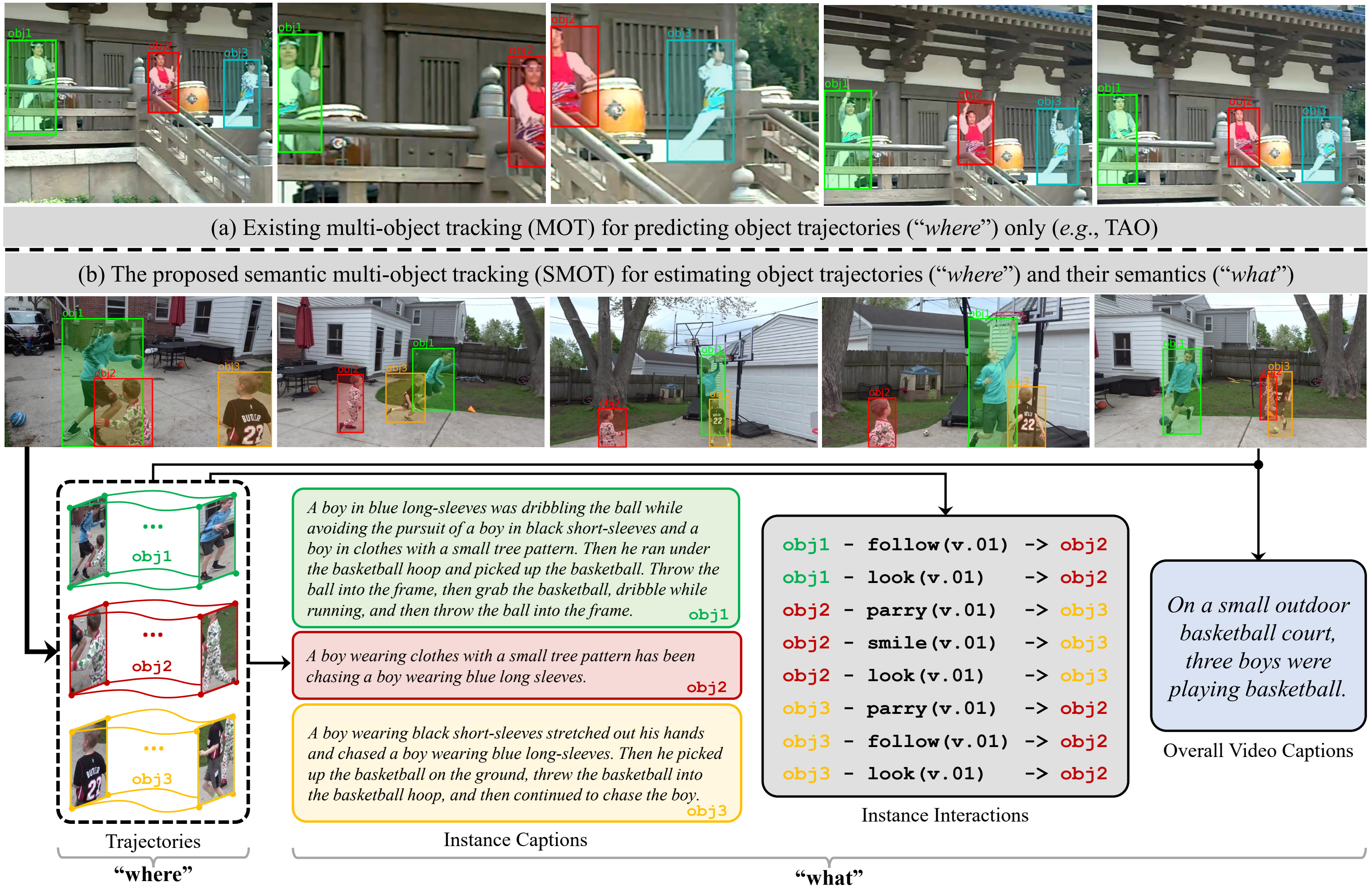
Beyond MOT: Semantic Multi-Object Tracking

Yunhao Li [1,2], Qin Li [1], Hao Wang [2], Xue Ma [1], Jiali Yao [3], Shaohua Dong [4],
Heng Fan [4,*], Libo Zhang [1,2,3,*,†]

[1] Institute of Software Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
[4] Department of Computer Science & Engineering, University of North Texas

Semantic Multi-Object Tracking （SMOT）



(a) Existing multi-object tracking (MOT) for predicting object trajectories (“*where*”) only (*e.g.*, TAO)

(b) The proposed semantic multi-object tracking (SMOT) for estimating object trajectories (“*where*”) and their semantics (“*what*”)

