Are Socially-aware Human Trajectory Prediction Models Really Socially-aware?

Saeed Saadatnejad*,1 , Mohammadhossein Bahari *,1, Pedram Khorsandi 2,
Mohammad Saneian 2, Seyed-Mohsen Moosavi-Dezfooli 3, Alexandre Alahi 1
[1] *EPFL*, [2] *Sharif university of technology*, [3] *ETH Zurich*

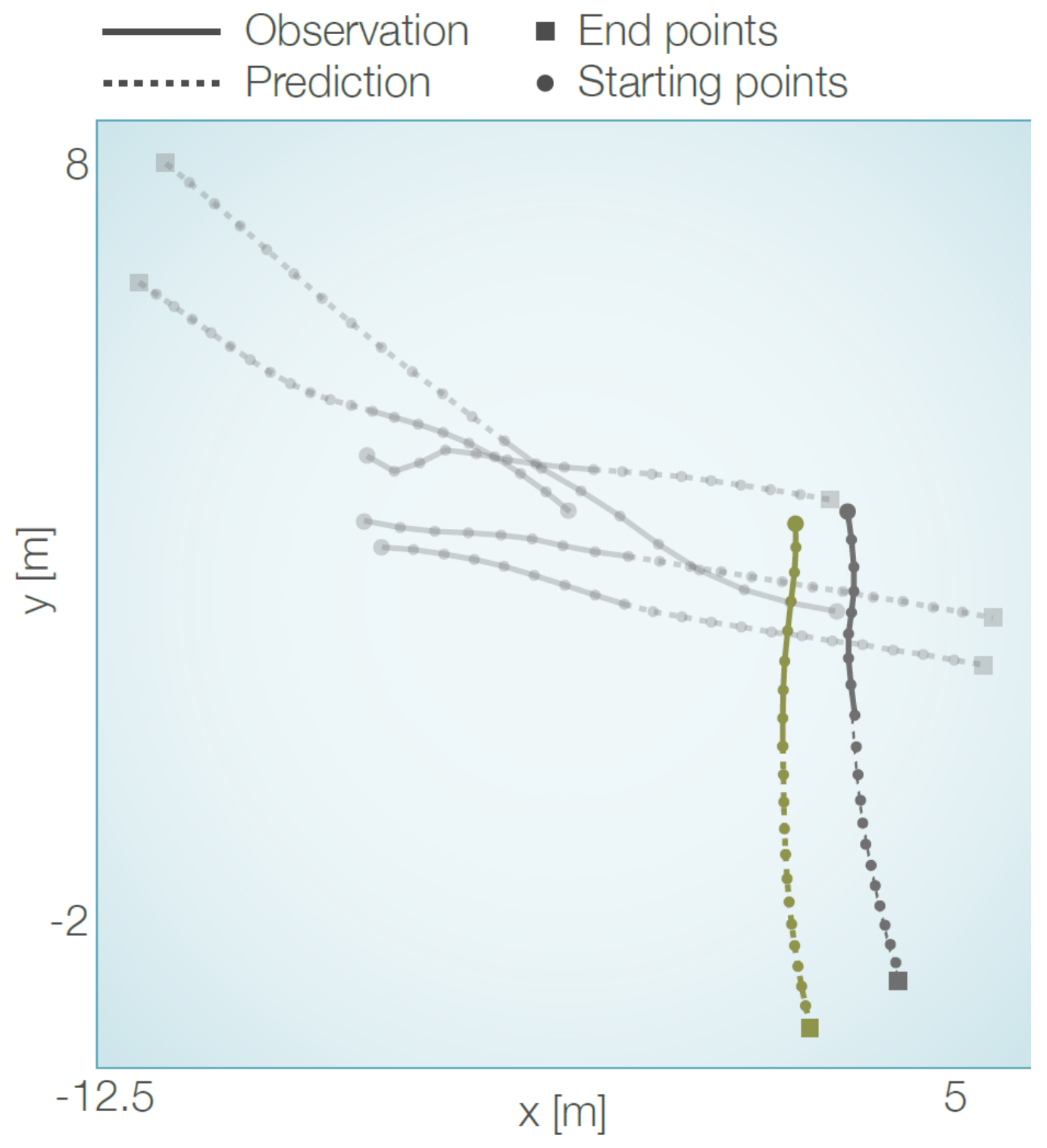
2021 ICCV OCTOBER 11-17 VIRTUAL

Abstract:

