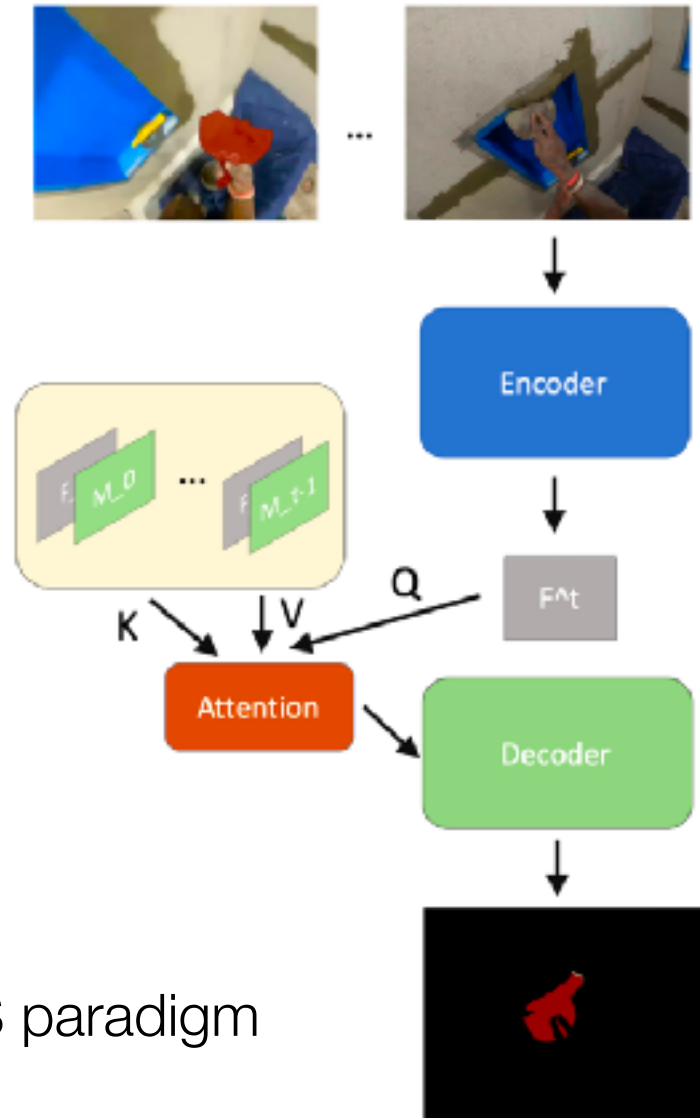
Memory-based VOS paradigm

How do VOTS methods work?



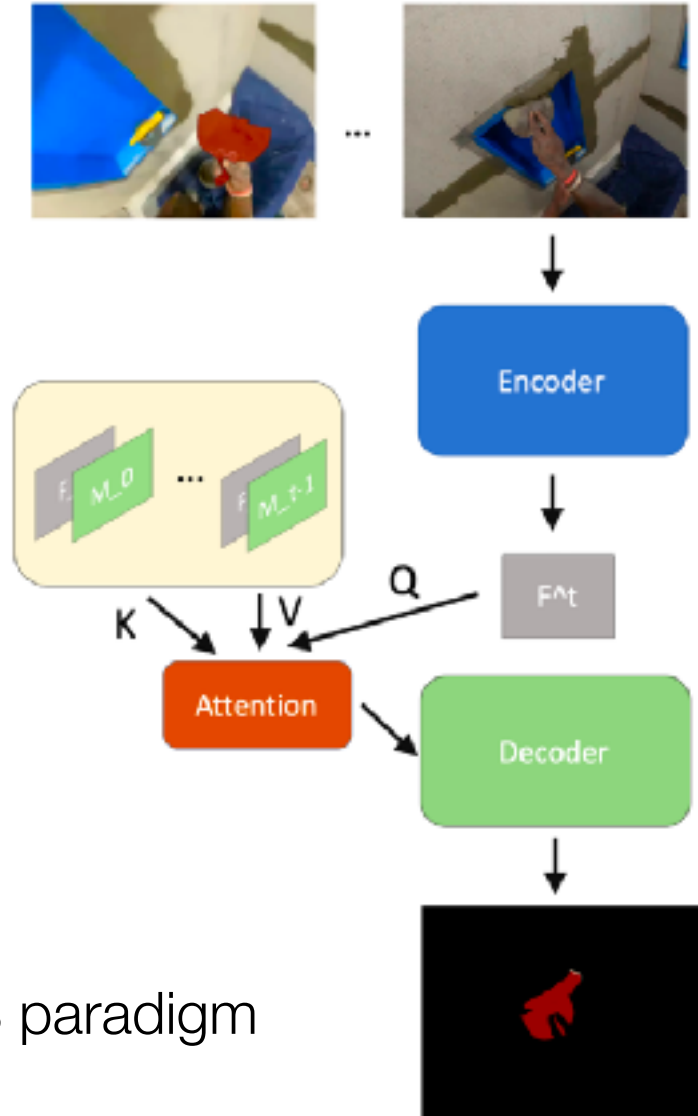
Memory-based VOS paradigm

How do VOTS methods work?



Memory-based VOS paradigm

How do VOTS methods work?

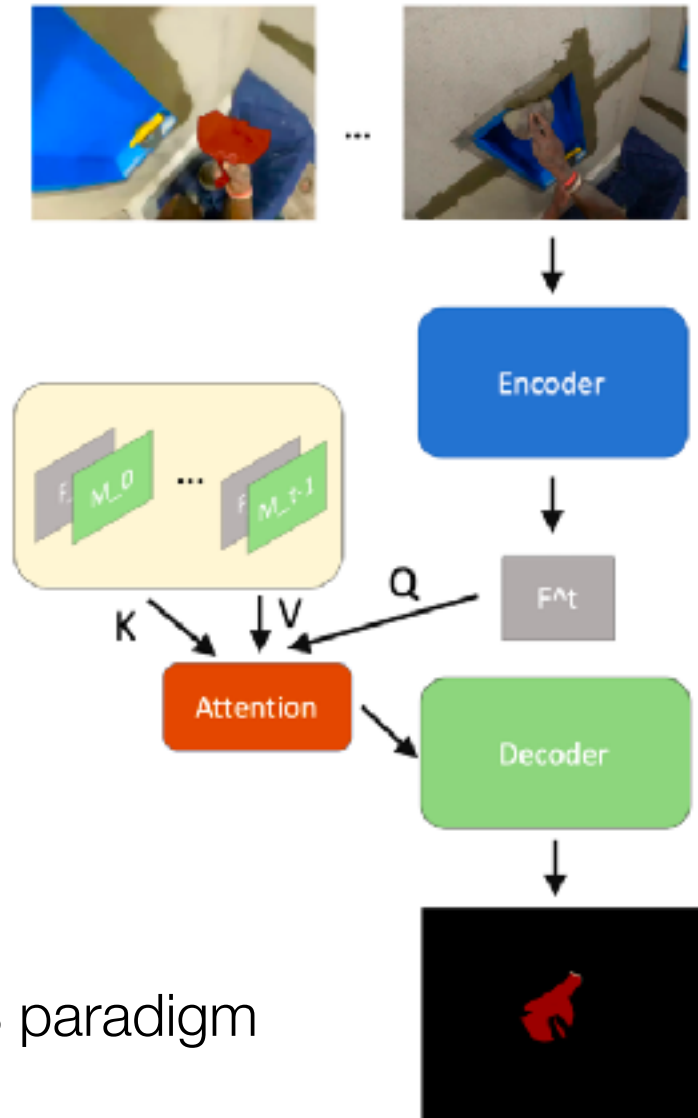


Memory-based VOS paradigm



AOT [Yang et al., NeurIPS'21]

