中国科学院大学
University of Chinese Academy of Sciences
Institute of Automation
Chinese Academy of Sciences
智能系统与工程研究中心
Center for Research on Intelligent System and Engineering

Global Instance Tracking: Locating Target More Like Humans

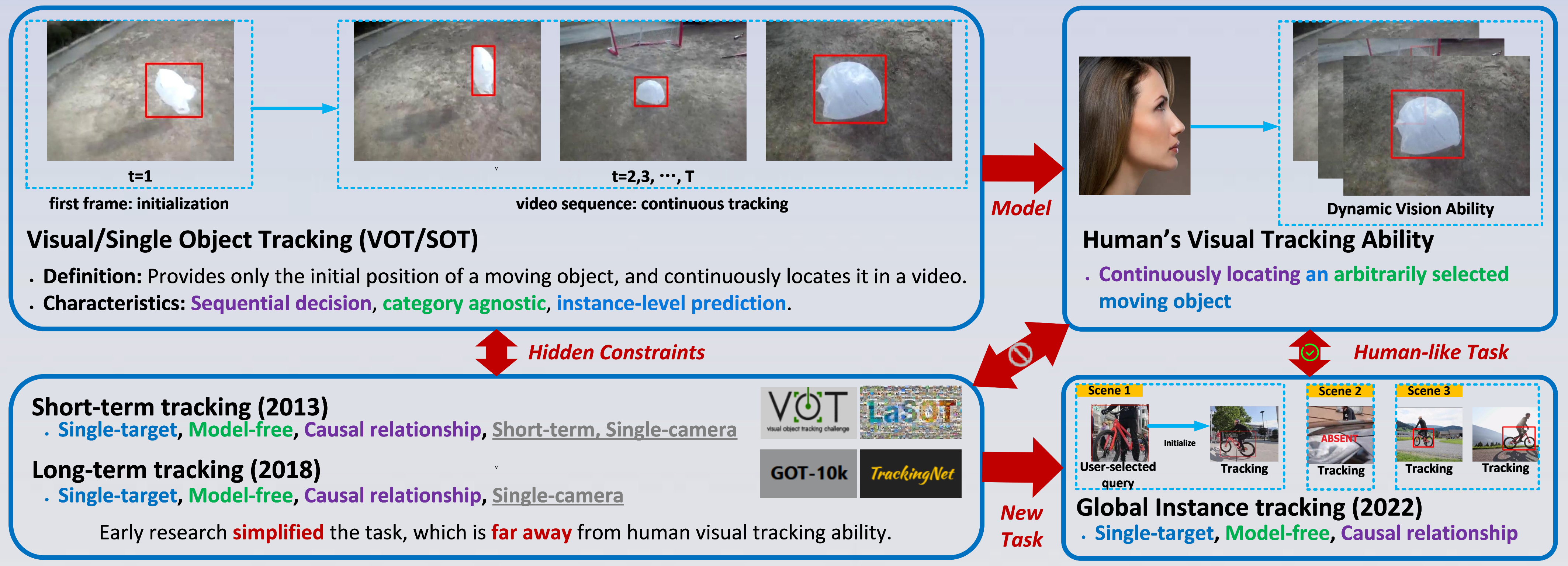
Shiyu Hu[1,2], Xin Zhao[1,2], Lianghua Huang[2], Kaiqi Huang[1,2,3]

1 School of Artificial Intelligence, University of Chinese Academy of Sciences; 2 Institute of Automation, Chinese Academy of Sciences; 3 Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences

VALSE 2024 重庆
视觉与学习青年学者研讨会
VISION AND LEARNING SEMINAR

Motivation: how to scientifically measure the tracking intelligence of an algorithm?

What are the abilities of humans? → Designing more human-like task to model the dynamic vision ability.



Visual/Single Object Tracking (VOT/SOT)

- **Definition:** Provides only the initial position of a moving object, and continuously locates it in a video.
- **Characteristics:** Sequential decision, category agnostic, instance-level prediction.

Human's Visual Tracking Ability

- Continuously locating an arbitrarily selected moving object

Short-term tracking (2013)

- Single-target, Model-free, Causal relationship, Short-term, Single-camera

Long-term tracking (2018)