A boy in blue long-sleeves was dribbling the ball while avoiding the pursuit of a boy in black short-sleeves and a boy in clothes with a small tree pattern. Then he ran under the basketball hoop and picked up the basketball. Throw the ball into the frame, then grab the basketball, dribble while running, and then throw the ball into the frame. obj1

A boy wearing clothes with a small tree pattern has been chasing a boy wearing blue long sleeves. obj2

A boy wearing black short-sleeves stretched out his hands and chased a boy wearing blue long-sleeves. Then he picked up the basketball on the ground, threw the basketball into the basketball hoop, and then continued to chase the boy. obj3

obj1 - follow(v.01) -> obj2
obj1 - look(v.01) -> obj2
obj2 - parry(v.01) -> obj3
obj2 - smile(v.01) -> obj3
obj2 - look(v.01) -> obj3
obj3 - parry(v.01) -> obj2
obj3 - follow(v.01) -> obj2
obj3 - look(v.01) -> obj2

On a small outdoor basketball court, three boys were playing basketball.

Trajectories

Instance Captions

Instance Interactions

Overall Video Captions

“where”

“what”

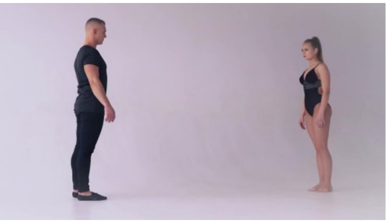
Extend the MOT task from merely “*where*” to “*what*”