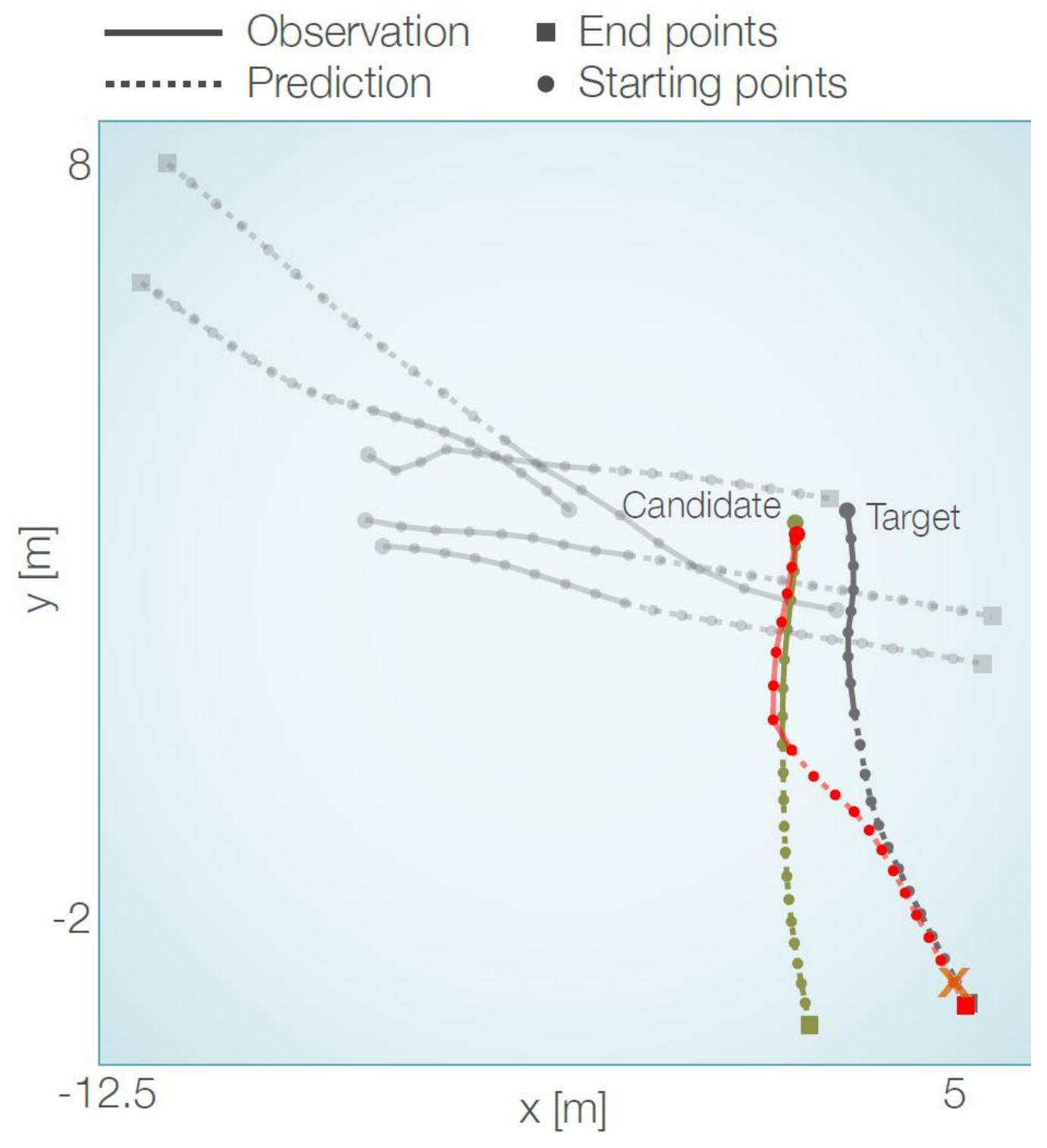
- Our field has recently witnessed an arms race of neural network-based trajectory predictors. While these predictors are at the core of many applications, their adversarial robustness has not been carefully studied. In this paper, we introduce a socially-attended attack to assess the social understanding of prediction models in terms of collision avoidance.
- An attack is a small yet carefully-crafted perturbation to fail predictors. Technically, we define collision as a failure mode of the output, and propose attention-based mechanisms to guide our attack.
- Thanks to our attack, we shed light on the limitations of the current models in terms of their social understanding. Finally, we show that our attack can be employed to increase the social understanding of state-of-the-art models

Motivation:

- Diverse approaches based on neural networks are proposed to learn human social behavior and showed to have accurate predictions (Fig a)
- But with a slight perturbation, can current prediction models still predict reasonably? (Fig b)
- Motivations:
  (1) it is an evaluation method for the previously-proposed predictors.
  (2) leverage adversarial examples to train models with better collision-avoidance.



