NEURAL INFORMATION PROCESSING SYSTEMS

A Multi-modal Global Instance Tracking Benchmark (MGIT): Better Locating Target in Complex Spatio-temporal and Causal Relationship

Shiyu Hu[1,2], Dailing Zhang[1,2], Meiqi Wu[3], Xiaokun Feng[1,2], Xuchen Li[4], Xin Zhao[1,2], Kaiqi Huang[1,2,5]

1 School of Artificial Intelligence, University of Chinese Academy of Sciences; 2 Institute of Automation, Chinese Academy of Sciences; 3 School of Computer Science and Technology, University of Chinese Academy of Sciences; 4 School of Computer Science, Beijing University of Posts and Telecommunications; 5 Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences

中国科学院大学 University of Chinese Academy of Sciences

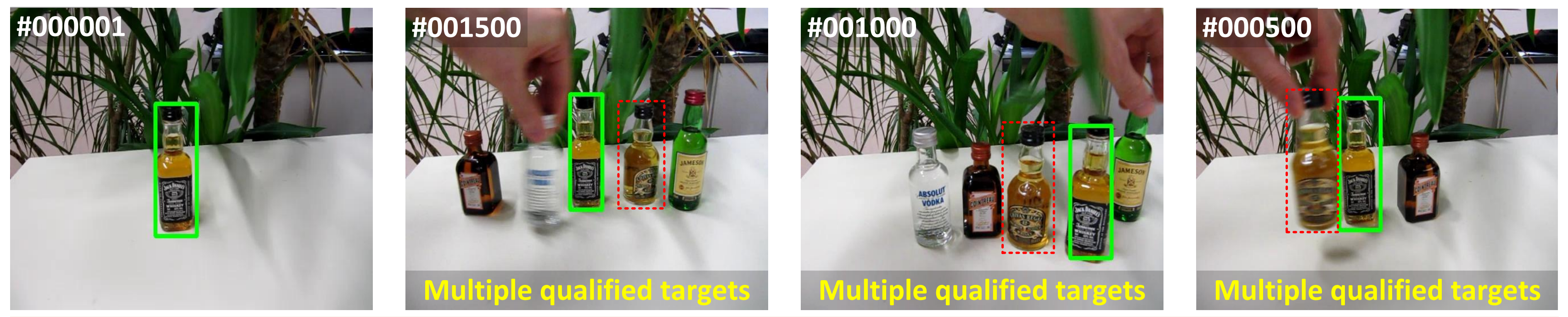
Institute of Automation Chinese Academy of Sciences

智能系统与工程研究中心 Center for Research on Intelligent System and Engineering

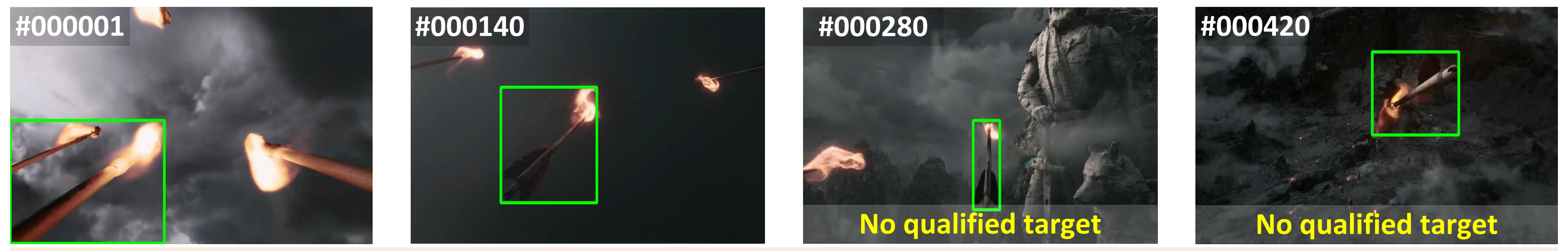
Platform for more information

Motivation

Limitations in existing visual-language tracking benchmarks


#000001
#001500
#001000
#000500
Multiple qualified targets
Multiple qualified targets
Multiple qualified targets

Language description: brown liquor bottle



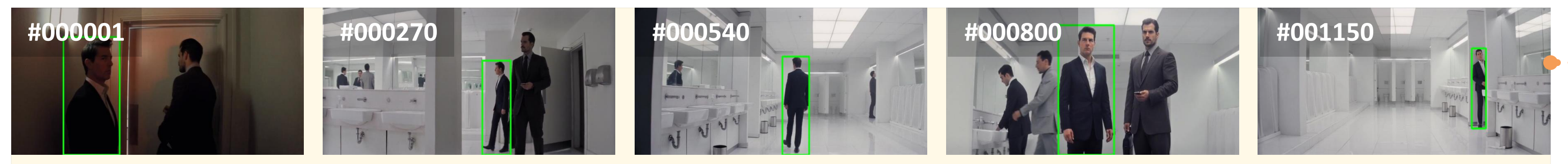
Language description: the second arrow from left to right