Benchmark for SMOT （BenSMOT）

Attending_zoo_museum

Bandage

Camping


Dance

Decorating

Dressing_bathing_child

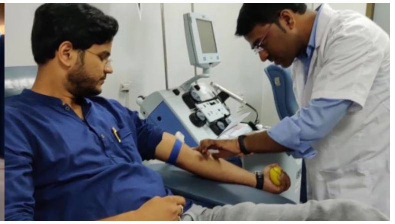
Baptism

Bathing_dog

Building_house

Cutting_child_hair

Dancing_with_child


Denoting_blood

Facial

Interview

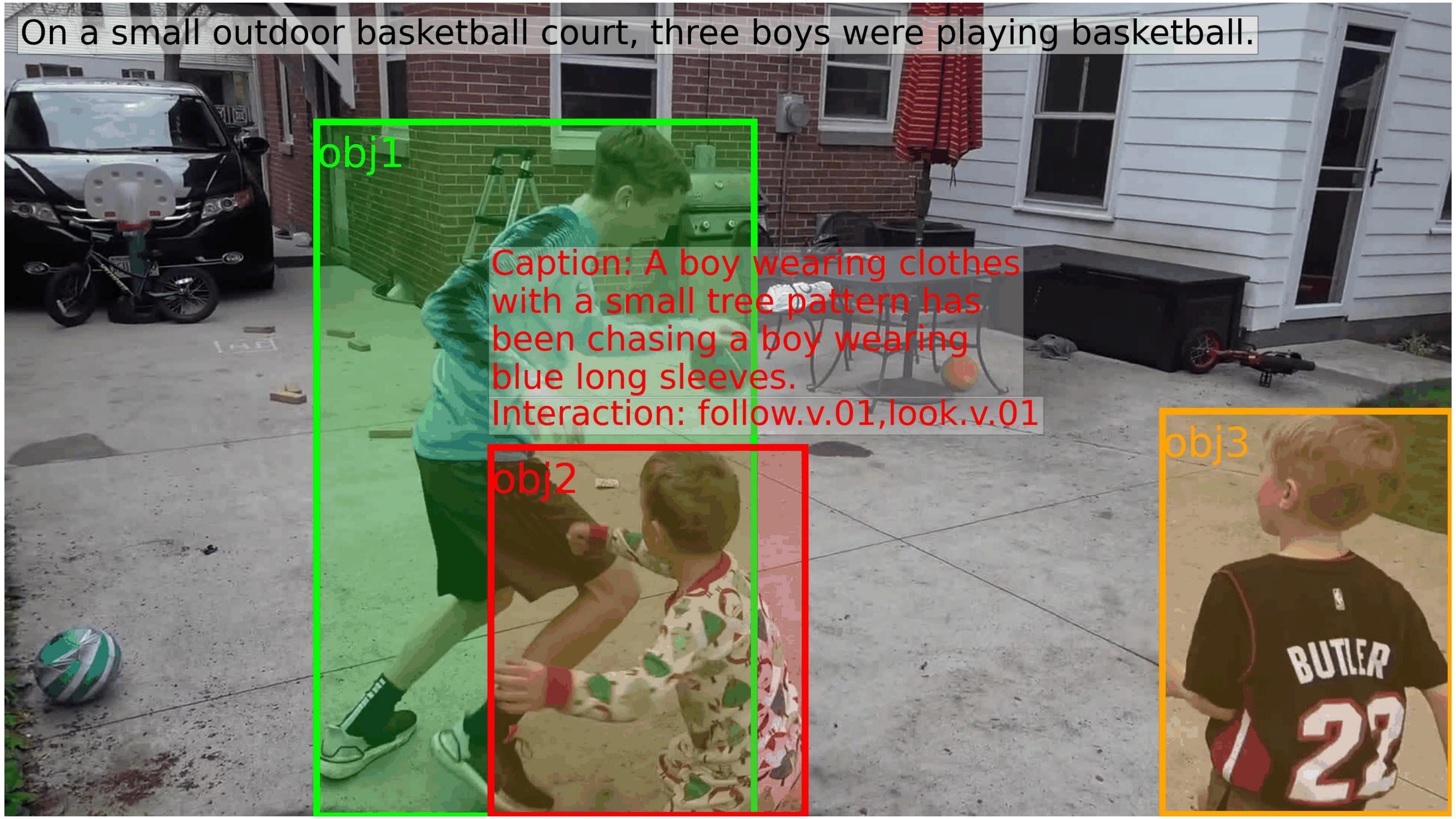
Helping_homelee_people

Haircut

Feeding_child

Grooming_pets

Benchmark for SMOT （BenSMOT）



On a small outdoor basketball court, three boys were playing basketball.

Caption: A boy wearing clothes with a small tree pattern has been chasing a boy wearing blue long sleeves.