(a) No perturbation



(b) With perturbation

Method:

- Collision can happen at any timestep between any two agents,
- The choice of them impacts the final perturbation's size,
- To address that, we introduce an attention-guided adversarial attack, named SATTack, which learns the best collision points:

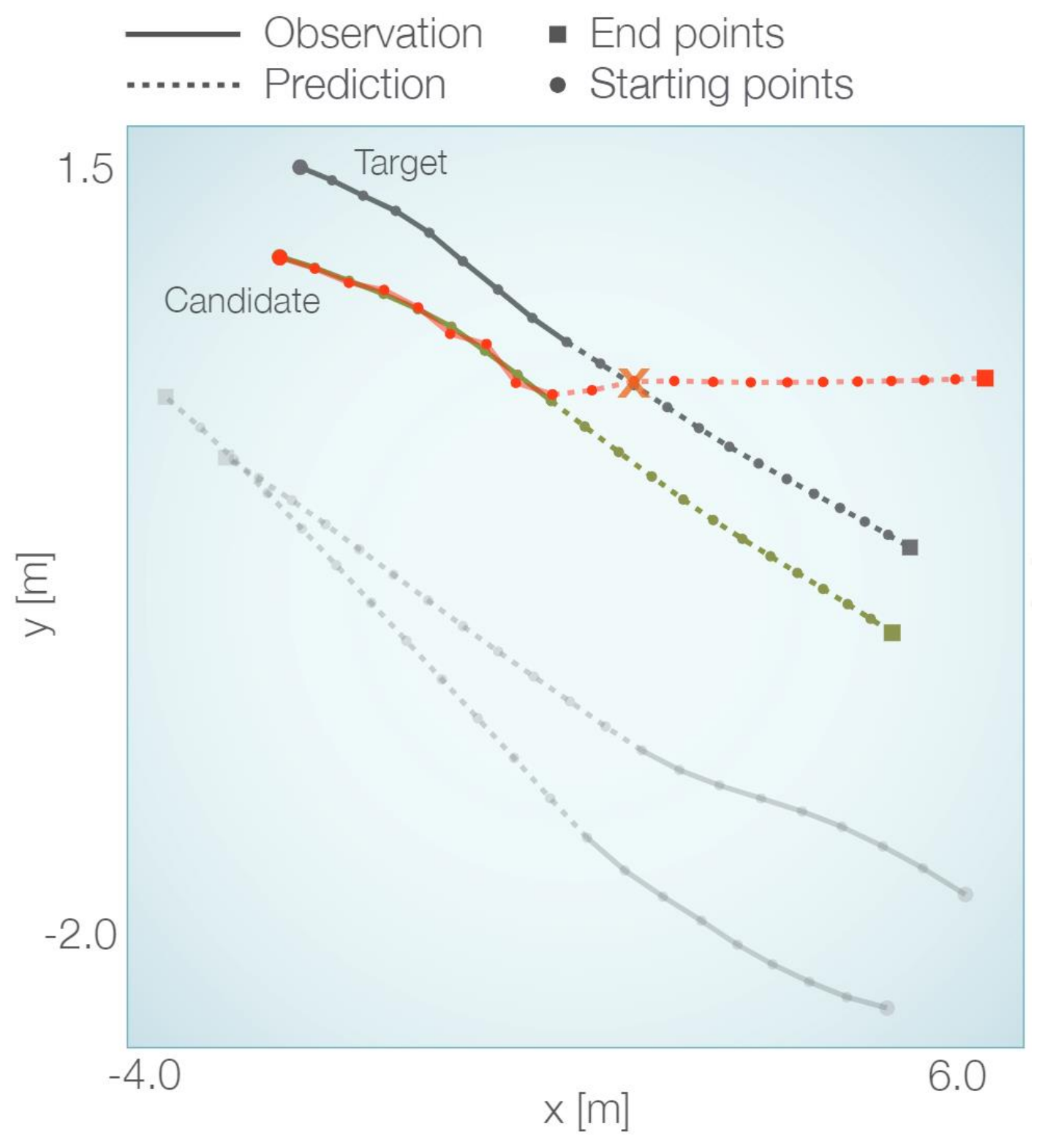
$$\min_{R,W} \operatorname{Tr}\left(W^{\top}\tanh(D(R))\right) + \lambda_r \|R\|_F - \lambda_w \|W\|_F\,,$$