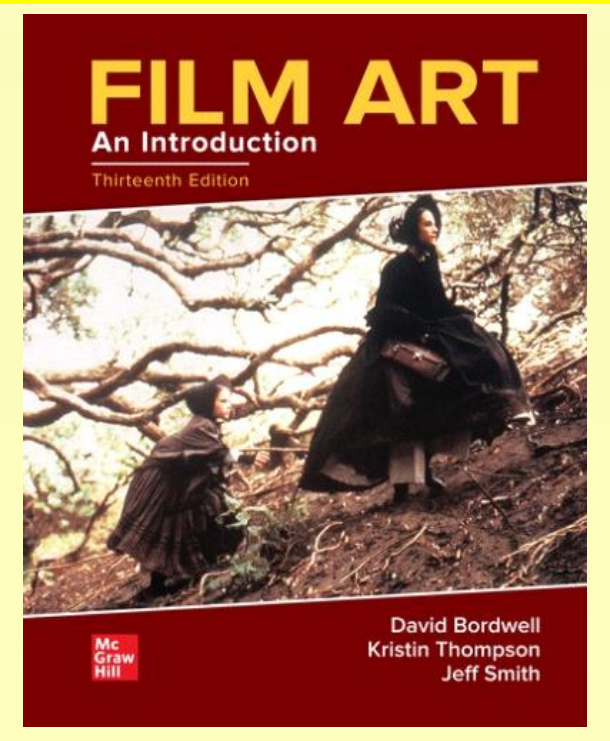
- Single-target, Model-free, Causal relationship, Single-camera

Early research simplified the task, which is far away from human visual tracking ability.

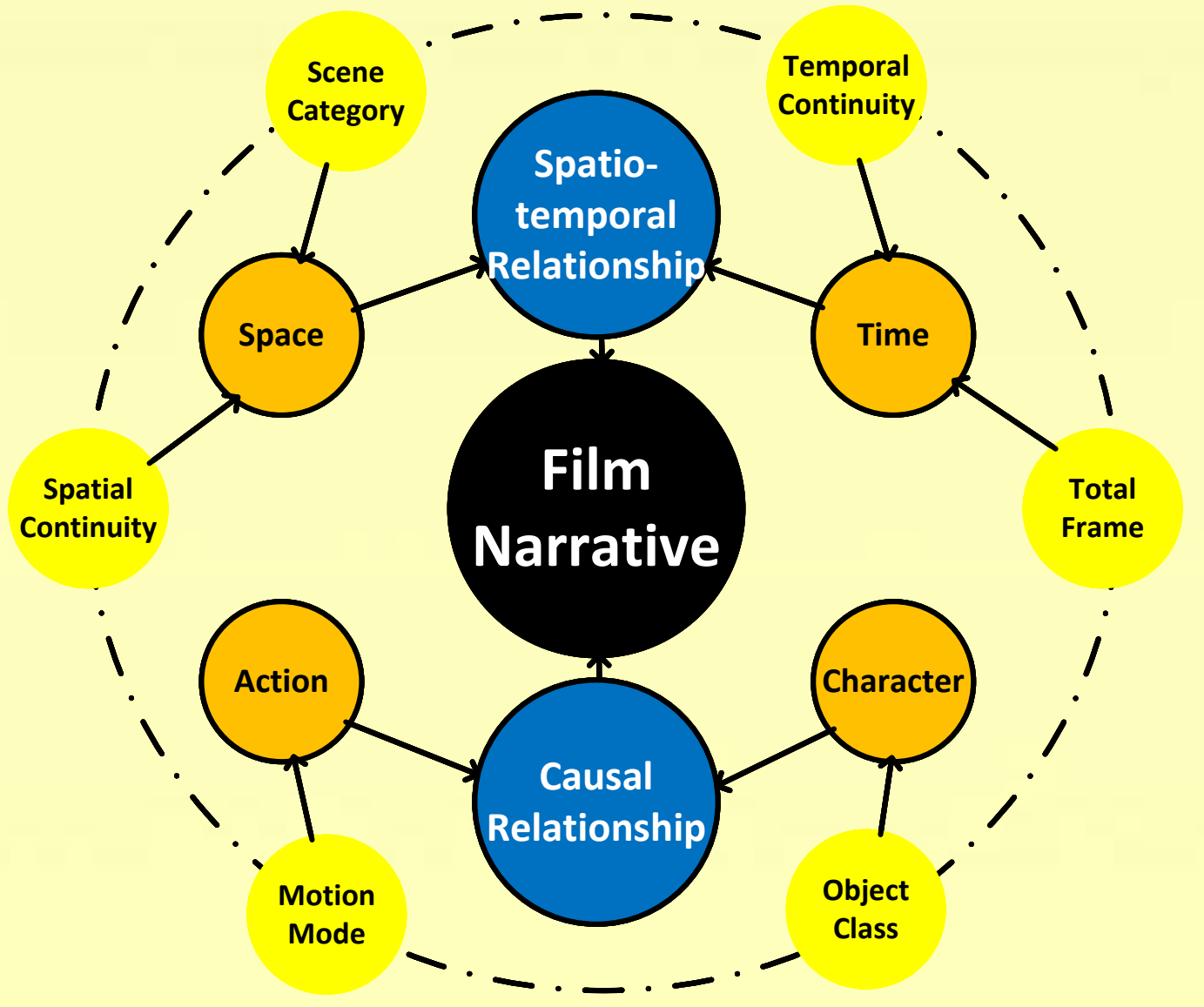
Global Instance tracking (2022)

