Ego3DT: Tracking Every 3D Object in Ego-centric Videos

Paper   Presentation

Shengyu Hao[1]*, Wenhao Chai[2]*, Zhonghan Zhao[1]*, Meiqi Sun[1], Wendi Hu[1], Jieyang Zhou[1], Yixian Zhao[1], Qi Li[1], Yizhou Wang[2], Xi Li[1†], Gaoang Wang[1†]

1 Zhejiang University 2 University of Washington

ACM Multimedia 2024 Melbourne, Australia

Motivation & Contribution

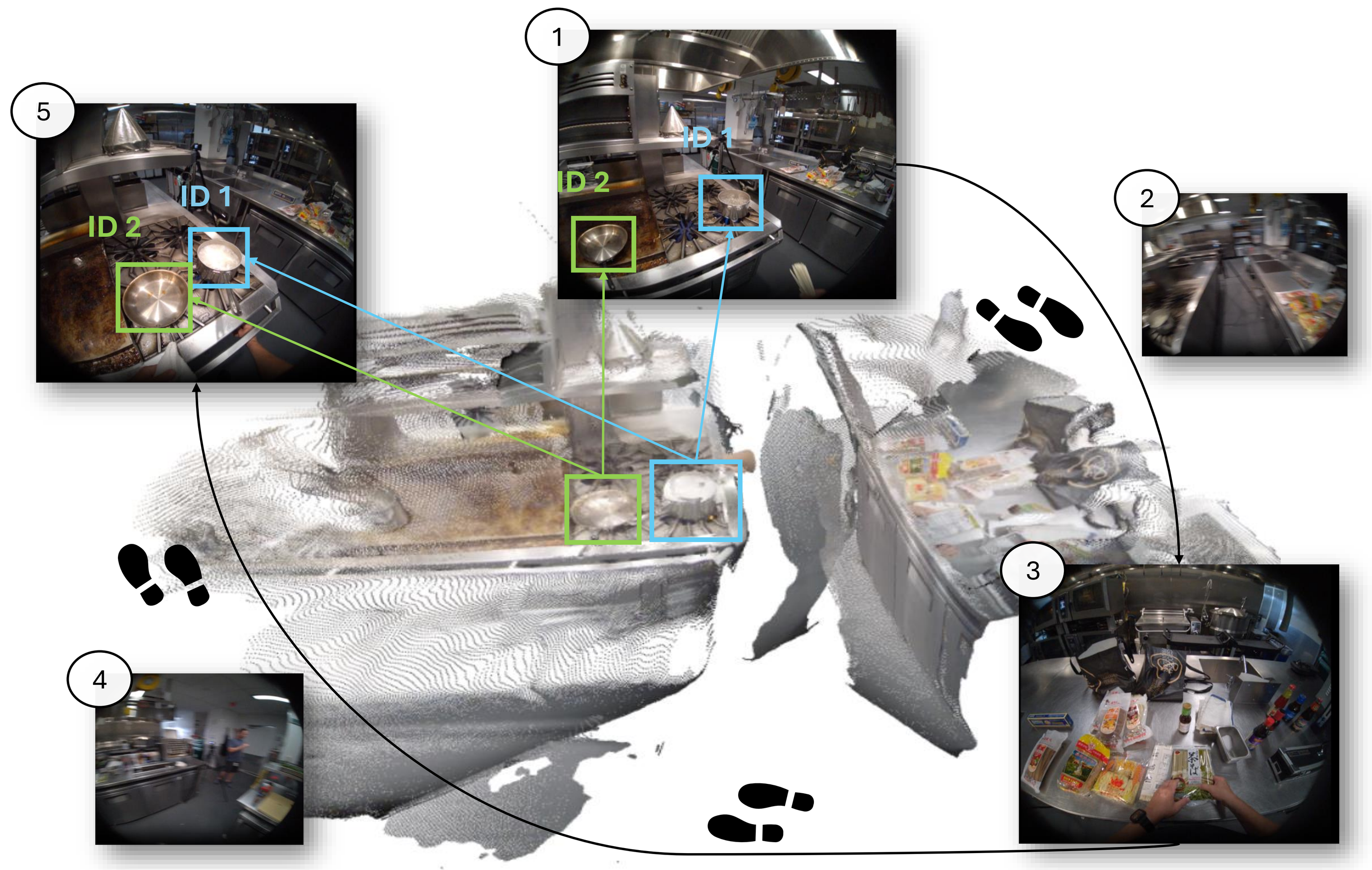
Motivation

Differing from traditional third-person videos, ego-centric videos often capture a wide range of activities, objects, and locations without a specific focus. Large head movements from the camera wearer frequently cause objects to exit and re-enter the field of view, and objects manipulated by hands may undergo frequent occlusions, along with rapid changes in scale, pose, and even state or appearance. These unique aspects make object tracking significantly more demanding than in scenarios typically presented in existing datasets, highlighting a critical gap in current evaluation methodologies. Traditional MOT tasks, when applied to ego-centric videos, often result in poor tracking accuracy.