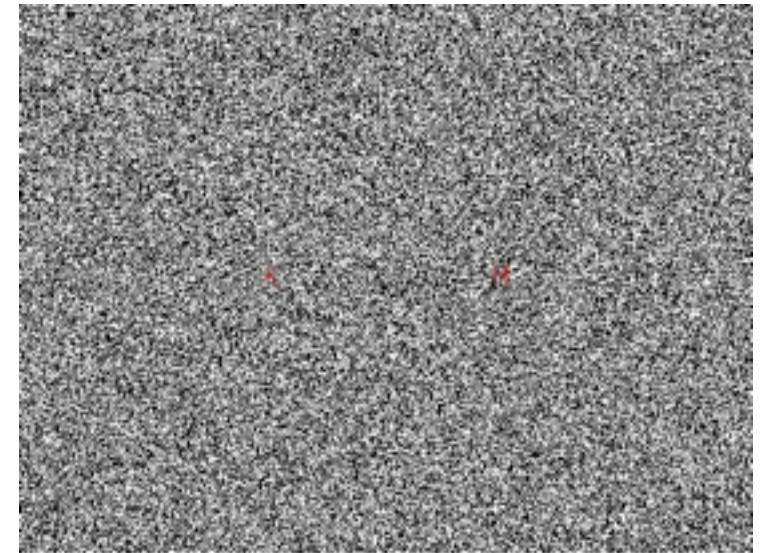
How do VOTS methods work?



Memory-based VOS paradigm



AOT [Yang et al., NeurIPS'21]



[Gao and Sholl, Visual Cognition'10]

Video Object Segmentation under Transformations (VOST)



Annotations visualization at 5 fps

Cut corn

VOST ground-truth annotations

Video Object Segmentation under Transformations (VOST)



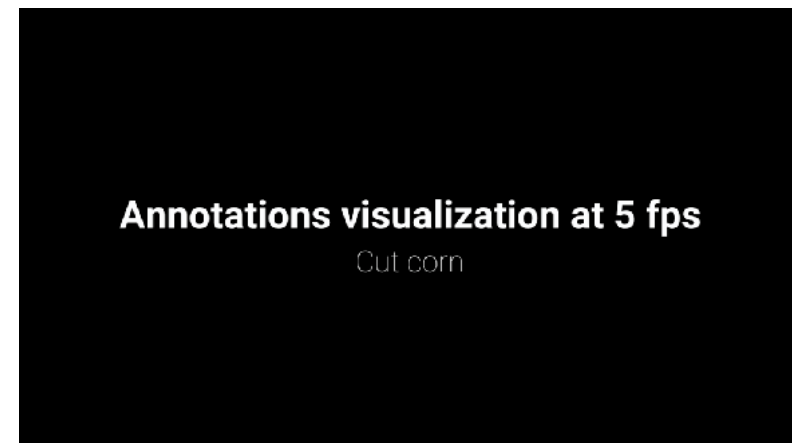
Annotations visualization at 5 fps

Cut corn

VOST ground-truth annotations

Video Object Segmentation under Transformations (VOST)

- 713 videos (572 train / 70 val / 71 test)
- Average video length of 21.2 seconds, labeled at 5 fps
- 75 547 frames
- 51 transformations over 154 object categories
- Held-out test set



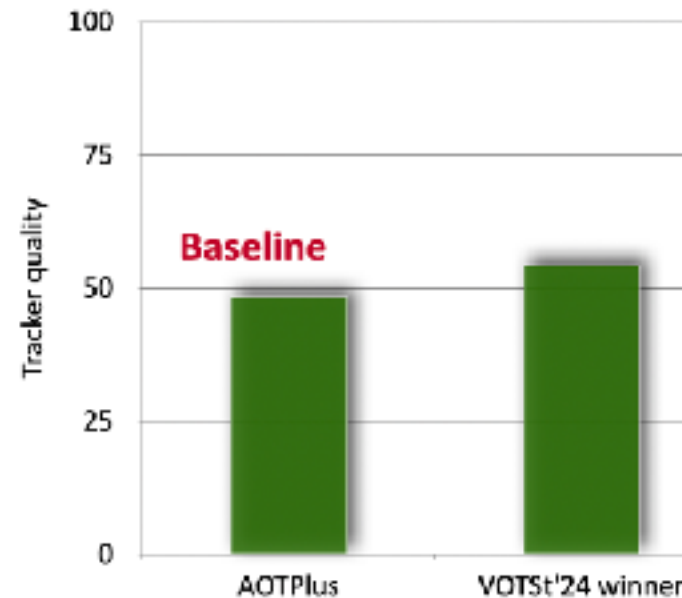
VOST ground-truth annotations

The challenge series

- **5th LSVOS Challenge at ICCV 2023**
 - Joint challenge for YouTube-VOS and VOST

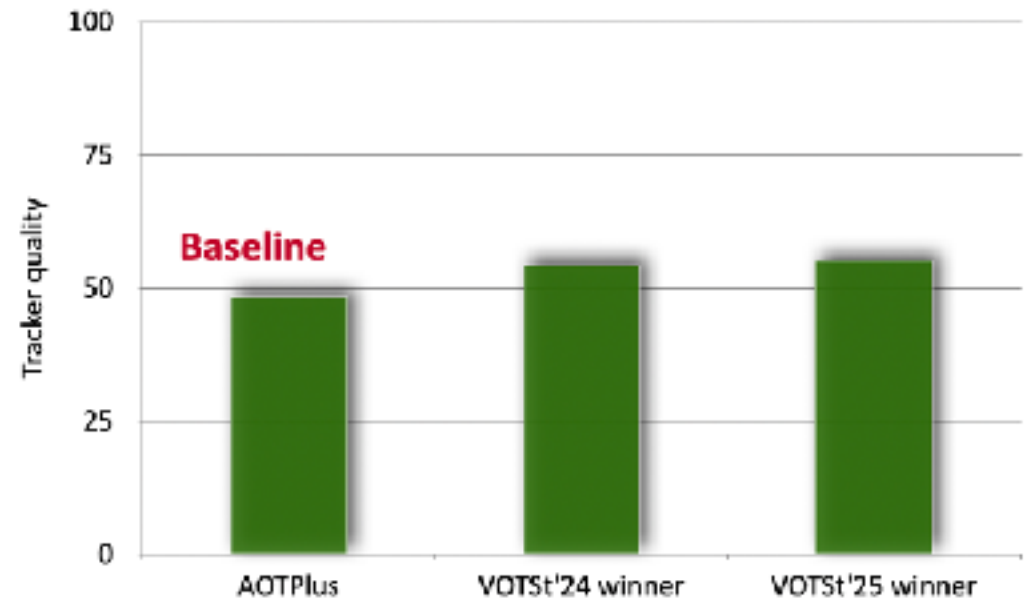
The challenge series

- **5th LSVOS Challenge at ICCV 2023**
 - Joint challenge for YouTube-VOS and VOST
- **2nd VOTS Challenge at ECCV 2024**