- Single-target, Model-free, Causal relationship

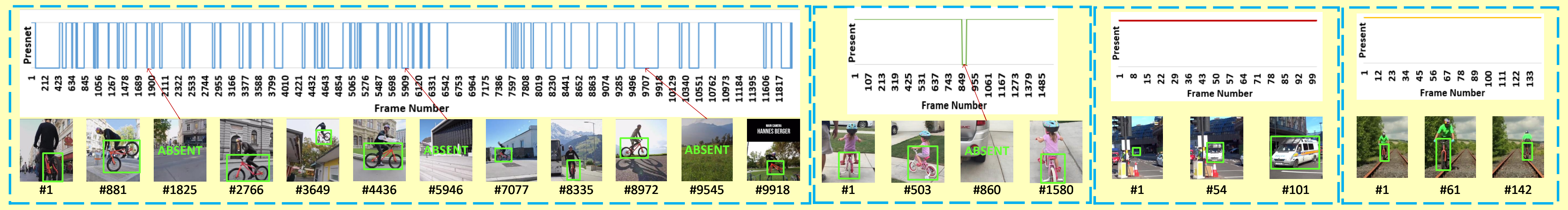
What are the living environments of humans? → Constructing more comprehensive and realistic datasets.



Film narrative is a chain of events in cause-effect relationship occurring in space and time.



The 6D principle of data collection



(a) VideoCube (b) LaSOT (c) GOT-10k (d) OTB100

Benchmark	Year	Videos	Min Frame	Mean Frame	Median Frame	Max Frame	Total Frame	Total Duration	Label Density	Attribute Classes (Absent)	Object Classes	Motion Modes	Scene Categories
OTB2013 [34]	2013	51	71	578	392	3872	29K	16.4m	30Hz	11(✗)	10	n/a	n/a
OTB2015 [1]	2015	100	71	590	393	3872	59K	32.8m	30Hz	11(✗)	16	n/a	n/a
TC-128 [41]	2015	129	71	429	365	3872	55K	30.7m	30Hz	11(✗)	27	n/a	n/a
NUS-PRO [42]	2015	365	146	371	300	5040	135K	75.2m	30Hz	n/a	8	n/a	n/a
UAV123 [43]	2016	123	109	915	882	3085	113K	75.2m	30Hz	12(✗)	9	n/a	n/a
VOT-2017 [4]	2017	60	41	356	293	1500	21K	11.9m	30Hz	n/a	24	n/a	n/a
Nfs [44]	2017	100	169	3830	2448	20665	383K	26.6m	240Hz	9(✗)	17	n/a	n/a
TrackingNet [2]	2018	30643	-	498	-	-	14M	141h	1Hz(30Hz)^a	15(✗)	27	n/a	n/a
GOT-10k [5]	2019	10000	29	149	101	1418	1.45M	40h	10Hz^b	6(✓)	563^c	87	n/a
UAV20L [43]	2016	20	1717	2934	2626	5527	59K	32.6m	30Hz	12(✗)	5	n/a	n/a
OxUvA [46]	2018	366	900	4320	2628	37740	1.55M	14.4h	1Hz^d	(✓)^e	22	n/a	n/a
LaSOT [3]	2020	1550	1000	2502	2145	11397	3.87M	35.8h	30Hz	14(✓)	85	n/a	n/a
VideoCube	2022	500	4008	14920	14162	29834	7.46M	69.1h	10Hz(30Hz)^f	12(✓)	9(89)^g	61	8(55)^h