Interaction: follow.v.01,look.v.01

Benchmark for SMOT （BenSMOT）

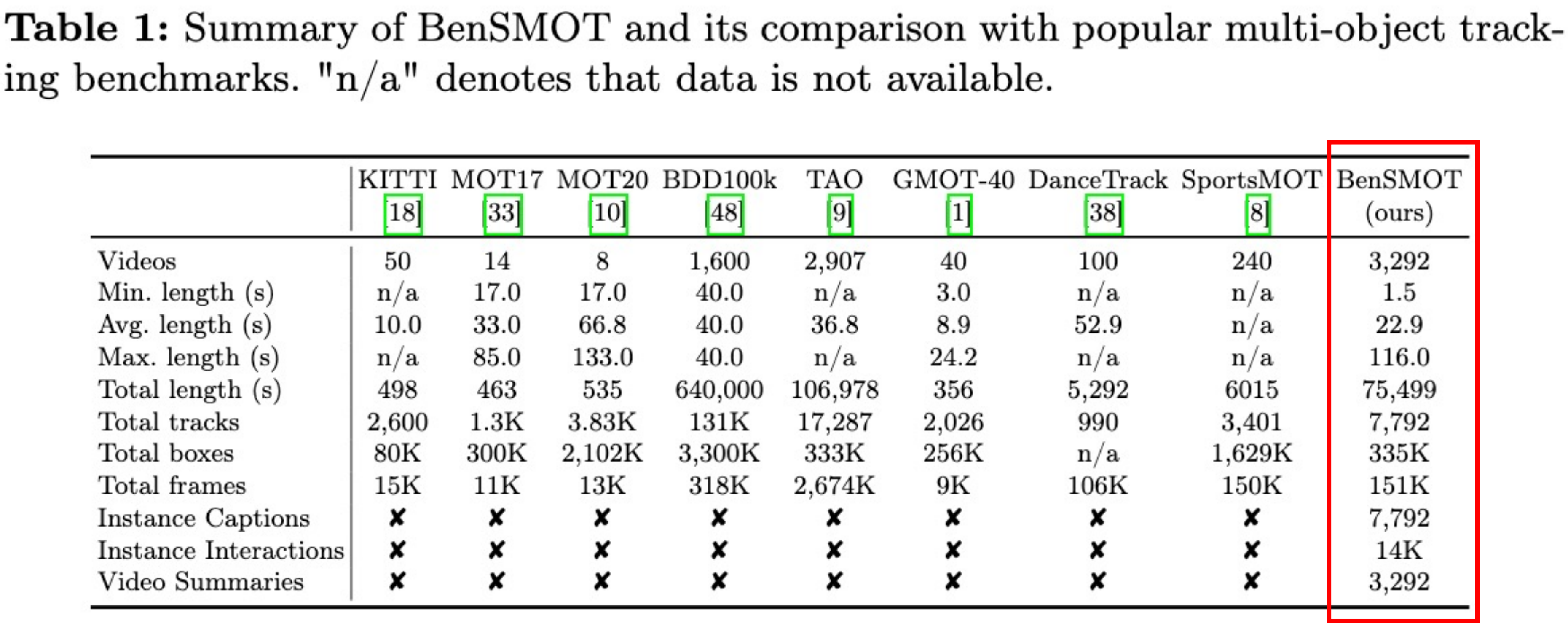


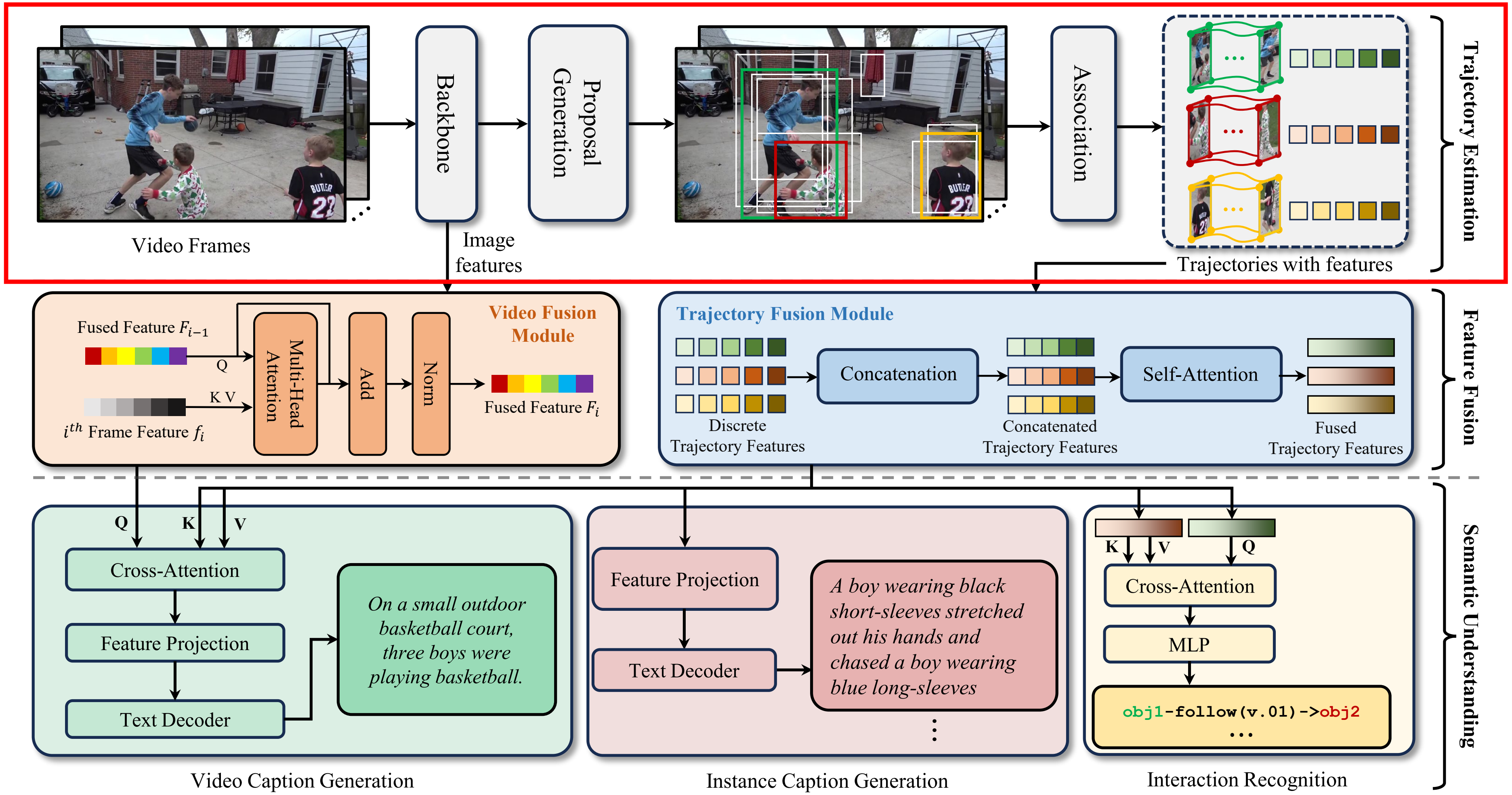
Table 1: Summary of BenSMOT and its comparison with popular multi-object tracking benchmarks. "n/a" denotes that data is not available.