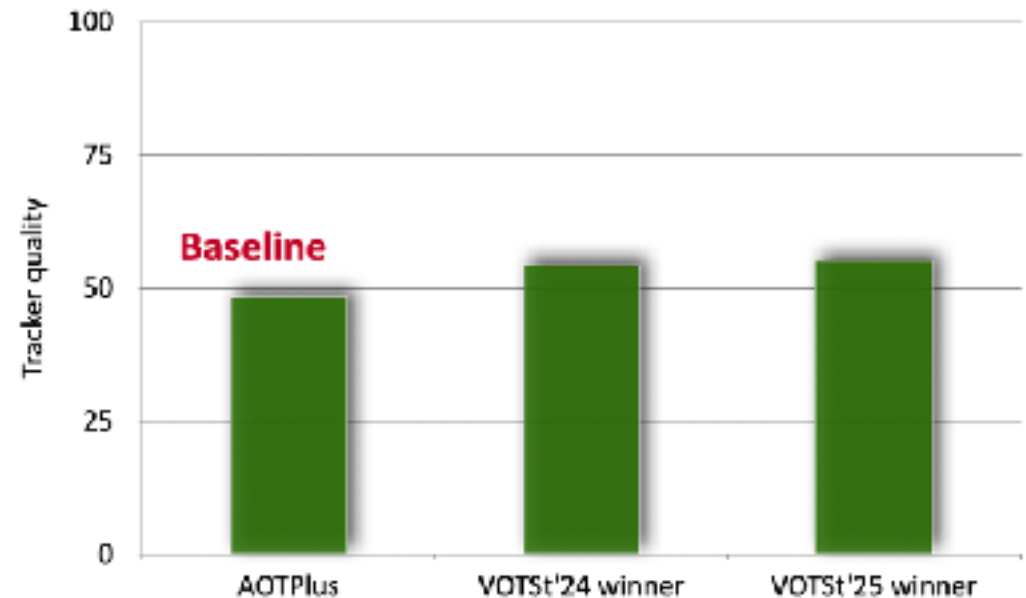
The challenge series

- **5th LSVOS Challenge at ICCV 2023**
 - Joint challenge for YouTube-VOS and VOST
- **2nd VOTS Challenge at ECCV 2024**
- **3rd VOTS Challenge at ICCV 2025**



The challenge series

- **5th LSVOS Challenge at ICCV 2023**
 - Joint challenge for YouTube-VOS and VOST
- **2nd VOTS Challenge at ECCV 2024**
- **3rd VOTS Challenge at ICCV 2025**
- **4th VOTS Challenge at ECCV 2026**
 - This workshop



Outline

- VOTSt dataset and challenge
- **VOTSt 2026 Challenge results**
- VOTSr dataset and challenge
- VOTSr 2026 Challenge results
- Conclusions

Participation

- 16 registered participants
- 8 valid submissions
- All rely on the memory-based paradigm
- Code publicly available for all the valid submissions

VOTSt track leaderboard

Tracker	Q	A	R	NRE	DRE	ADQ
▲ musmem	0.545①	0.543①	0.779②	0.119	0.102①	0.335
⊕ ccm4sam_vost	0.520②	0.531②	0.798①	0.077②	0.125	0.177
● deaotsam	0.511③	0.517③	0.747	0.132	0.121③	0.453②
● RMem	0.504	0.517	0.775③	0.092③	0.133	0.341
▶ SAM2TC	0.483	0.467	0.684	0.213	0.103②	0.429③
■ SAM3DirectMemSel	0.464	0.479	0.748	0.076①	0.176	0.357
✖ VOTSTracker	0.373	0.359	0.599	0.103	0.298	0.386
★ SAM2VOTStPlus	0.372	0.388	0.612	0.166	0.222	0.488①

VOTSt track leaderboard

Three submissions outperformed RMem baseline

Tracker	Q	A	R	NRE	DRE	ADQ
▲ musmem	0.545①	0.543①	0.779②	0.119	0.102①	0.335
⊕ ccm4sam_vost	0.520②	0.531②	0.798①	0.077②	0.125	0.177
● deaotsam	0.511③	0.517③	0.747	0.132	0.121③	0.453②
● RMem	0.504	0.517	0.775③	0.092③	0.133	0.341
▶ SAM2TC	0.483	0.467	0.684	0.213	0.103②	0.429③
■ SAM3DirectMemSel	0.464	0.479	0.748	0.076①	0.176	0.357
✖ VOTSTracker	0.373	0.359	0.599	0.103	0.298	0.386
★ SAM2VOTStPlus	0.372	0.388	0.612	0.166	0.222	0.488①

VOTSt track leaderboard

Winning method focuses on careful memory design

Tracker	Q	A	R	NRE	DRE	ADQ
▲musmem	0.545①	0.543①	0.779②	0.119	0.102①	0.335
⊕ccm4sam_vost	0.520②	0.531②	0.798①	0.077②	0.125	0.177
●deaotsam	0.511③	0.517③	0.747	0.132	0.121③	0.453②
●RMem	0.504	0.517	0.775③	0.092③	0.133	0.341
▶SAM2TC	0.483	0.467	0.684	0.213	0.103②	0.429③
■SAM3DirectMemSel	0.464	0.479	0.748	0.076①	0.176	0.357
✖VOTSTracker	0.373	0.359	0.599	0.103	0.298	0.386
★SAM2VOTStPlus	0.372	0.388	0.612	0.166	0.222	0.488①

VOTSt track leaderboard

SAM-based methods underperform

Tracker	Q	A	R	NRE	DRE	ADQ
▲musmem	0.545①	0.543①	0.779②	0.119	0.102①	0.335
✚ccm4sam_vost	0.520②	0.531②	0.798①	0.077②	0.125	0.177
●deaotsam	0.511③	0.517③	0.747	0.132	0.121③	0.453②
●RMem	0.504	0.517	0.775③	0.092③	0.133	0.341
▶SAM2TC	0.483	0.467	0.684	0.213	0.103②	0.429③
■SAM3DirectMemSel	0.464	0.479	0.748	0.076①	0.176	0.357
✖VOTSTracker	0.373	0.359	0.599	0.103	0.298	0.386
★SAM2VOTStPlus	0.372	0.388	0.612	0.166	0.222	0.488①



VOTSt2026 challenge

Winners:

musmem by: [Elham Soltan Kazemi](#), [H. Rahmon](#), [I. Toubal](#), [J. Mogollon](#),
[K. Palaniappan](#)

“Associating Visual Objects Through Transformers”

Winners' talk next in Session II

Outline

- VOTSt dataset and challenge
- VOTSt 2026 Challenge results
- **VOTSr dataset and challenge**
- VOTSr 2026 Challenge results
- Conclusions

Referring Video Object Segmentation



A bear moving towards left

[Khoreva et al., ACCV'18]

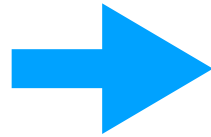
Referring Video Object Segmentation



A bear moving towards left

[Khoreva et al., ACCV'18]