ABILITY

TASK

=

ENVIRONMENT

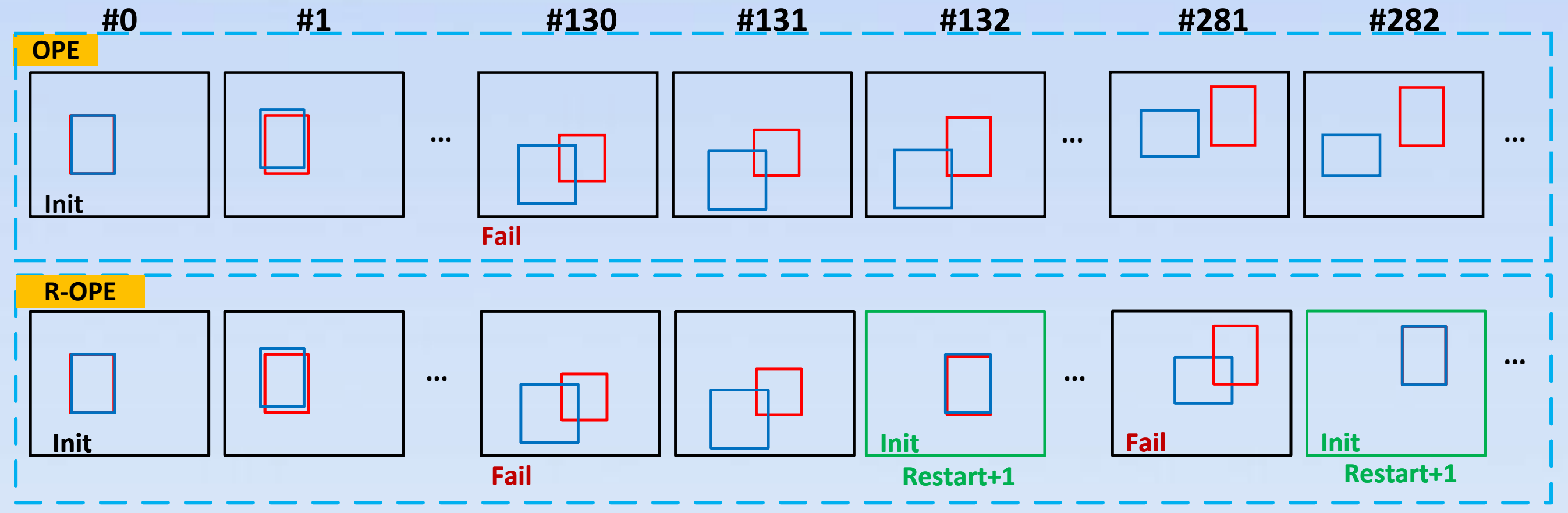
+

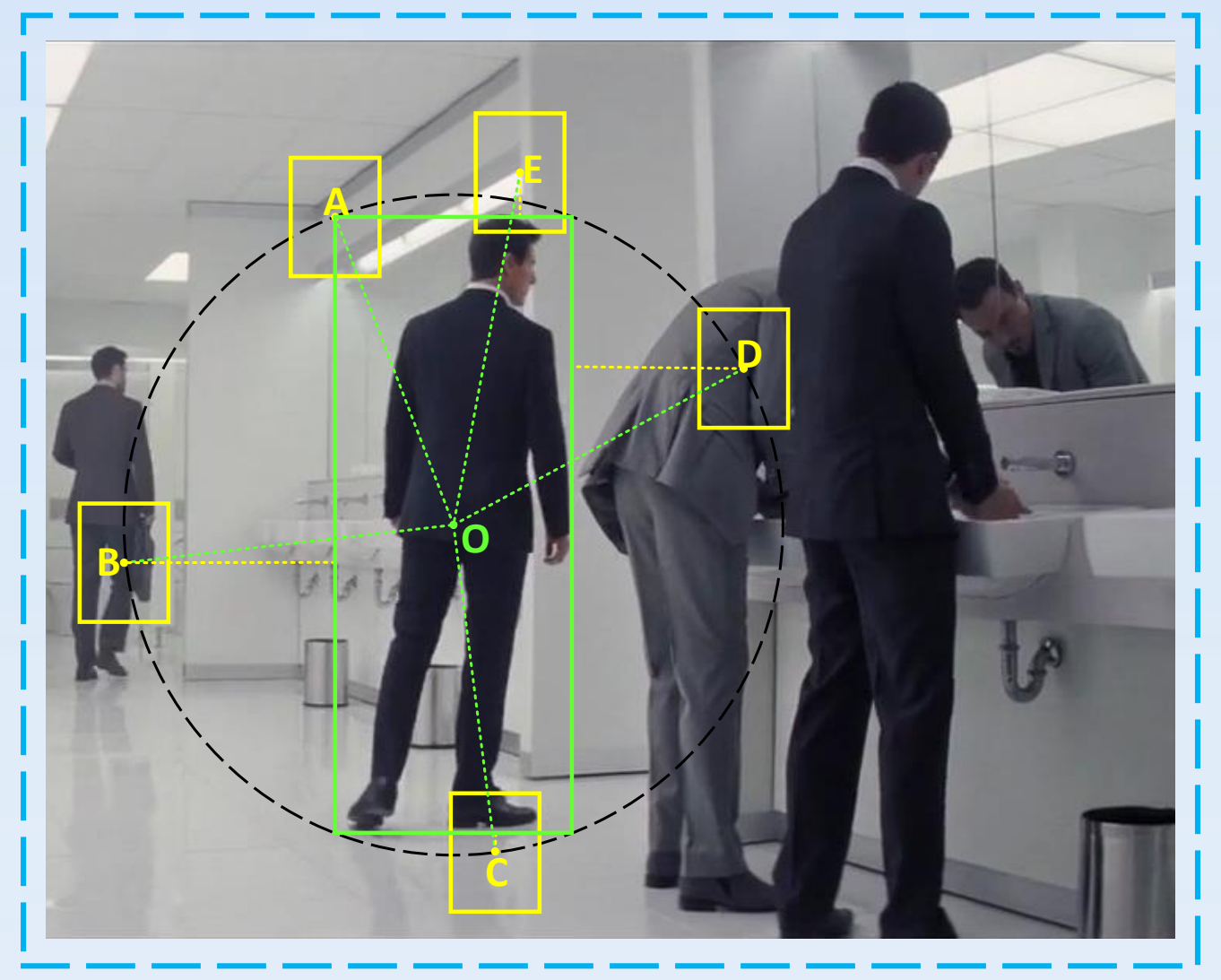
EVALUATION

+

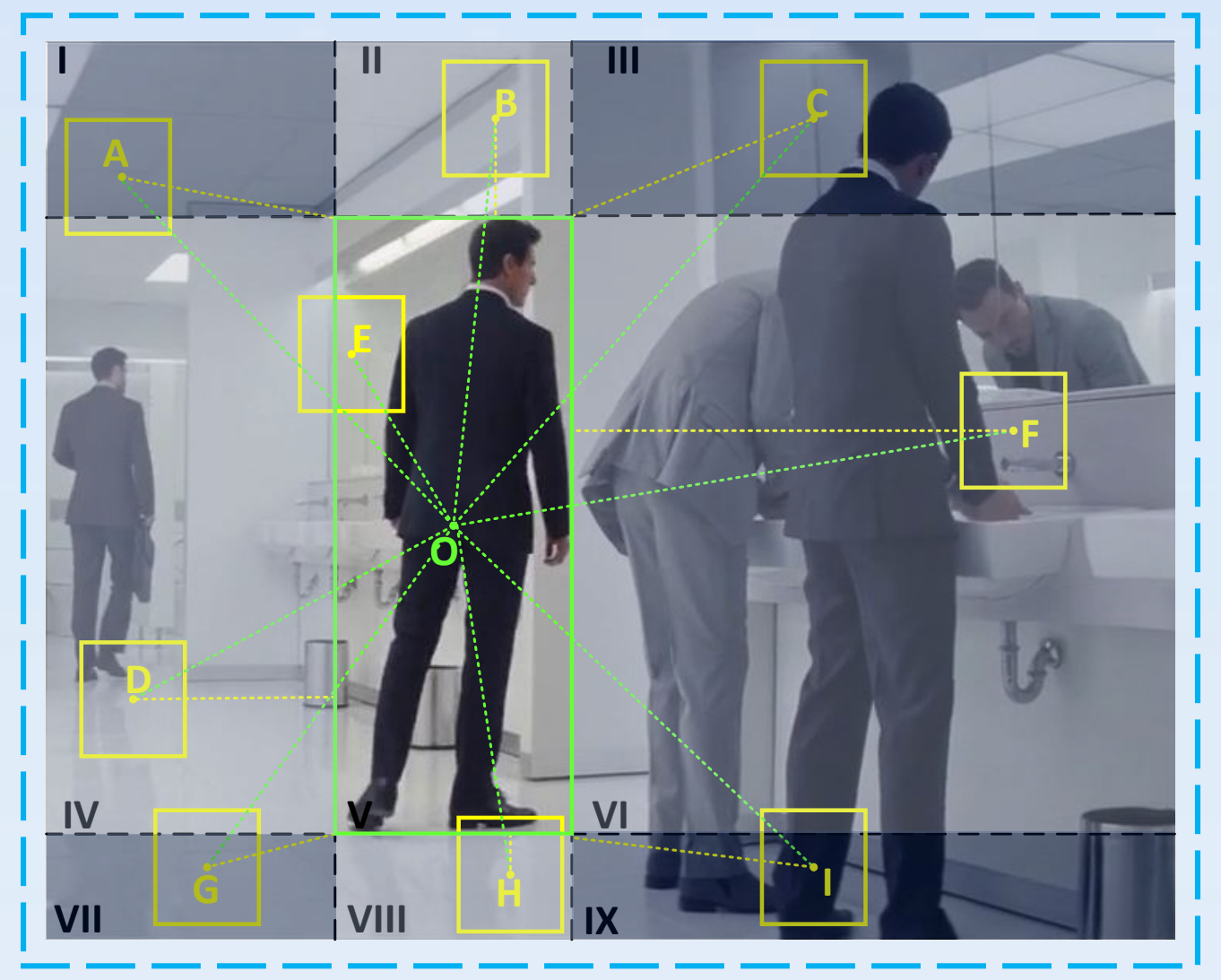
EXECUTOR

How do we evaluate human and machine visual tracking abilities? → Designing a universal framework for fair evaluation.





(a) Insufficiency of existing precision indicators



(b) The computation process of the normalized precision

How far the gap is between human and machine in dynamic vision ability? → Utilizing human as a baseline to evaluate machine intelligence.