Video Short sequences with uncomplicated spatio-temporal and causal relationships

Language Simple semantic descriptions (multiple qualified targets, no qualified target)

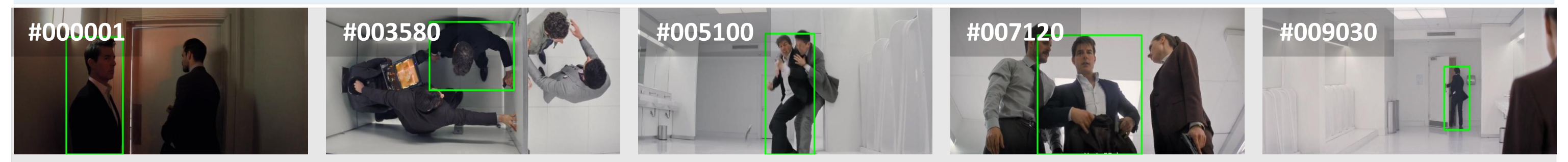
Existing trackers always perform poorly in complex environments (e.g., longer videos with more complicated narrative content)

Method

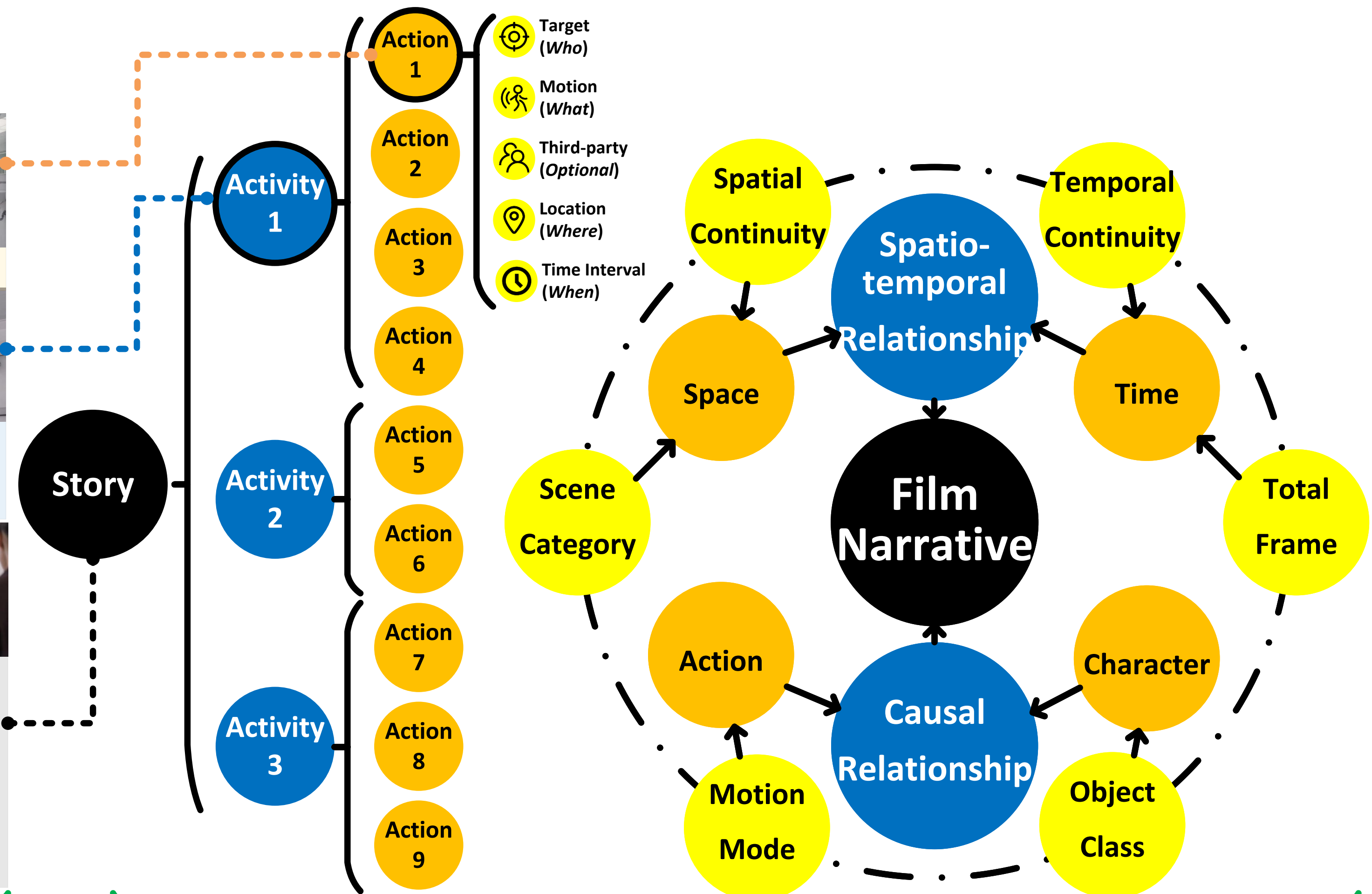


Action 1: A male secret agent wearing a black suit walks in the washroom.