	KITTI	MOT17	MOT20	BDD100k	TAO	GMOT-40	DanceTrack	SportsMOT	BenSMOT
	[18]	[33]	[10]	[48]	[9]	[1]	[38]	[8]	(ours)
Videos	50	14	8	1,600	2,907	40	100	240	3,292
Min. length (s)	n/a	17.0	17.0	40.0	n/a	3.0	n/a	n/a	1.5
Avg. length (s)	10.0	33.0	66.8	40.0	36.8	8.9	52.9	n/a	22.9
Max. length (s)	n/a	85.0	133.0	40.0	n/a	24.2	n/a	n/a	116.0
Total length (s)	498	463	535	640,000	106,978	356	5,292	6015	75,499
Total tracks	2,600	1.3K	3.83K	131K	17,287	2,026	990	3,401	7,792
Total boxes	80K	300K	2,102K	3,300K	333K	256K	n/a	1,629K	335K
Total frames	15K	11K	13K	318K	2,674K	9K	106K	150K	151K
Instance Captions	✘	✘	✘	✘	✘	✘	✘	✘	7,792
Instance Interactions	✘	✘	✘	✘	✘	✘	✘	✘	14K
Video Summaries	✘	✘	✘	✘	✘	✘	✘	✘	3,292

We collect video templates from online video platforms and manually label them with four types of annotations, including *bounding box*, *instance caption*, *instance interaction*, and the *overall video caption*.