Trackers	Venue	PRE (20 pixels)	N-PRE (percent in area V)	SR-IoU (mean)	SR-GIoU (mean)	SR-DIoU (mean)	FPS
Ocean	ECCV'20	0.179	0.523	0.328	0.316	0.312	11.665
SiamRCNN	CVPR'20	0.424	0.662	0.536	0.529	0.524	2.39
SuperDiMP	CVPR'20	0.28	0.602	0.444	0.436	0.433	6.423
PrDiMP	CVPR'20	0.26	0.617	0.421	0.412	0.408	5.791
LTMU	CVPR'20	0.276	0.641	0.446	0.435	0.433	2.512
SiamCAR	CVPR'20	0.095	0.321	0.151	0.144	0.142	11.063
SiamFCPP	AAAI'20	0.112	0.418	0.261	0.251	0.251	79.067
GlobalTrack	AAAI'20	0.262	0.688	0.434	0.424	0.42	3.712
DiMP	ICCV'19	0.176	0.52	0.356	0.346	0.342	6.252
SPLT	ICCV'19	0.135	0.532	0.325	0.311	0.309	11.436
SiamDW	CVPR'19	0.075	0.463	0.146	0.132	0.134	8.766
SiamRPNPP	CVPR'19	0.198	0.538	0.351	0.34	0.339	6.363
ATOM	CVPR'19	0.115	0.425	0.251	0.242	0.238	6.963
DaSiamRPN	CVPR'18	0.115	0.453	0.281	0.271	0.268	24.37
SiamRPN	CVPR'18	0.119	0.456	0.283	0.274	0.271	44.454
CSRT	IJCV'18	0	0.013	0.001	0.001	0.001	79.831
ECO	CVPR'17	0.024	0.255	0.116	0.102	0.101	7.176
SiamFC	ECCV'16	0.025	0.12	0.056	0.053	0.053	47.928
KCF	TPAMI'15	0.01	0.026	0.079	0.062	0.062	162.962
TLD	TPAMI'11	0.018	0.266	0.026	0.022	0.019	5.99