Activity 1: A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, then the male secret agent wins, and lifts the insensible grey-suit man to the washroom cubicle.



Story: A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, and the male secret agent wins. He then lifts the insensible grey-suit man to the washroom cubicle. The male secret agent crouches in the washroom cubicle and checks the insensible grey-suit man. Suddenly, the grey-suit man wakes up, and they fight together again in the washroom. Eventually, the male secret agent wins the fight. After the male secret agent talks with a woman wearing a brown suit, he again lifts the insensible grey-suit man to the washroom cubicle. Finally, the male secret agent lefts the washroom after talking with the brown-suit woman.



Video Long sequences with complex spatial-temporal variation and casual relationship

Language Multi-granular annotation strategy based on the hierarchical structure of human cognitive

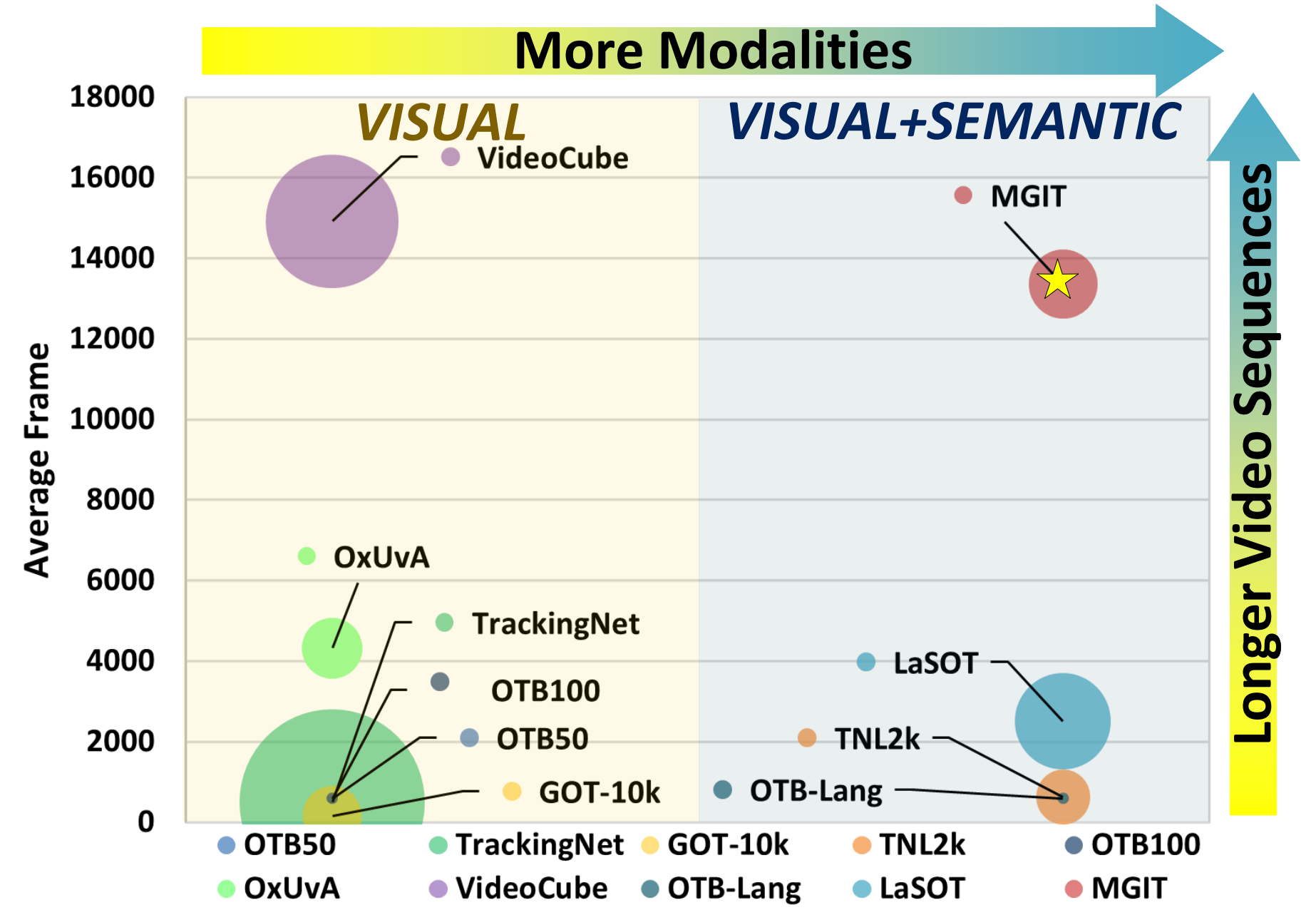
Coupling human causal reasoning ability into multi-modal information → Helping algorithms understand video content from a multi-modal perspective