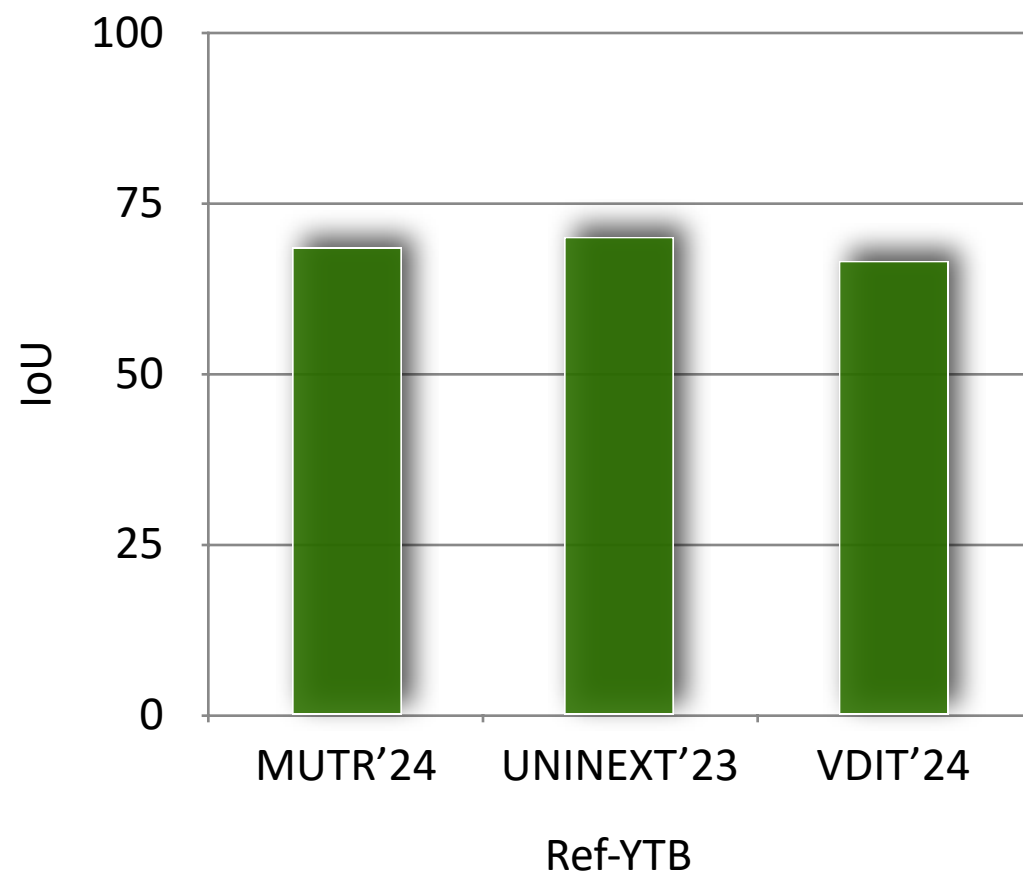
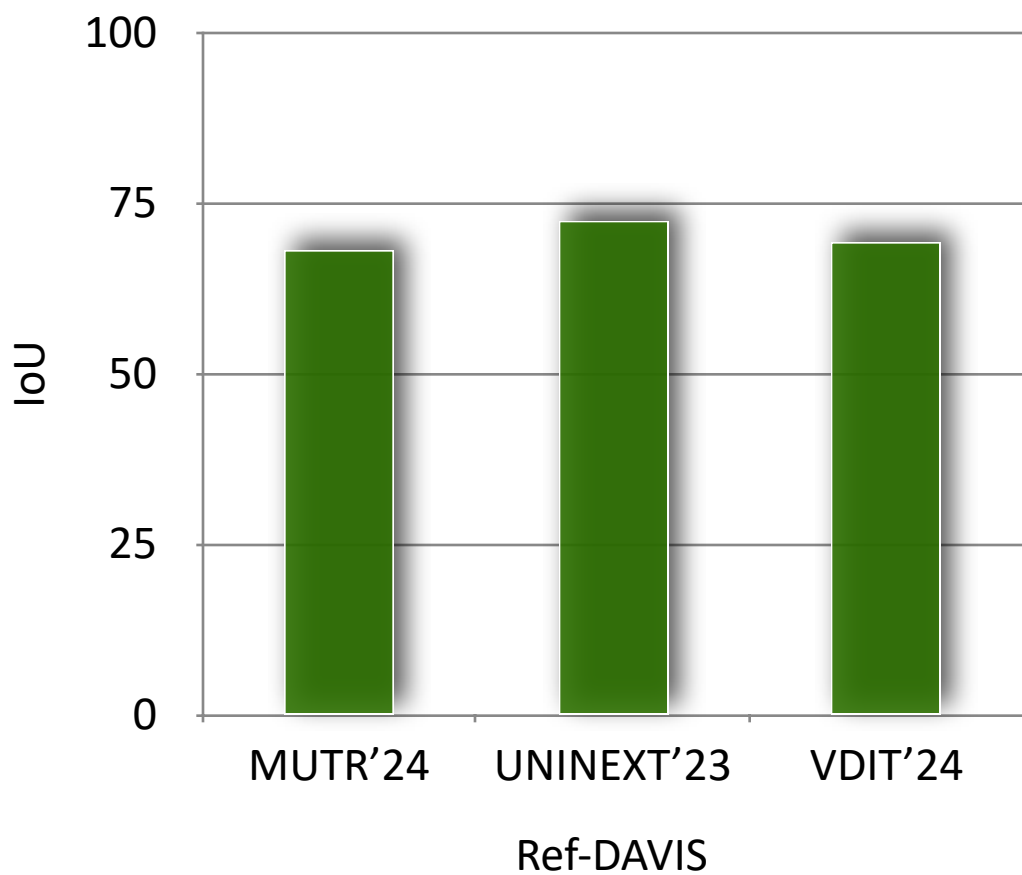
Referring Video Object Segmentation



A bear moving towards left

[Khoreva et al., ACCV'18]

Progress on existing benchmarks



Why limit the problem to objects?

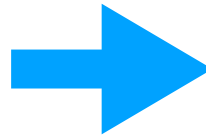


Why limit the problem to objects?



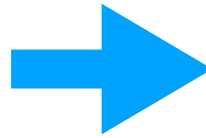
Why limit the problem to objects?

The woman smoking a cigarette



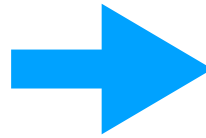
Why limit the problem to objects?

The cigarette being smoked by a woman



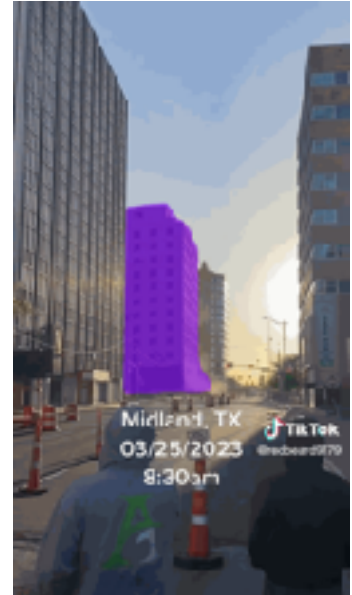
Why limit the problem to objects?