$$\text{s.t.} \sum_{j,t} w_{j,t} = 1, \quad w_{j,t} \geq 0,$$

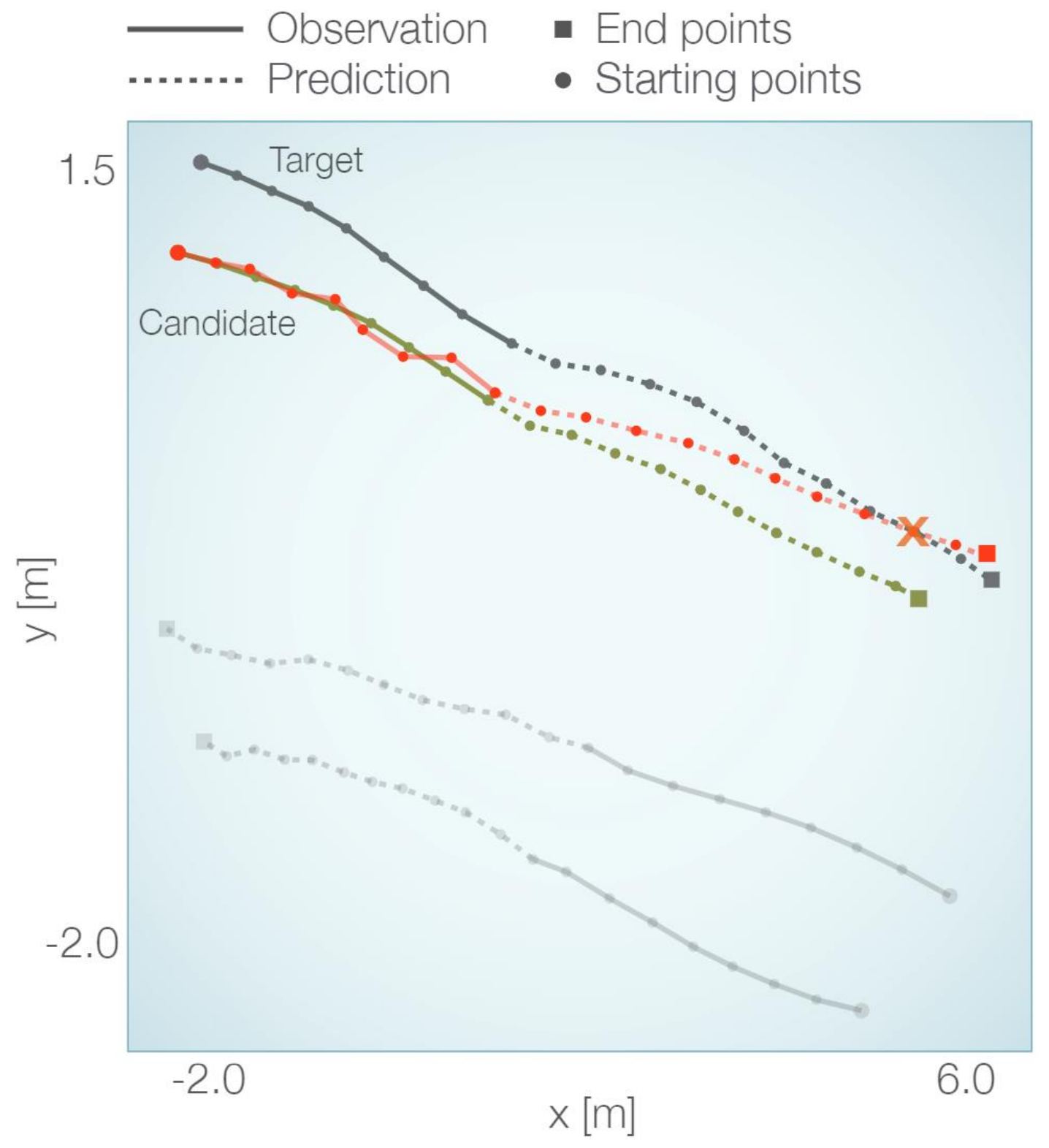
- W is the attention weight matrix and $w_{j,t}$ is the attention weight for the agent j at timestep t,
- R is the perturbation,
- $D(R)$ is the distance matrix, $d_{j,t} = \hat{Y}_t^j - \hat{Y}_t^1$, Y are predictions