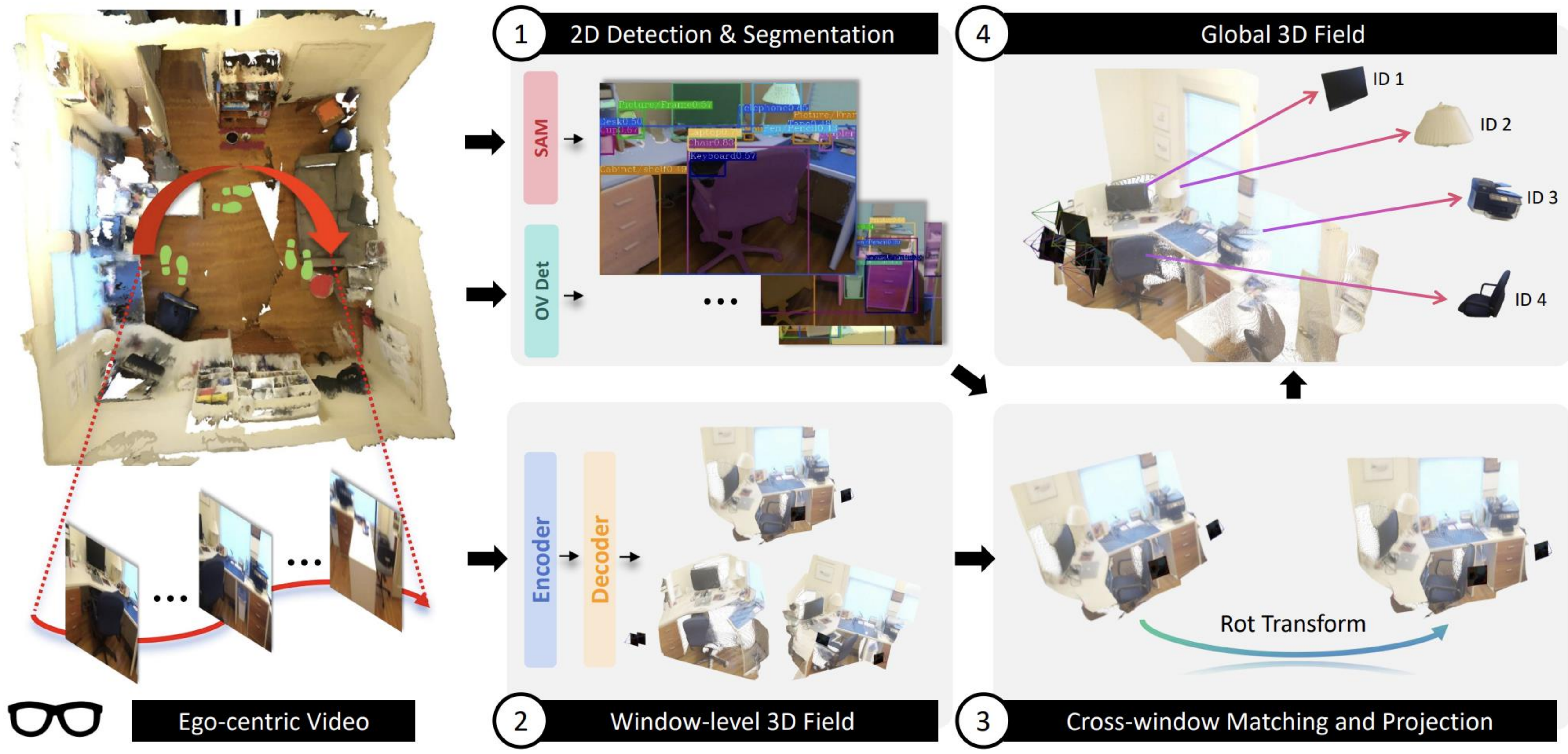
T=1   T=5   T=10

Contribution

- We propose a method for constructing a 3D scene from an ego-centric video and achieving open-vocabulary object tracking, which requires only RGB videos as input and is a zero-shot approach.
- We implement object 3D position matching through a dynamic cross-window matching method, thereby alleviating the instability caused by relying solely on 2D image tracking.
- Our method achieves state-of-the-art performance on the open-vocabulary multi-object tracking in ego-centric videos, with 1.04x-2.90x in HOTA.



Method



Algorithm 1 Cross-window Matching Process $\mathcal{M}$

Input: Video frames $X = \{I_i\}_{i=1}^{N}$, Initial 3D coordinates O_{3D}^{1}, Window size W, Overlap size T
Output: Tracked objects Y with IDs
Initialize: Buffer $\mathcal{B} \leftarrow \emptyset$, Detector **Det**, Segmenter **Seg**, 3D Estimator $\mathcal{G}$
$Y_0 \leftarrow Hungarian(\mathbf{PointMatch}(O_{3D}^{1}))$
Add Y_0 to $\mathcal{B}$ // Save to memory.
// Cross-window matching in the overlap
for $t = 1$ to T **do**
    $O_{3D}^{t} \leftarrow \mathcal{G}(X, \mathbf{Seg}(\mathbf{Det}(I_t)))$
    Align 3D scenes: $O_{3D}^{t} \leftarrow \mathcal{A}(O_{3D}^{t-1}, O_{3D}^{t})$
end for
for $t = 1$ to W **do**
    $Y_t \leftarrow \mathbf{PointMatch}(O_{3D}^{t-1}, O_{3D}^{t})$ // Matching 3D points
    Add Y_t with IDs to $\mathcal{B}$ // Save to memory.
end for
Convert buffer $\mathcal{B}$ to the output space Y
return Y