The smoke exhaled by a woman



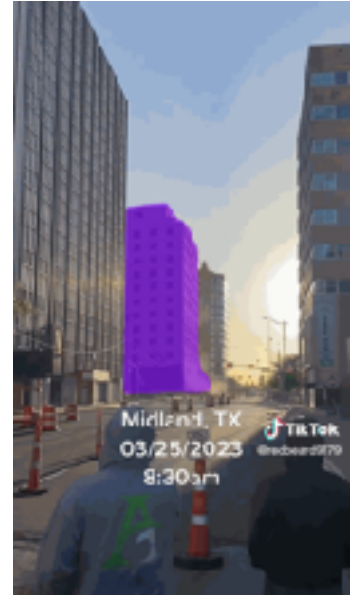
Referring Video Process Segmentation (Ref-VSP)

- 141 videos (0 train / 70 val / 71 test)
- 39 unique, dynamic concepts
- Each video labeled by 2 annotators (4 expressions)
- Held-out test set for the challenge



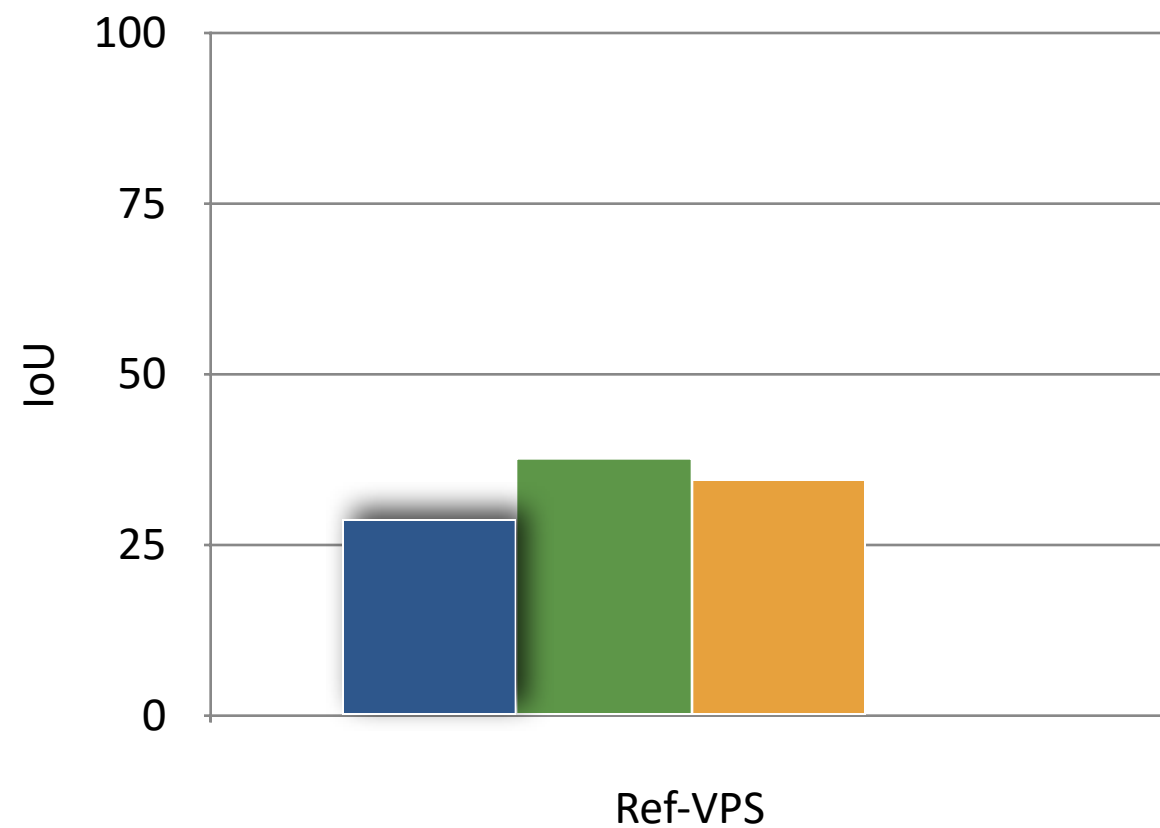
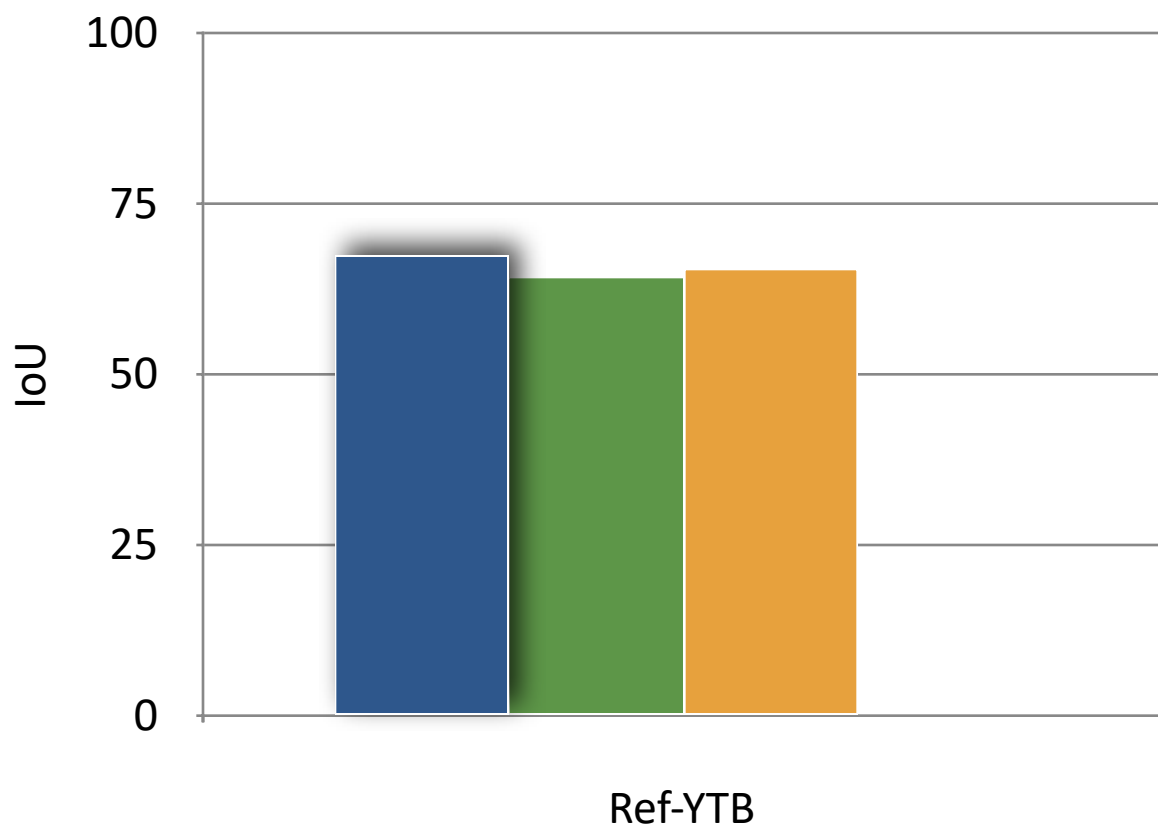
Referring Video Process Segmentation (Ref-VSP)

- 141 videos (0 train / 70 val / 71 test)
- 39 unique, dynamic concepts
- Each video labeled by 2 annotators (4 expressions)
- Held-out test set for the challenge