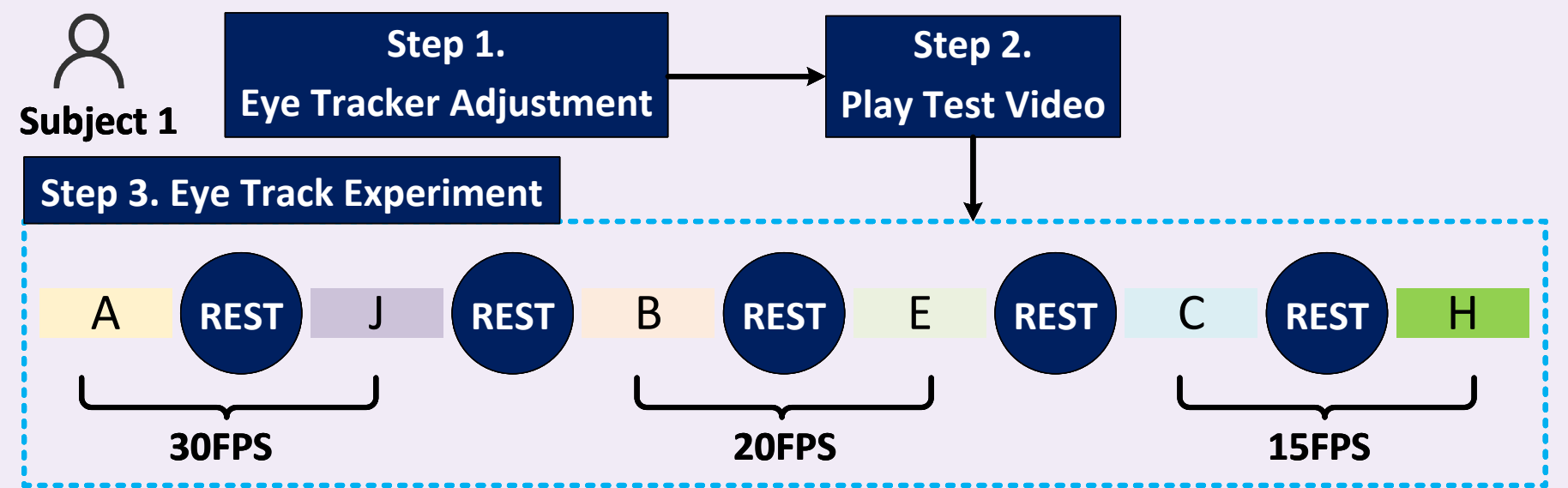
Results in OPE

Trackers	Venue	PRE (20 pixels)	N-PRE (percent in area V)	SR-IoU (mean)	SR-GIoU (mean)	SR-DIoU (mean)	Robust
Ocean	ECCV'20	0.379	0.73	0.505	0.493	0.495	0.7349
SiamRCNN	CVPR'20	0.548	0.785	0.643	0.637	0.635	0.7426
SuperDiMP	CVPR'20	0.432	0.781	0.597	0.59	0.588	0.7405
PrDiMP	CVPR'20	0.404	0.78	0.571	0.563	0.561	0.7397
LTMU	CVPR'20	0.398	0.778	0.562	0.553	0.553	0.7397
SiamCAR	CVPR'20	0.34	0.701	0.476	0.464	0.466	0.7332
SiamFCPP	AAAI'20	0.316	0.713	0.494	0.481	0.484	0.7338
GlobalTrack	AAAI'20	0.353	0.706	0.519	0.51	0.508	0.7404
DiMP	ICCV'19	0.364	0.753	0.55	0.54	0.54	0.7375
SPLT	ICCV'19	0.258	0.7	0.461	0.447	0.448	0.7321
SiamDW	CVPR'19	0.272	0.714	0.458	0.441	0.446	0.7313
SiamRPNPP	CVPR'19	0.375	0.734	0.525	0.516	0.515	0.7365
ATOM	CVPR'19	0.338	0.737	0.517	0.507	0.506	0.7355
DaSiamRPN	CVPR'18	0.317	0.71	0.495	0.484	0.484	0.7335
SiamRPN	CVPR'18	0.316	0.712	0.496	0.485	0.486	0.7336
CSRT	IJCV'18	0	0.019	0	0	0	0.6848
ECO	CVPR'17	0.294	0.725	0.469	0.453	0.457	0.7318
SiamFC	ECCV'16	0.25	0.514	0.345	0.334	0.338	0.7254
KCF	TPAMI'15	0.223	0.621	0.391	0.373	0.376	0.7224
TLD	TPAMI'11	0.017	0.261	0.026	0.022	0.019	0.6868