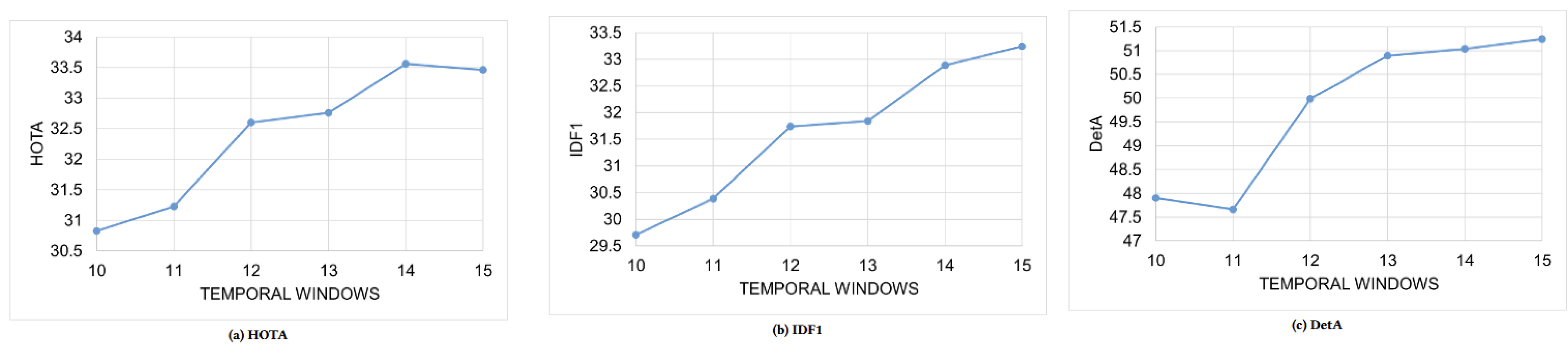
Experiment

MOT performance on different methods & detectors

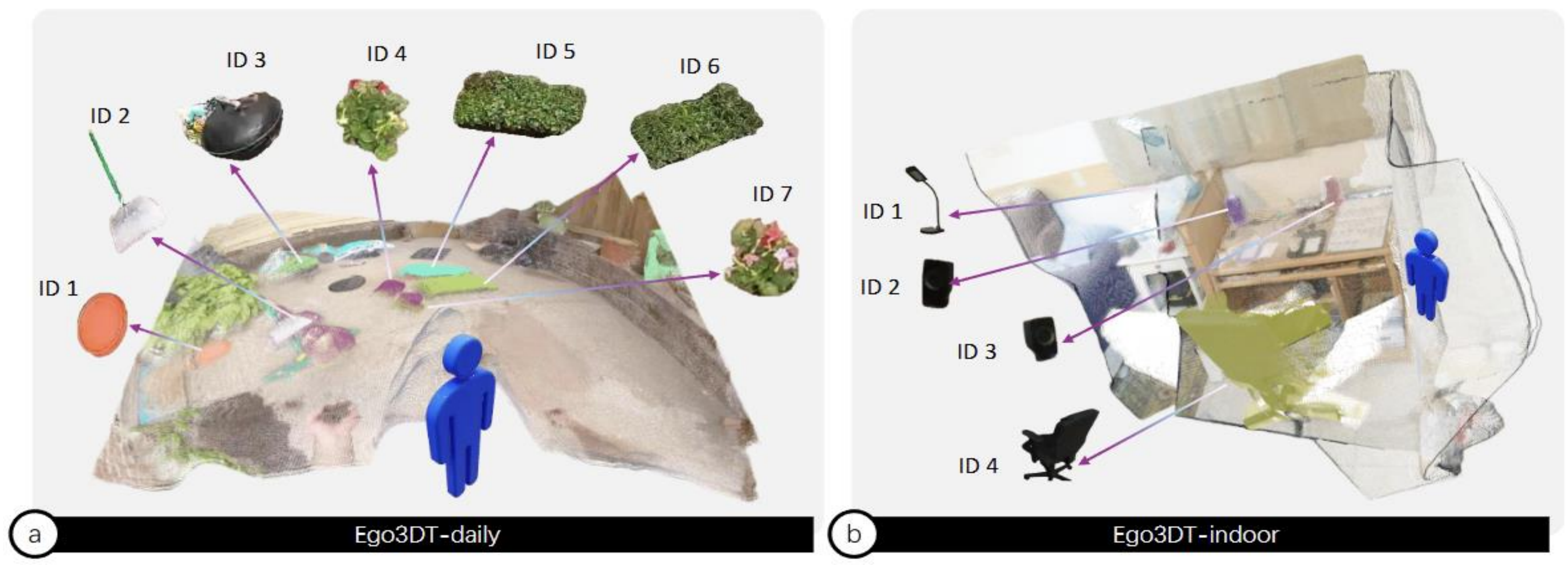
Tracker	Detector	Association	HOTA (↑)	IDF1 (↑)	DetA (↑)	MT (↑)	ML (↓)	Frag (↓)
ByteTrack [70]	YOLO-World [4]	2D box	19.14	18.77	17.11	23	78	775
	GLEE [61]	2D box	29.58	**31.28**	29.10	**30**	73	1217
DeepSort [60]	YOLO-World [4]	2D box + f	10.63	9.63	11.15	9	106	637
	GLEE [61]	2D box + f	15.91	15.79	18.00	9	90	710
OVTrack [32]	OVTrack [32]	2D box + f	15.40	15.15	12.90	6	123	816
TET [31]	TET [31]	2D box + f	13.94	13.34	11.41	5	134	583
Ego3DT (Ours)	OVTrack [32]	3D point	13.44	12.90	13.79	5	138	512
	TET [31]		12.40	11.62	13.24	5	134	**463**
	YOLO-World [4]		16.28	15.28	19.43	14	78	1196
	GLEE [61]		**30.83**	29.71	**47.91**	24	**49**	1217