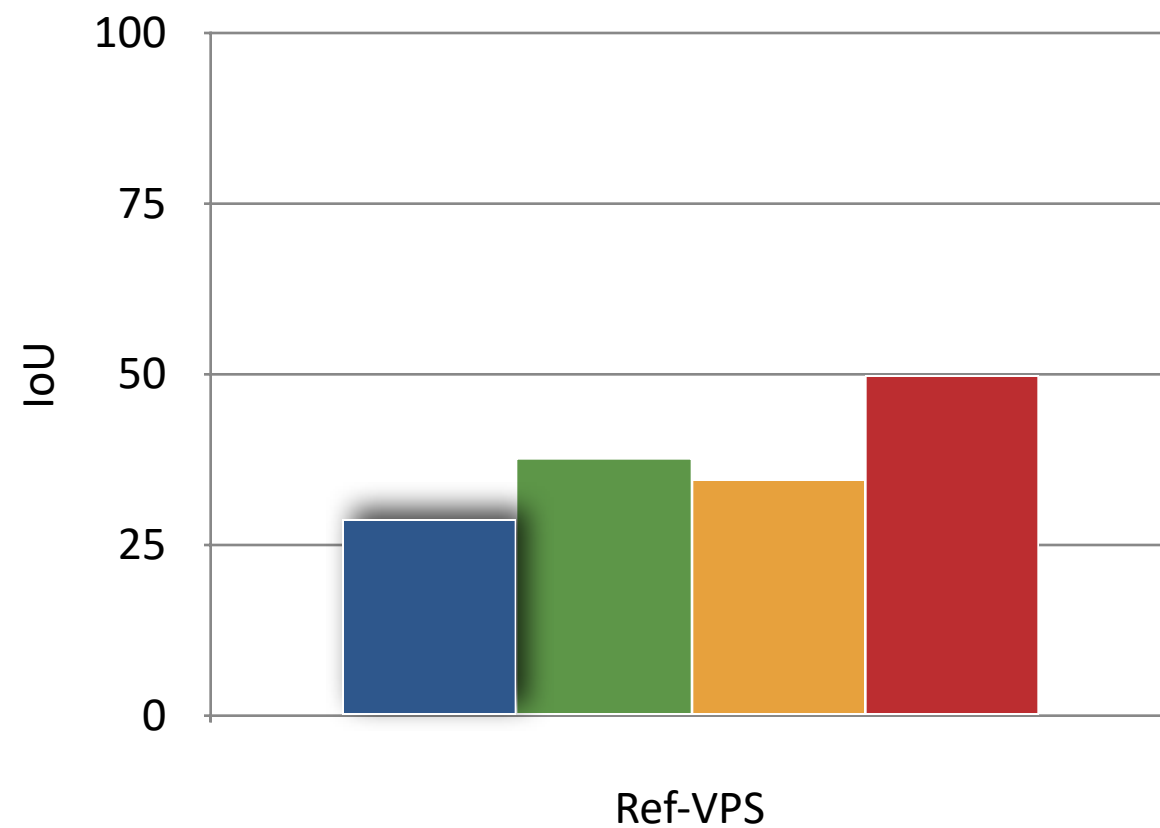
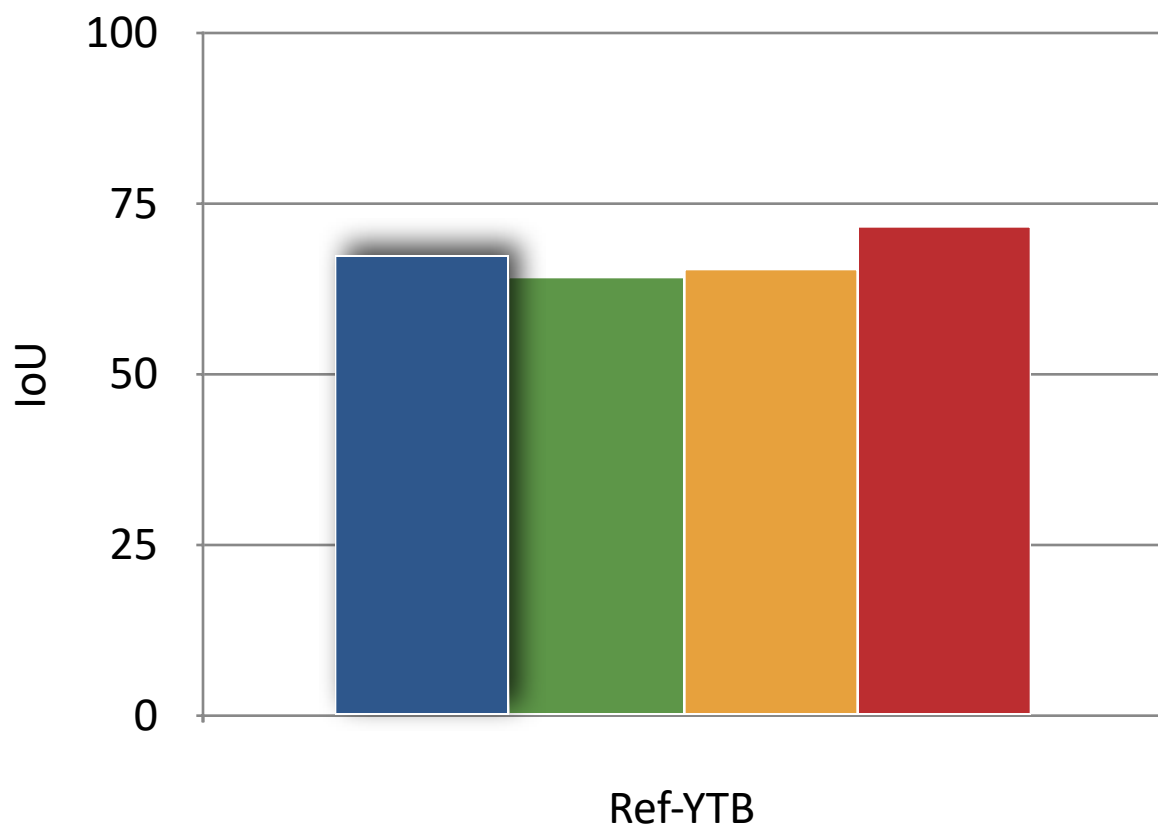
Complexity of Ref-VPS

UNINEXT'23
VDIT'24
GLUS (SAM2)'25



Complexity of Ref-VPS

- UNINEXT'23
- VDIT'24
- GLUS (SAM2)'25
- REM (Ours)



Outline

- VOTSt dataset and challenge
- VOTSt 2026 Challenge results
- VOTSr dataset and challenge
- **VOTSr 2026 Challenge results**
- Conclusions

Participation

- 15 registered participants
- 6 valid submissions
- Most rely heavily on Internet-scale VLM
- Code publicly available for all the valid submissions

VOTSr track leaderboard

Tracker	Q	A	R	NRE	DRE	ADQ
▶ ReflexTrack	0.697①	0.662①	0.982①	0.008①	0.010①	0.615①
▲ RSTracker	0.652②	0.634②	0.965②	0.016②	0.019	0.477
★ seed2lite_sam3	0.619③	0.585③	0.913③	0.073	0.014③	0.330
■ VLM-SAFEDAM	0.570	0.528	0.894	0.094	0.012②	0.523③
● REM	0.500	0.462	0.839	0.123	0.039	0.588②
● CLIPSegSAM2VOTSr	0.278	0.276	0.746	0.058③	0.196	0.413
⊕ Qwen7bSam3Hybrid	0.049	0.064	0.490	0.253	0.257	0.504

VOTSr track leaderboard

Four out of six submissions outperformed the baseline

Tracker	Q	A	R	NRE	DRE	ADQ
▶ ReflexTrack	0.697①	0.662①	0.982①	0.008①	0.010①	0.615①
▲ RSTracker	0.652②	0.634②	0.965②	0.016②	0.019	0.477
★ seed2lite_sam3	0.619③	0.585③	0.913③	0.073	0.014③	0.330
■ VLM-SAFEDAM	0.570	0.528	0.894	0.094	0.012②	0.523③
● REM	0.500	0.462	0.839	0.123	0.039	0.588②
● CLIPSegSAM2VOTSr	0.278	0.276	0.746	0.058③	0.196	0.413
⊕ Qwen7bSam3Hybrid	0.049	0.064	0.490	0.253	0.257	0.504

VOTSr track leaderboard

Winning method improves the Q score by 20 points!