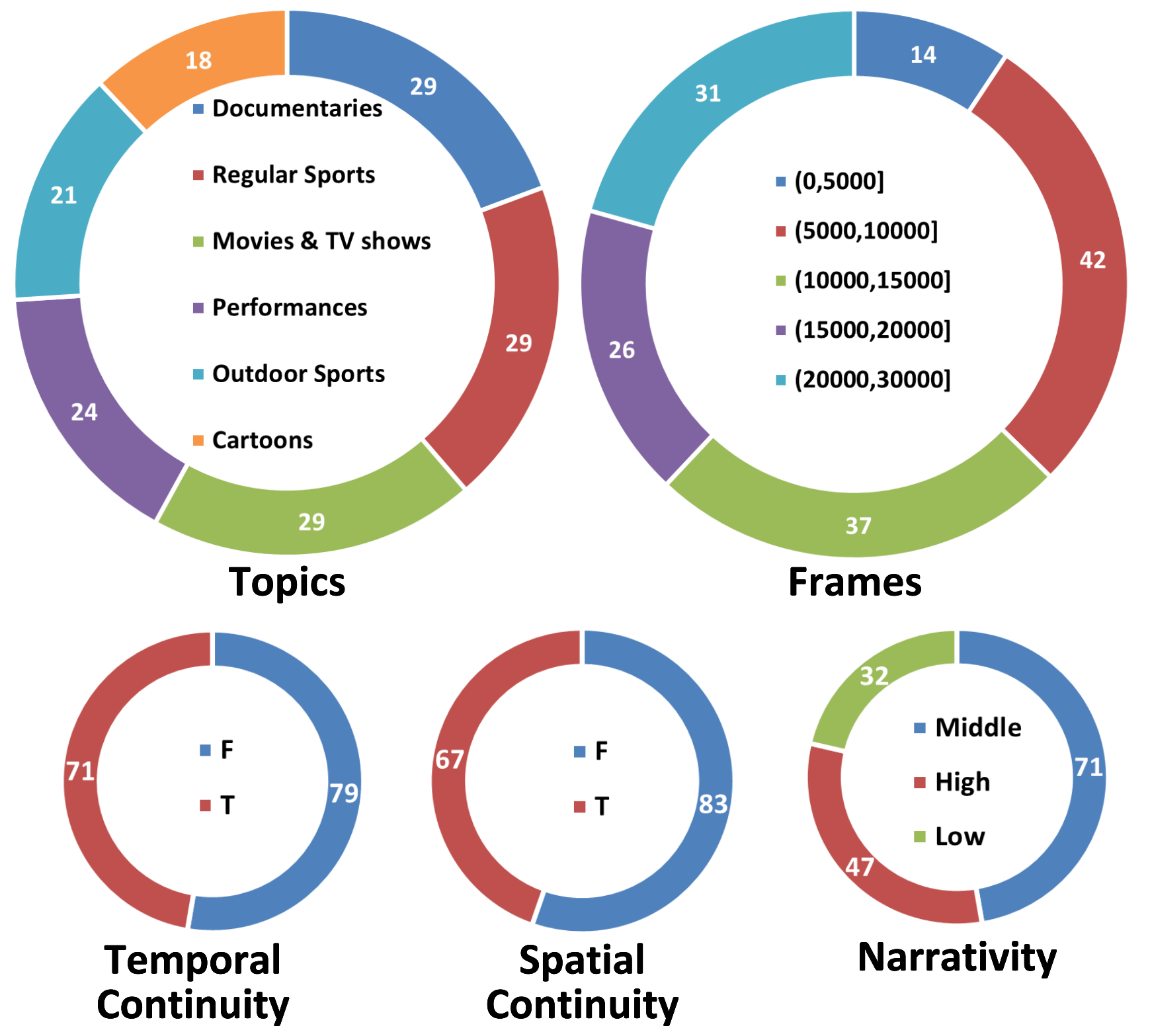
Contributions

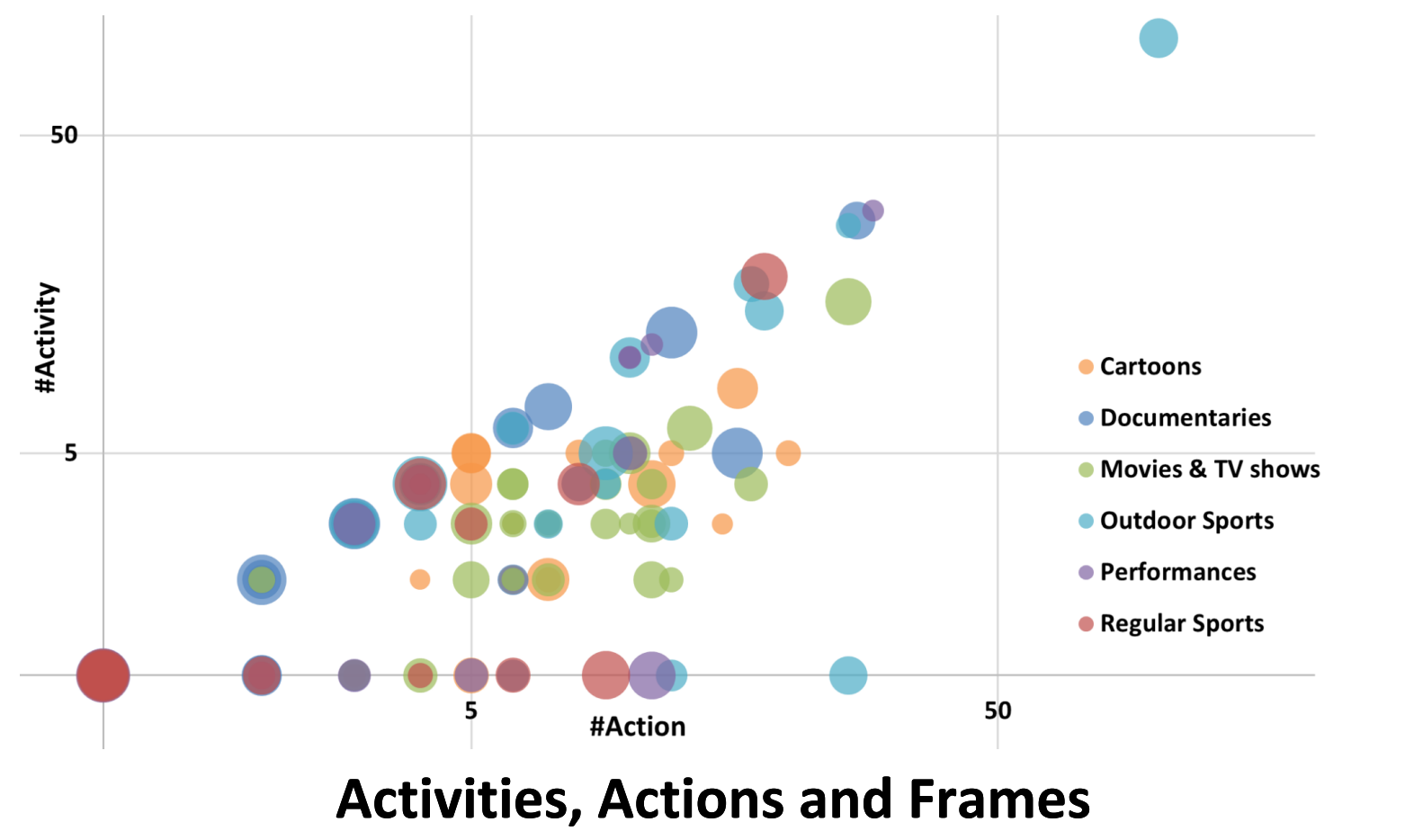
- We propose a new multi-modal benchmark named MGIT. It consists of 150 long videos with a total of 2.03 million frames, and the average length of a single sequence is 5-22 times longer than existing multi-modal benchmarks.
- We design a multi-granular annotation strategy for providing scientific semantic information.
- We execute comparative experiments on other benchmarks, and conduct detailed experimental analyses on MGIT.