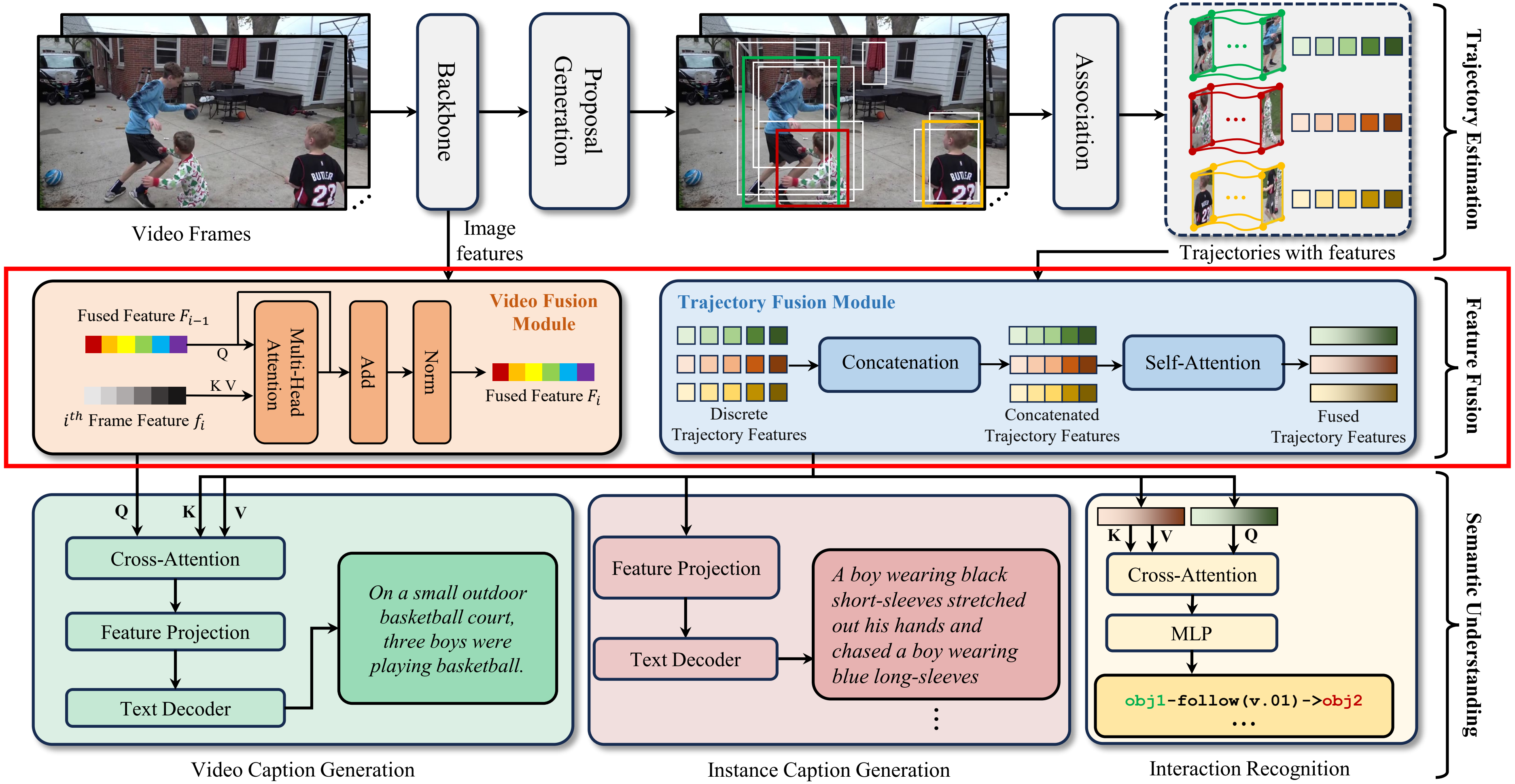
Model Architecture

Trajectory Estimation



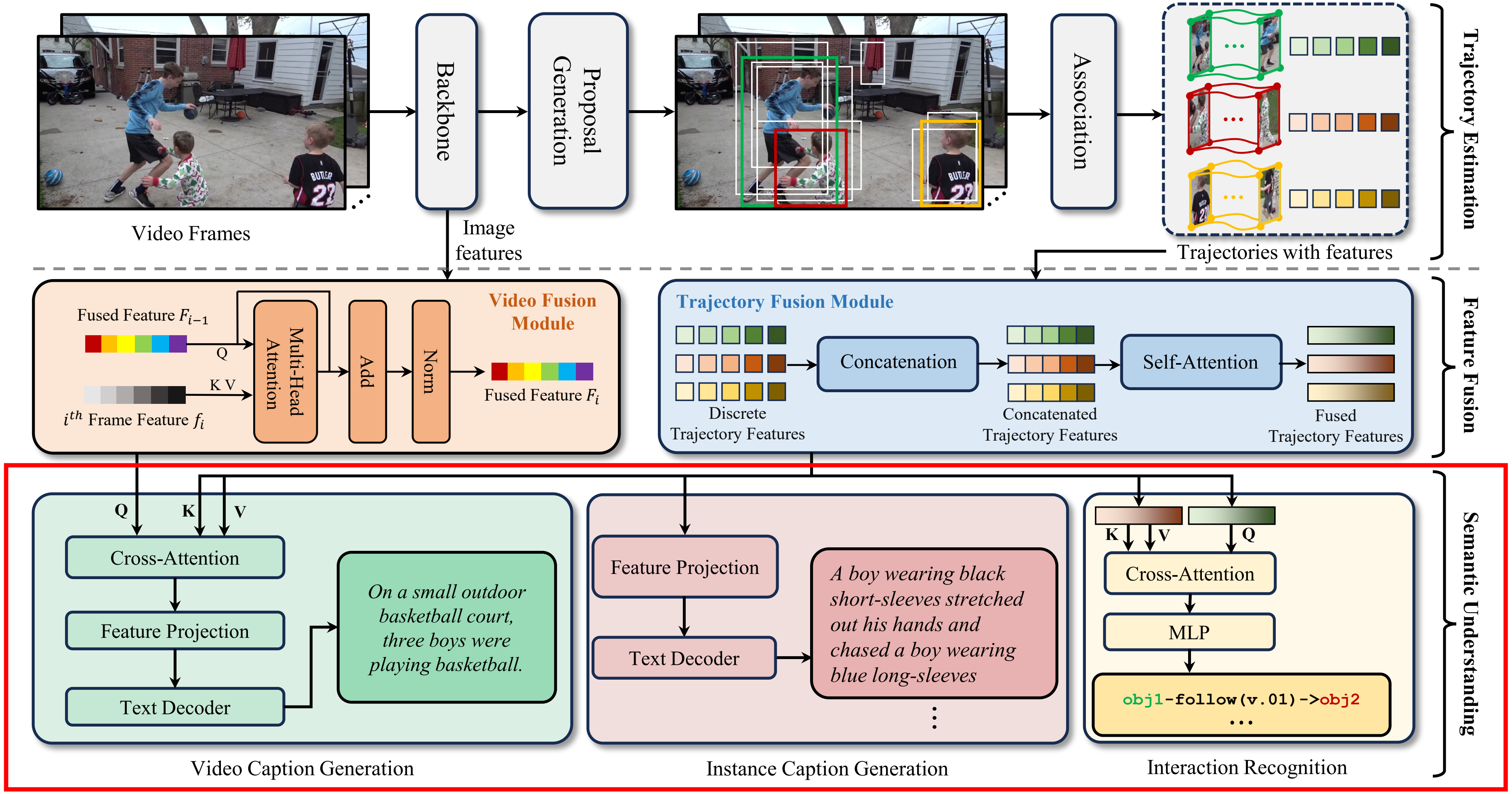
Model Architecture

Feature Fusion



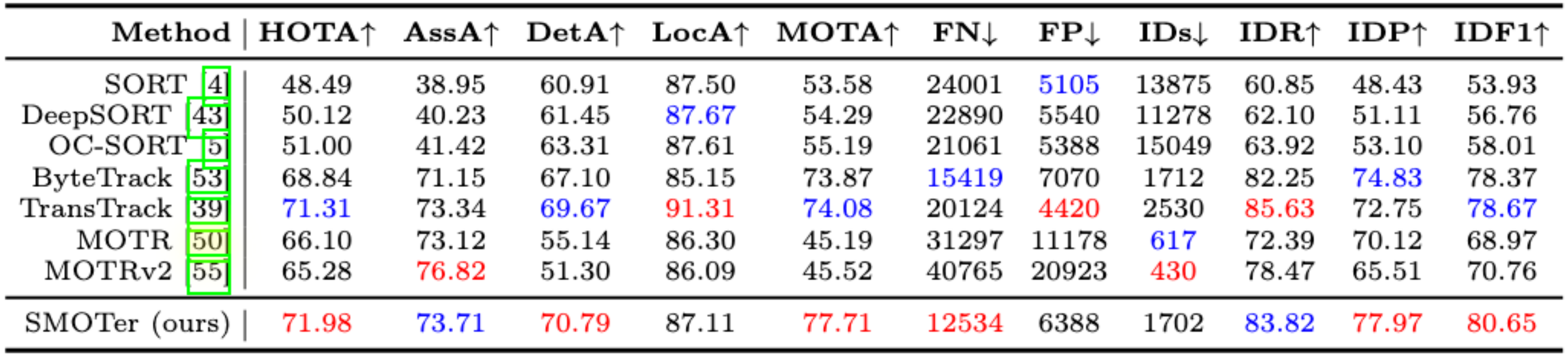