Tracker	Q	A	R	NRE	DRE	ADQ
▶ ReflexTrack	0.697①	0.662①	0.982①	0.008①	0.010①	0.615①
▲ RSTracker	0.652②	0.634②	0.965②	0.016②	0.019	0.477
★ seed2lite_sam3	0.619③	0.585③	0.913③	0.073	0.014③	0.330
■ VLM-SAFEDAM	0.570	0.528	0.894	0.094	0.012②	0.523③
● REM	0.500	0.462	0.839	0.123	0.039	0.588②
● CLIPSegSAM2VOTSr	0.278	0.276	0.746	0.058③	0.196	0.413
⊕ Qwen7bSam3Hybrid	0.049	0.064	0.490	0.253	0.257	0.504

VOTSr track leaderboard

Training-free approach: VLM agent for grounding and SAM3 for segmentation

Tracker	Q	A	R	NRE	DRE	ADQ
▶ ReflexTrack	0.697①	0.662①	0.982①	0.008①	0.010①	0.615①
▲ RSTracker	0.652②	0.634②	0.965②	0.016②	0.019	0.477
★ seed2lite_sam3	0.619③	0.585③	0.913③	0.073	0.014③	0.330
■ VLM-SAFEDAM	0.570	0.528	0.894	0.094	0.012②	0.523③
● REM	0.500	0.462	0.839	0.123	0.039	0.588②
● CLIPSegSAM2VOTSr	0.278	0.276	0.746	0.058③	0.196	0.413
⊕ Qwen7bSam3Hybrid	0.049	0.064	0.490	0.253	0.257	0.504



VOTSr2026 challenge
Winners:

ReflexTrack by: [Yuanjia Li, T. Xu, Z. Tang, H. Wang, X. Zhu, X. Wu, J. Kittler](#)

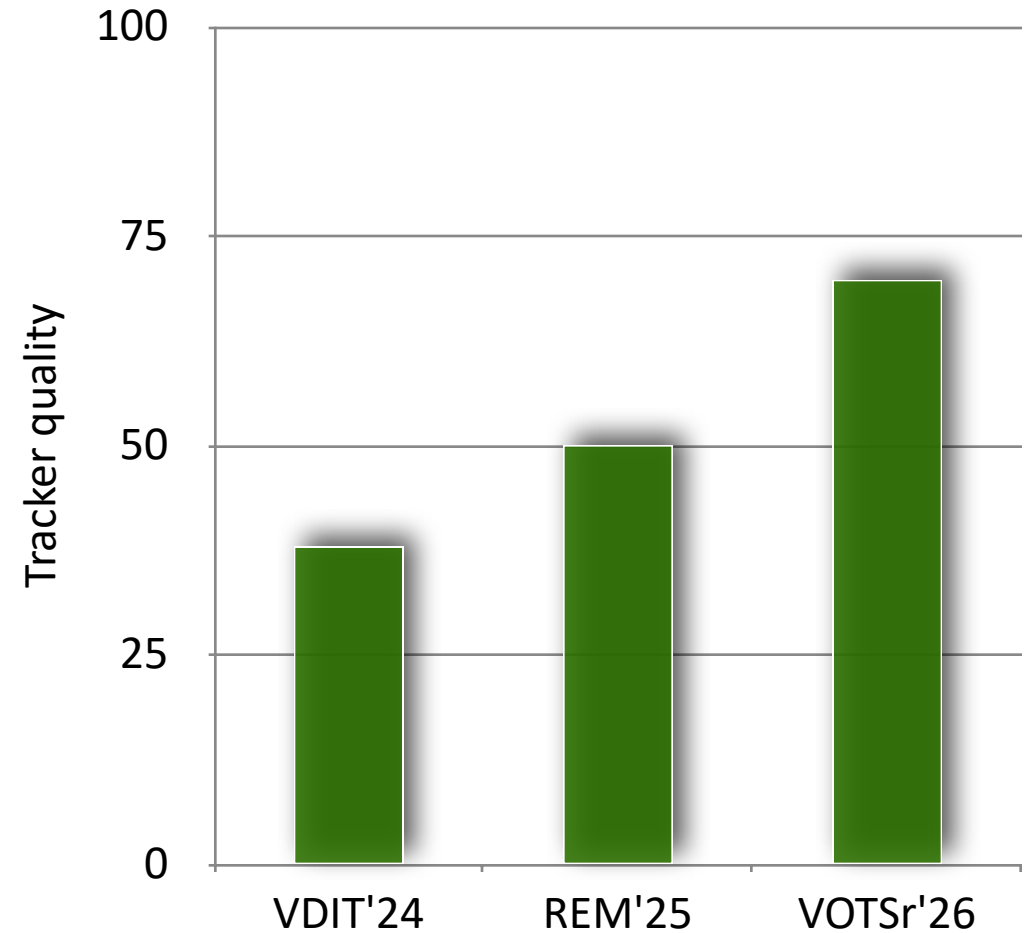
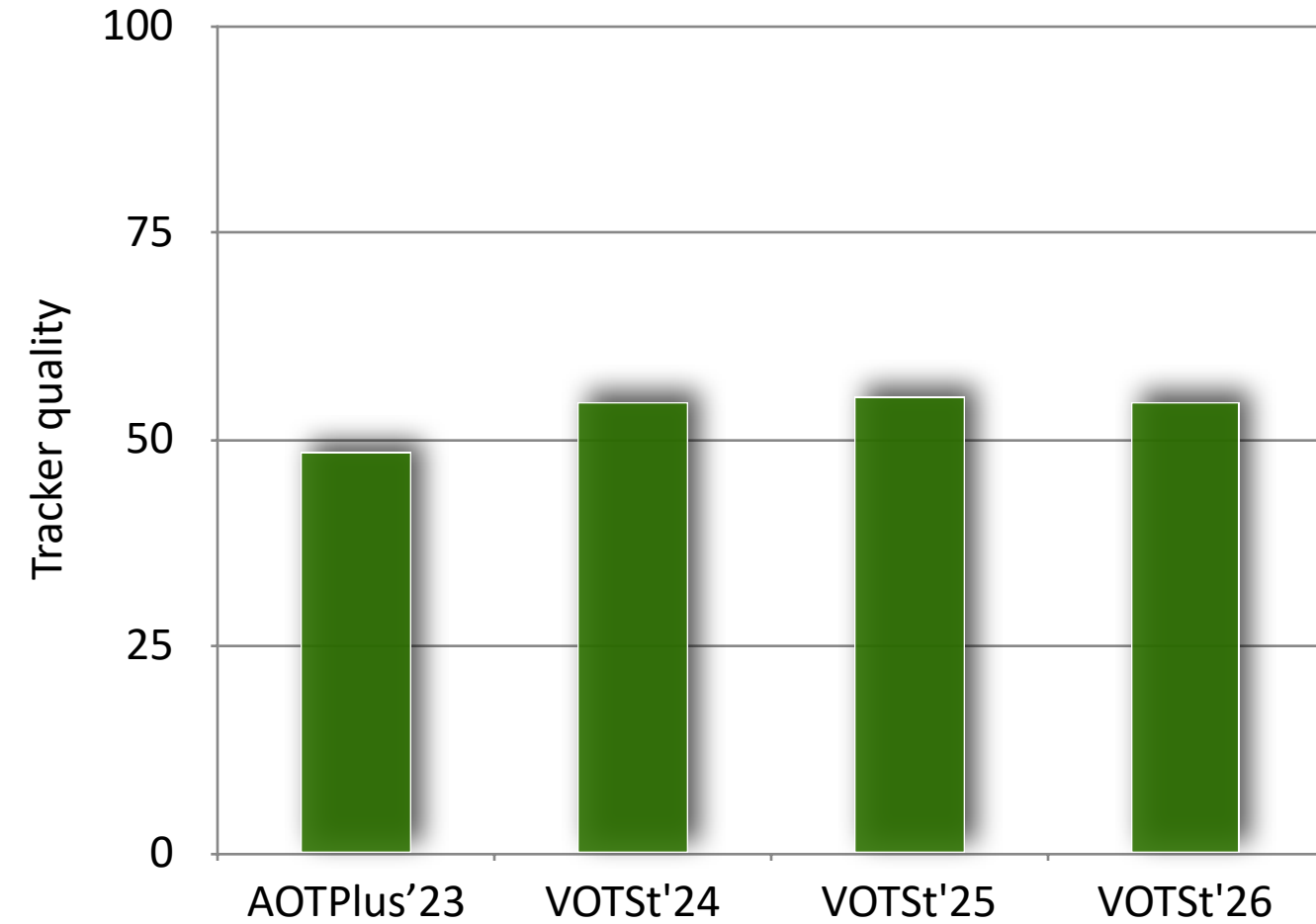
“A Feedback-Driven Agent for Training-Free Referring Video Object Segmentation”