MGIT Benchmark

Dataset Distribution

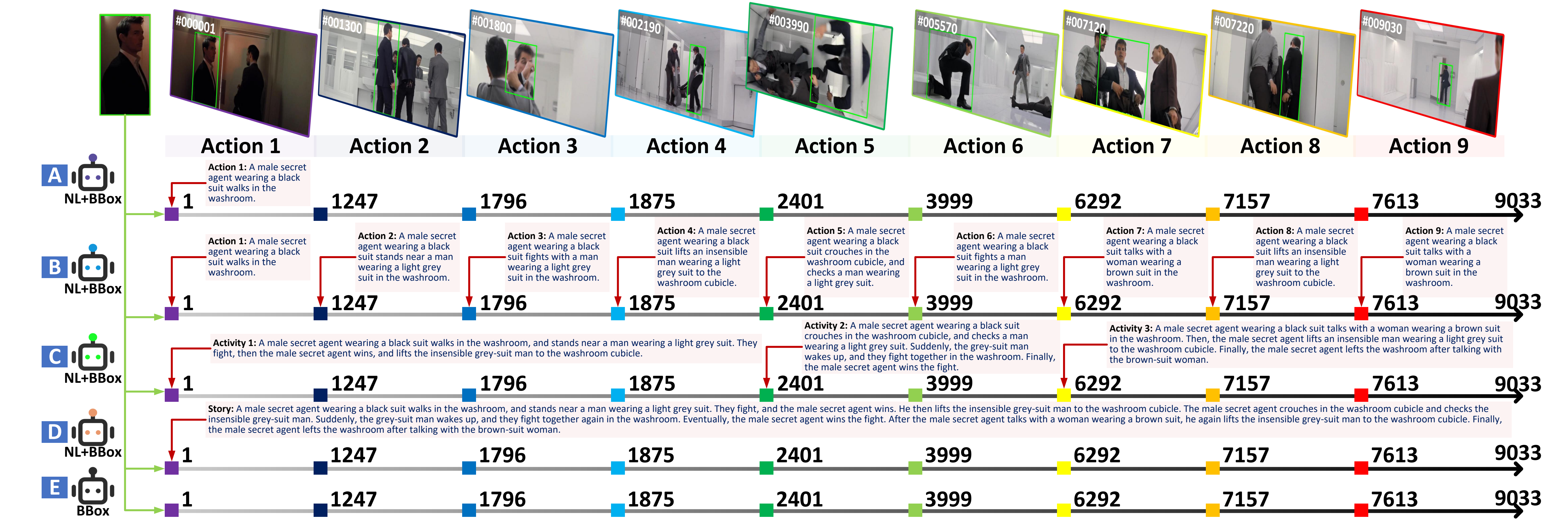






Word Cloud

Evaluation System



A: **Action 1:** A male secret agent wearing a black suit walks in the washroom.

B: **Action 1:** A male secret agent wearing a black suit walks in the washroom. **Action 2:** A male secret agent wearing a black suit stands near a man wearing a light grey suit in the washroom. **Action 3:** A male secret agent wearing a black suit fights with a man wearing a light grey suit in the washroom. **Action 4:** A male secret agent wearing a black suit lifts an insensible man wearing a light grey suit to the washroom cubicle. **Action 5:** A male secret agent wearing a black suit crouches in the washroom cubicle, and checks a man wearing a light grey suit. **Action 6:** A male secret agent wearing a black suit fights a man wearing a light grey suit in the washroom. **Action 7:** A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom. **Action 8:** A male secret agent wearing a black suit lifts an insensible man wearing a light grey suit to the washroom cubicle. **Action 9:** A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom.

C: **Activity 1:** A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, then the male secret agent wins, and lifts the insensible grey-suit man to the washroom cubicle. **Activity 2:** A male secret agent wearing a black suit crouches in the washroom cubicle, and checks a man wearing a light grey suit. Suddenly, the grey-suit man wakes up, and they fight together in the washroom. Finally, the male secret agent wins the fight. **Activity 3:** A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom. Then, the male secret agent lifts an insensible man wearing a light grey suit to the washroom cubicle. Finally, the male secret agent lefts the washroom after talking with the brown-suit woman.

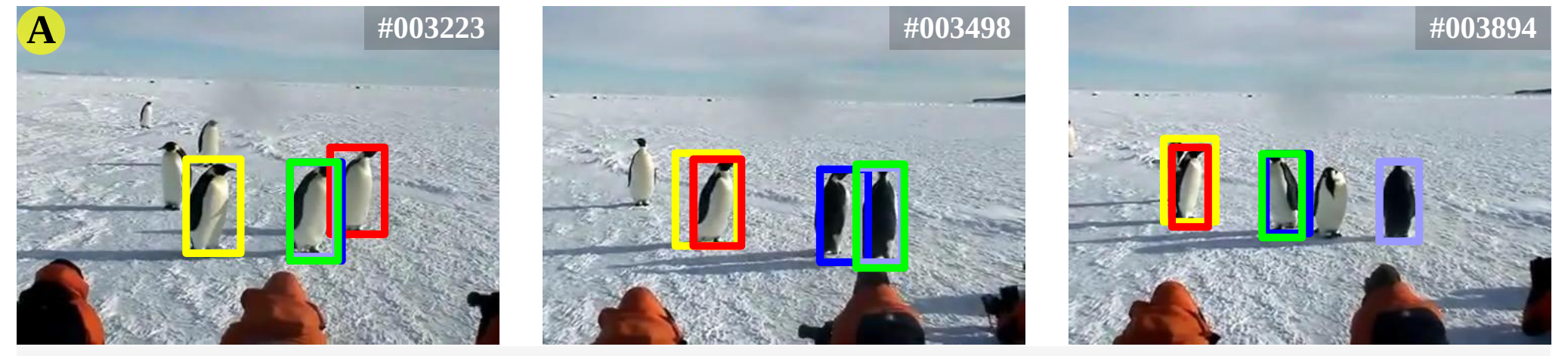
D: **Story:** A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, and the male secret agent wins. He then lifts the insensible grey-suit man to the washroom cubicle. The male secret agent crouches in the washroom cubicle and checks the insensible grey-suit man. Suddenly, the grey-suit man wakes up, and they fight together again in the washroom. Eventually, the male secret agent wins the fight. After the male secret agent talks with a woman wearing a brown suit, he again lifts the insensible grey-suit man to the washroom cubicle. Finally, the male secret agent lefts the washroom after talking with the brown-suit woman.