Model Architecture

Semantic Understanding



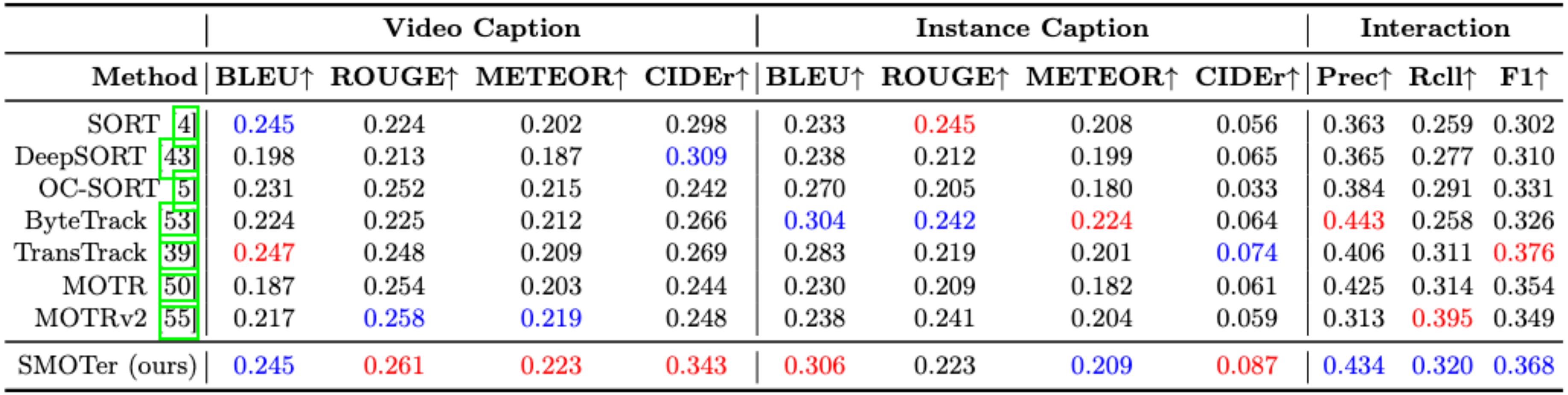
Experiments

- Comparison between SMOTer and two-stage MOT methods regarding tracking performance on BenSMOT.