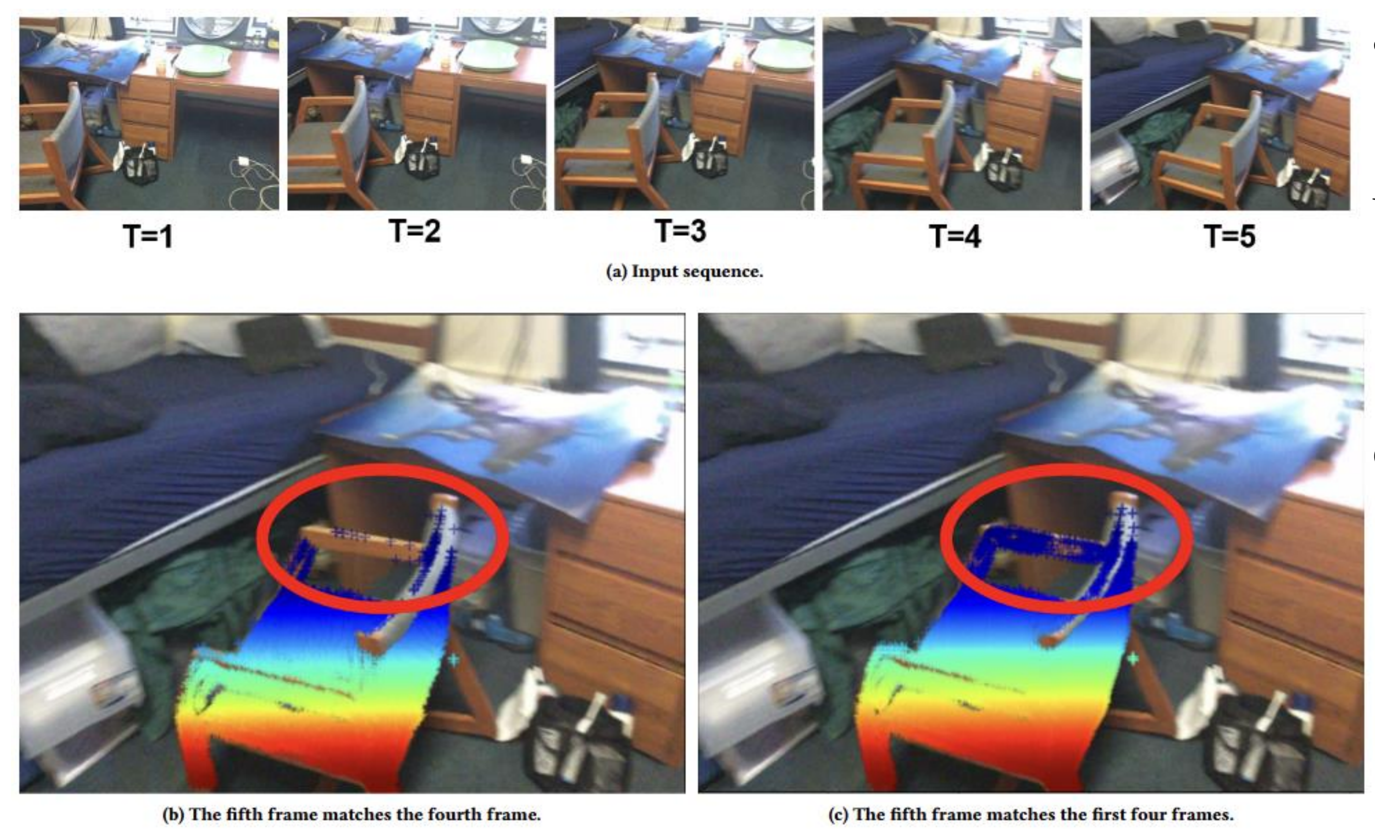
Setting		HOTA (↑)	IDF1 (↑)	DetA (↑)	MT (↑)	ML (↓)	Frag (↓)
Detector	YOLO-World [4]	16.28	15.28	19.43	14	78	**1196**
	GLEE [61]	**30.83**	**29.71**	**47.91**	**24**	**49**	1217
Memory	w/o Memory	29.13	28.68	44.56	21	**49**	**1216**
	30 Frames	**30.83**	**29.71**	**47.91**	**24**	**49**	1217
	Full Frames	27.60	28.54	38.60	18	109	1241

Impact of Temporal Windows



Visualization of 3D Fields & Memory Mechanism





a) The input sequence of frames T=1 to T=5 displays the tracking scenario.
b) The fifth frame's matching process with only the fourth frame, showing limited temporal context, is circled in red.
c) An enhanced matching approach in which the fifth frame matches the first four frames demonstrates the extended memory's role in capturing a broader temporal context for more accurate tracking.