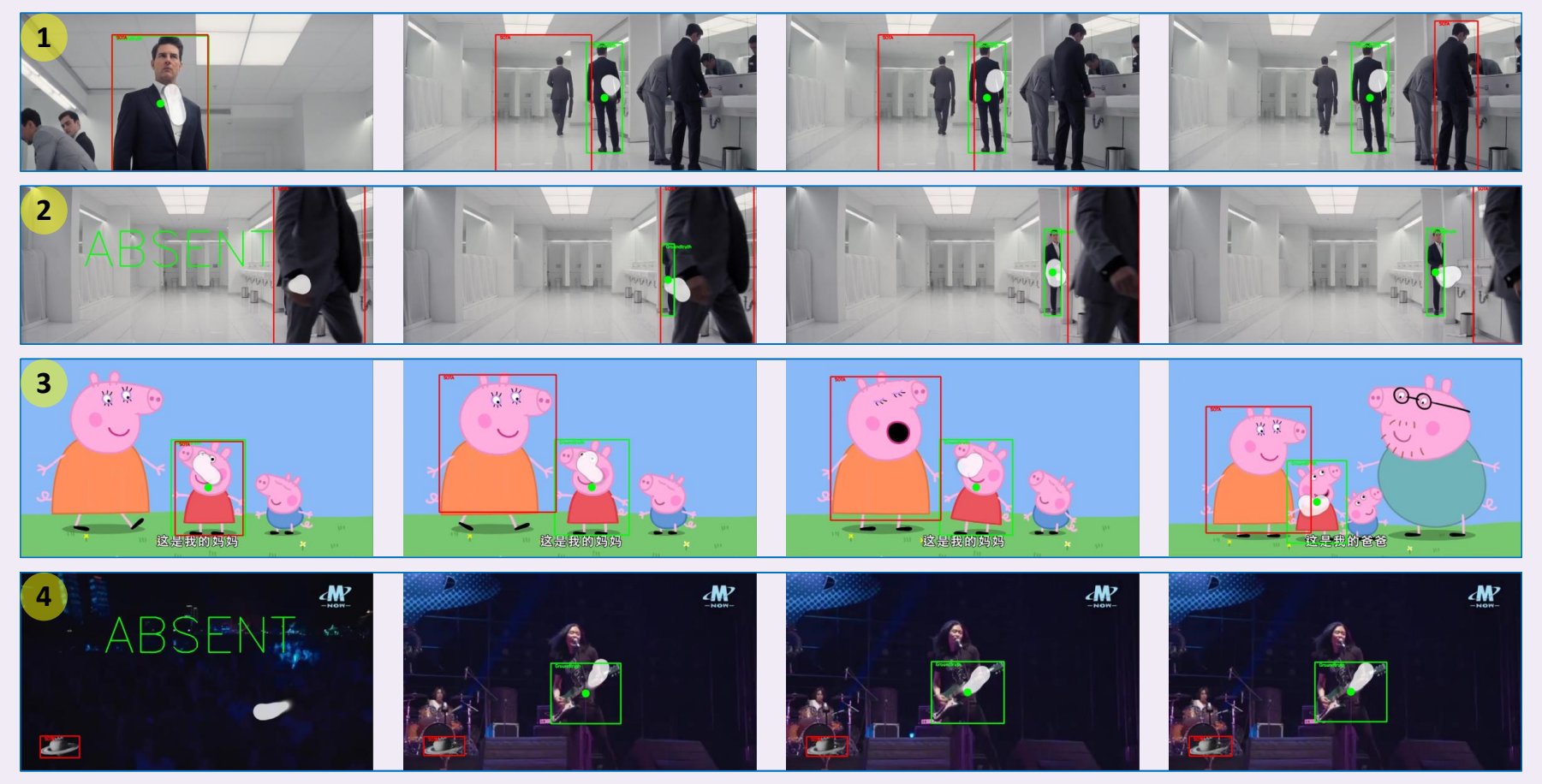
Results in R-OPE

Trackers	Venue	PRE (20 pixels)	N-PRE (percent in area V)
Ocean	ECCV'20	0.256	0.476
SiamRCNN	CVPR'20	0.551	0.71
SuperDiMP	CVPR'20	0.398	0.617
PrDiMP	CVPR'20	0.354	0.59
LTMU	CVPR'20	0.421	0.662
SiamCAR	CVPR'20	0.142	0.4
SiamFCPP	AAAI'20	0.153	0.412
GlobalTrack	AAAI'20	0.405	0.687
DiMP	ICCV'19	0.26	0.487
SPLT	ICCV'19	0.158	0.501
SiamDW	CVPR'19	0.106	0.431
SiamRPNPP	CVPR'19	0.262	0.521
ATOM	CVPR'19	0.151	0.408
DaSiamRPN	CVPR'18	0.136	0.39
SiamRPN	CVPR'18	0.132	0.371
CSRT	IJCV'18	0.001	0.116
ECO	CVPR'17	0.028	0.208
SiamFC	ECCV'16	0.044	0.143
KCF	TPAMI'15	0.005	0.141
TLD	TPAMI'11	0.019	0.293
Turing_15	Human	0.377	0.85
Turing_20	Human	0.243	0.805
Turing_30	Human	0.203	0.778

Results in VTT


Subject 1

Step 1. Eye Tracker Adjustment → Step 2. Play Test Video → Step 3. Eye Track Experiment



Some examples (human > machine)

Our series of work

Under Review
Remembering Target More Like Humans: A Robust Visual-Language Tracker with Adaptive Prompts
📌 Visual Language Tracking 📌 Human-like Memory Modeling 📌 Adaptive Prompts

Under Review
Target or Distractor? Rethinking Similar Object Interference in Single Object Tracking
📌 Visual Object Tracking 📌 Similar Object Interference 📌 Data Mining

CVPRW 2024
Diverse Text Generation for Visual Language Tracking Based on LLM
📌 Visual Language Tracking 📌 Large Language Model

TCSVT 2024
Finger in Camera Speaks Everything: Unconstrained Air-Writing for Real-World
📌 Air-writing Technique 📌 Large-scale Benchmark Construction 📌 Human-machine Interaction

JIG 2023
Visual Intelligence Evaluation Techniques for Single Object Tracking: A Survey (单目标跟踪中的视觉智能评估技术综述)
📌 Visual Object Tracking 📌 Intelligent Evaluation Technique 📌 AI4Science

IJCV 2023
BioDrone: A Bionic Drone-based Single Object Tracking Benchmark for Robust Vision
📌 Visual Object Tracking 📌 Drone-based Tracking 📌 Visual Robustness
🌐 http://biodrone.aitestunion.com/

IJCV 2023
SOTVerse: A User-defined Task Space of Single Object Tracking
📌 Visual Object Tracking 📌 Dynamic Open Environment Construction 📌 3E Paradigm
🌐 http://metaverse.aitestunion.com/

NIPS 2023
A Multi-modal Global Instance Tracking Benchmark (MGIT): Better Locating Target in Complex Spatio-temporal and causal Relationship
📌 Visual Language Tracking 📌 Long Video Understanding and Reasoning
🌐 http://videocube.aitestunion.com/

TPAMI 2023
Global Instance Tracking: Locating Target More Like Humans
📌 Visual Object Tracking 📌 Large-scale Benchmark Construction 📌 Intelligent Evaluation Technology
🌐 http://videocube.aitestunion.com/

TPAMI 2021
GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild
📌 Short-term Tracking 📌 Large-scale Benchmark Construction 📌 Visual Generalization
🌐 http://got-10k.aitestunion.com/

VideoCube Platform for More Information about GIT and MGIT

Visual Intelligence Interest Group for Communication and Collaboration