Winners' talk later in Session II

Outline

- VOTSt dataset and challenge
- VOTSt 2026 Challenge results
- VOTSr dataset and challenge
- VOTSr 2026 Challenge results
- **Conclusions**

Contrasting the two challenges



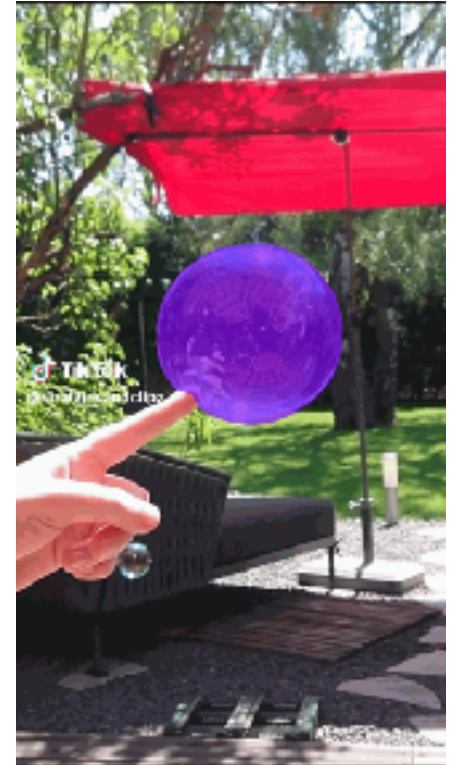
Sources of the gap

- Video length

VOTSt



VOTSr



The bubble being popped

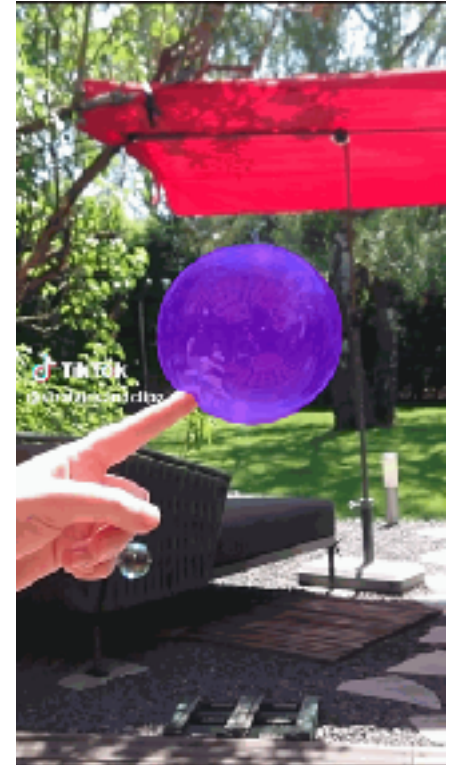
Sources of the gap

- Video length

VOTSt



VOTSr



The bubble being popped

Sources of the gap

- Video length
- Multi-instance labels

VOTSt



VOTSr



The bubble being popped

Sources of the gap

- Video length
- Multi-instance labels
- Strong ego-motion

VOTSt



VOTSr



The bubble being popped

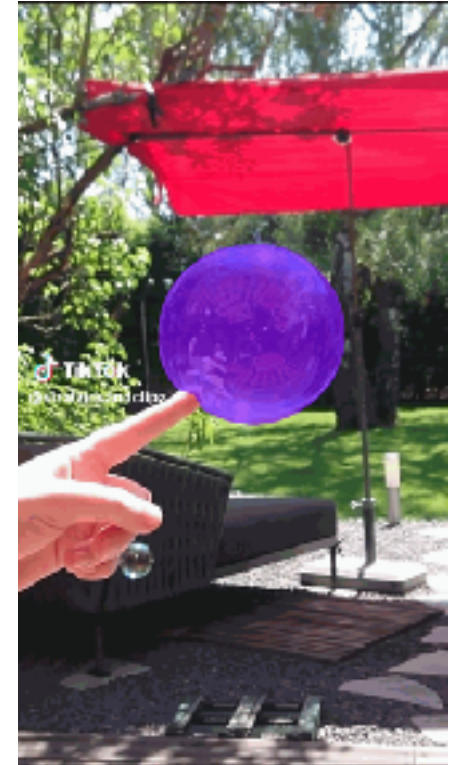
Sources of the gap

- Video length
- Multi-instance labels
- Strong ego-motion
- **Natural-language target descriptions**

VOTSt



VOTSr



The bubble being popped

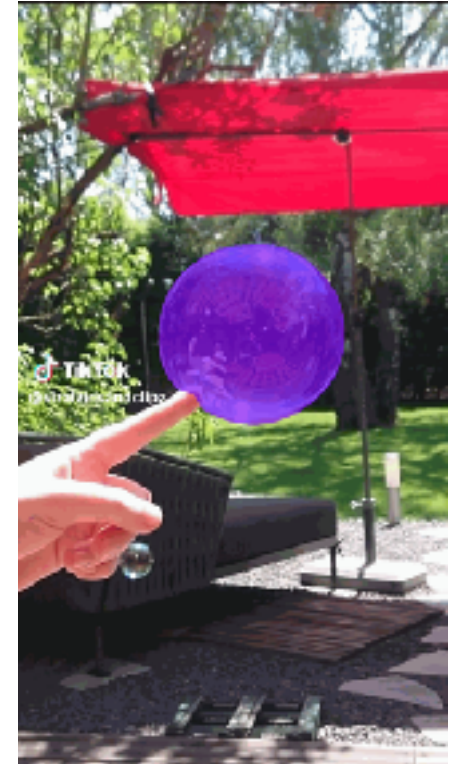
Sources of the gap

- Video length
- Multi-instance labels
- Strong ego-motion
- **Natural-language target descriptions**

VOTSt



VOTSr



The bubble being popped



The banana being peeled