Experiments

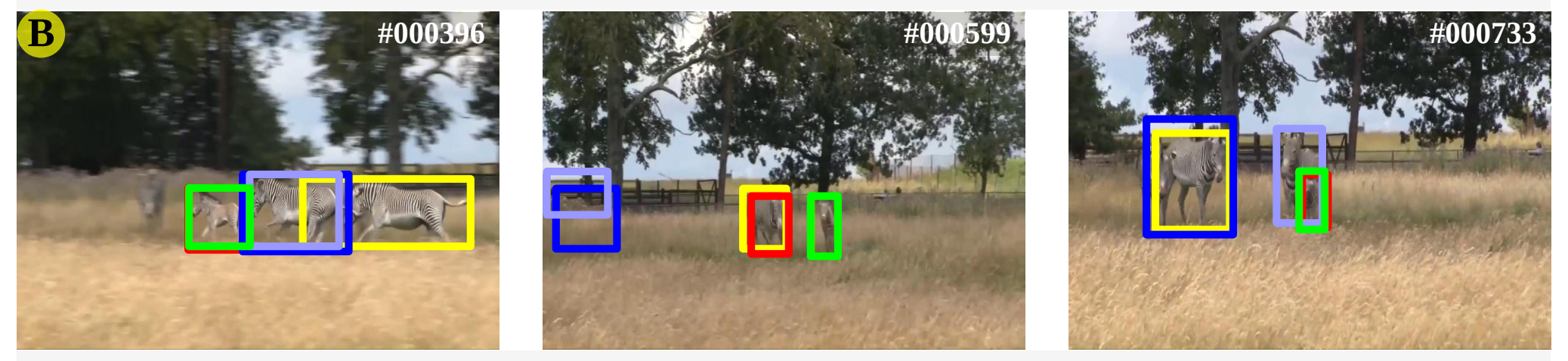
Results on different multi-modal benchmarks (mechanism A)

Tracker	OTB-Lang [1]		TNL2k [3]		LaSOT [2]		LaSOTExt [17]		LaSOTSub		LaSOTNLC		MGIT	
	PRE	SR	PRE	SR	PRE	SR	PRE	SR	PRE	SR	PRE	SR	PRE	SR
SNLT [46]	0.848	0.666	0.081	0.100	0.475	0.459	0.306	0.262	0.527	0.495	0.513	0.483	0.004	0.036
VLT_SCAR [42]	0.898	0.739	0.556	0.497	0.677	0.630	0.503	0.428	0.670	0.633	0.659	0.633	0.124	0.177
VLT_TT [42]	0.931	0.764	0.583	0.539	0.714	0.670	0.549	0.465	0.707	0.660	0.721	0.662	0.324	0.474
JointNLT [18]	0.856	0.653	0.598	0.552	0.640	0.607	0.457	0.398	0.624	0.583	0.707	0.651	0.433	0.603