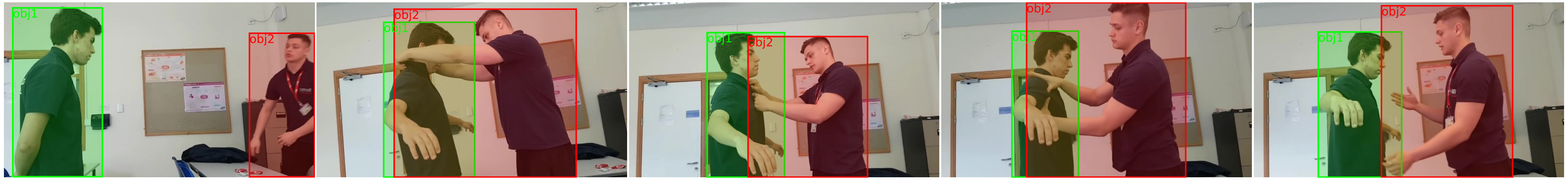
Method	HOTA↑	AssA↑	DetA↑	LocA↑	MOTA↑	FN↓	FP↓	IDs↓	IDR↑	IDP↑	IDF1↑
SORT [4]	48.49	38.95	60.91	87.50	53.58	24001	5105	13875	60.85	48.43	53.93
DeepSORT [43]	50.12	40.23	61.45	87.67	54.29	22890	5540	11278	62.10	51.11	56.76
OC-SORT [5]	51.00	41.42	63.31	87.61	55.19	21061	5388	15049	63.92	53.10	58.01
ByteTrack [53]	68.84	71.15	67.10	85.15	73.87	15419	7070	1712	82.25	74.83	78.37
TransTrack [39]	71.31	73.34	69.67	91.31	74.08	20124	4420	2530	85.63	72.75	78.67
MOTR [50]	66.10	73.12	55.14	86.30	45.19	31297	11178	617	72.39	70.12	68.97
MOTRv2 [55]	65.28	76.82	51.30	86.09	45.52	40765	20923	430	78.47	65.51	70.76
SMOTer (ours)	71.98	73.71	70.79	87.11	77.71	12534	6388	1702	83.82	77.97	80.65

- Comparison of SMOTer against two-stage methods based on MOT models regarding semantic understanding.



	Video Caption				Instance Caption				Interaction		
Method	BLEU↑	ROUGE↑	METEOR↑	CIDEr↑	BLEU↑	ROUGE↑	METEOR↑	CIDEr↑	Prec↑	Rcll↑	F1↑
SORT [4]	0.245	0.224	0.202	0.298	0.233	0.245	0.208	0.056	0.363	0.259	0.302
DeepSORT [43]	0.198	0.213	0.187	0.309	0.238	0.212	0.199	0.065	0.365	0.277	0.310
OC-SORT [5]	0.231	0.252	0.215	0.242	0.270	0.205	0.180	0.033	0.384	0.291	0.331
ByteTrack [53]	0.224	0.225	0.212	0.266	0.304	0.242	0.224	0.064	0.443	0.258	0.326
TransTrack [39]	0.247	0.248	0.209	0.269	0.283	0.219	0.201	0.074	0.406	0.311	0.376
MOTR [50]	0.187	0.254	0.203	0.244	0.230	0.209	0.182	0.061	0.425	0.314	0.354
MOTRv2 [55]	0.217	0.258	0.219	0.248	0.238	0.241	0.204	0.059	0.313	0.395	0.349
SMOTer (ours)	0.245	0.261	0.223	0.343	0.306	0.223	0.209	0.087	0.434	0.320	0.368

Visualization Results



GroundTruth: A man in the dark blue shirt with pimples on his face raises his arms to his sides as instructed by the man with the ID in front of him and patiently submits to being examined by the man with the ID.

Prediction: *In a black short-sleeved shirt holds a pair of scissors in her right hand, and a comb in her right hand, combing the man in a black scarf.*

obj1 caption

GroundTruth: A man wearing a dark blue shirt and a work permit around his neck asks the man with the pimples to raise his arms, first turning his collar with both hands and then pressing on his left and right sleeves and cuffs.

Prediction: *Wearing a black short-sleeved shirt with yellow letters checking the back of a man wearing a black short-sleeved.*

obj2 caption

GroundTruth: In a room, a man asks another man to raise his arms flat and perform a security check.

Prediction: *In a room, a man is tutoring a man.*

video caption

GroundTruth:

```
obj1 -> look.v.01  -> obj2
obj2 -> look.v.01  -> obj1
obj2 -> talk.v.02  -> obj1
obj2 -> frisk.v.02 -> obj1
```

Prediction:

```
obj1 -> look.v.01  -> obj2
obj2 -> look.v.01  -> obj1
obj2 -> talk.v.02  -> obj1
```

interaction

Thank You for Your Attention !