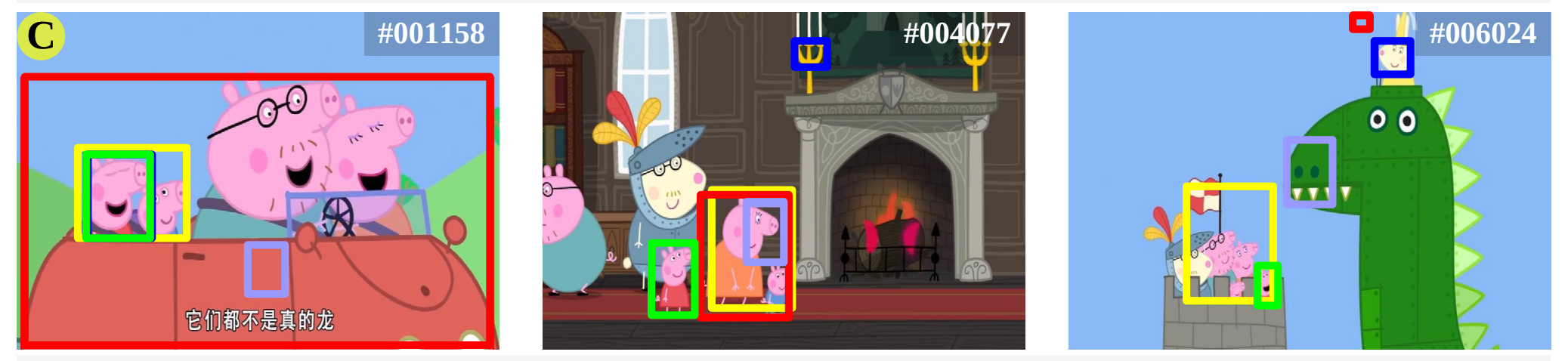
Bad case analysis



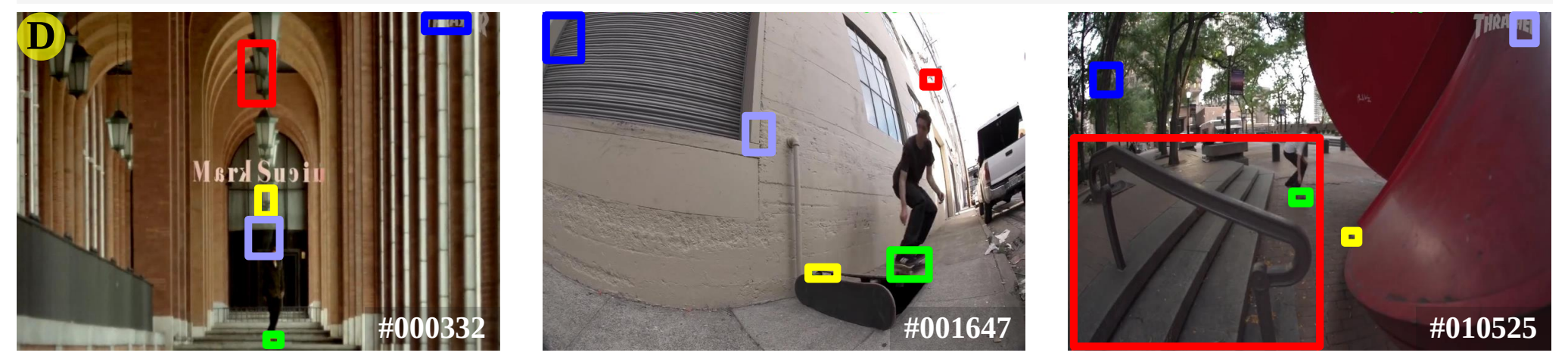
LaSOT bird-2 sequence: white bird walking on the among other birds



LaSOT zebra-16 sequence: zebra running on dry grass with other zebras



MGIT 012 sequence (mechanism A): A pink cartoon pig wearing red clothes talks to her family members on the grassland.



MGIT 006 sequence (mechanism A): A skateboard is slid by a man in black on the playground.

🟩GroundTruth 🟨JointNLT (CVPR23) 🟦VLT_SCAR (NeurIPS22) 🟥VLT_TT (NeurIPS22) 🟪SNLT (CVPR21)

Results of different trackers on MGIT (mechanism B-E)

Tracker	Architecture	Initialize	Mechanism	PRE	N-PRE	SR
SiamCAR [11]	SNN	BBox	E	0.116	0.378	0.183
SiamRCNN [10]	SNN	BBox		0.512	0.707	0.591
PrDiMP [12]	SNN+CF	BBox		0.296	0.602	0.453
KeepTrack [13]	SNN+CF	BBox		0.373	0.695	0.519
TransT [39]	Transformer	BBox		0.447	0.670	0.539
MixFormer [14]	Transformer	BBox		0.526	0.775	0.629
OSTrack [15]	Transformer	BBox		0.476	0.706	0.583
GRM [16]	Transformer	BBox		0.500	0.718	0.597
SNLT [46]	SNN	NL&BBox	Action (B)	0.004	0.226	0.036
			Activity (C)	0.004	0.234	0.038
			Story (D)	0.005	0.230	0.040
VLT_SCAR [42]	SNN	NL&BBox	Action (B)	0.116	0.354	0.167
			Activity (C)	0.124	0.382	0.180
			Story (D)	0.127	0.403	0.184
VLT_TT [42]	Transformer	NL&BBox	Action (B)	0.318	0.602	0.468
			Activity (C)	0.325	0.627	0.485
			Story (D)	0.322	0.616	0.480
JointNLT [18]	Transformer	NL&BBox	Action (B)	0.445	0.786	0.610
			Activity (C)	0.441	0.780	0.605
			Story (D)	0.433	0.773	0.600

Conclusions

- MGIT is a complex environment, the annotation strategy is a feasible solution for coupling human understanding into semantic labels.
- Existing trackers should improve the capability for processing long text and aligning multi-modal information.