Results:

Model	Original CR [%] ↓	Attacked CR [%] ↓	P-avg [m] ↓
S-LSTM [1]	7.8	89.8	0.031
S-Att [45]	9.4	86.4	0.057
S-GAN [20]	13.9	85.0	0.034
D-Pool [25]	7.3	88.0	0.042
S-STGCNN [32]	16.3	59.1	0.11
PECNet [31]	15.0	64.9	0.071



(c) Attack on S-GAN

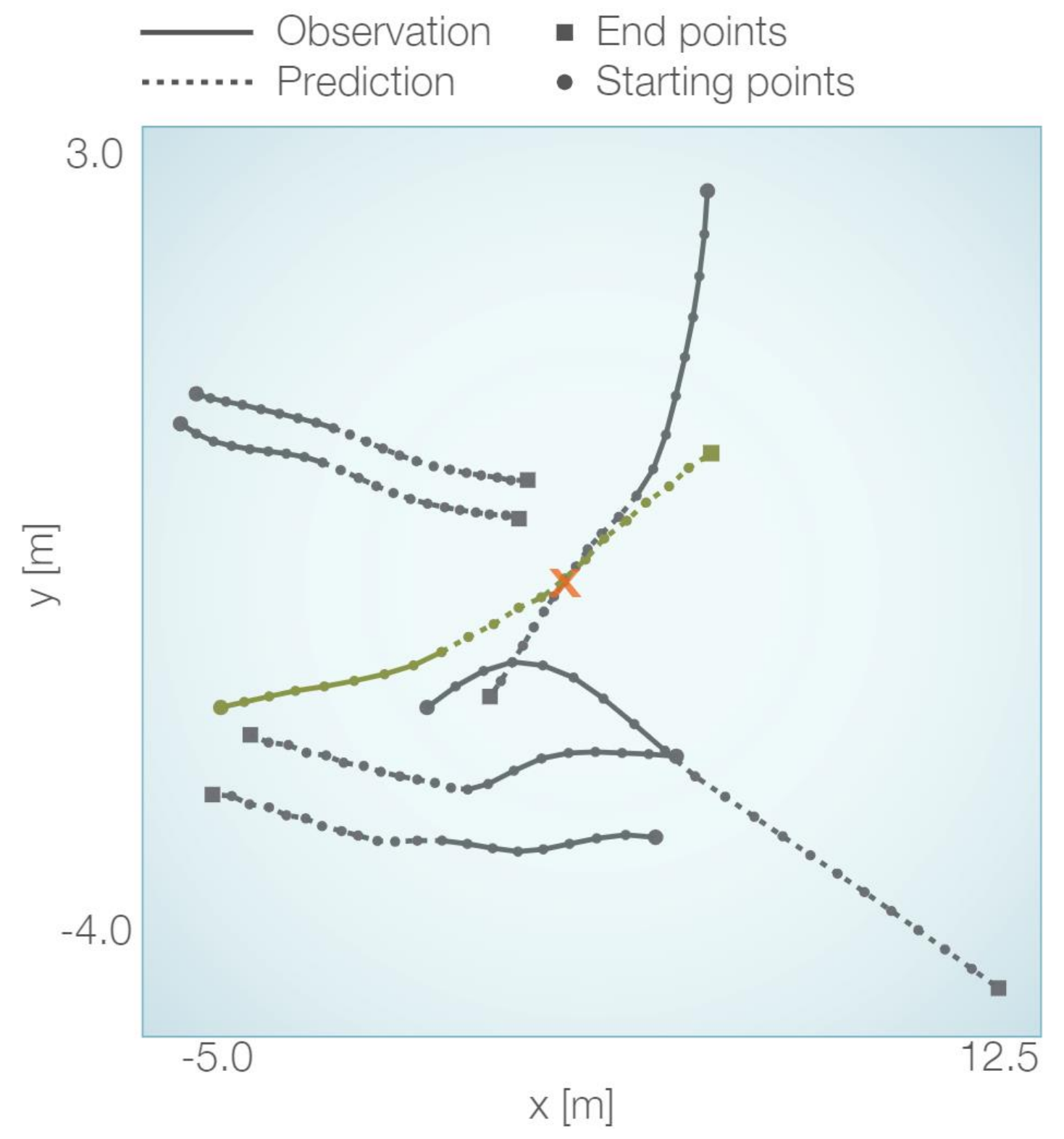


(d) Attack on S-STGCNN

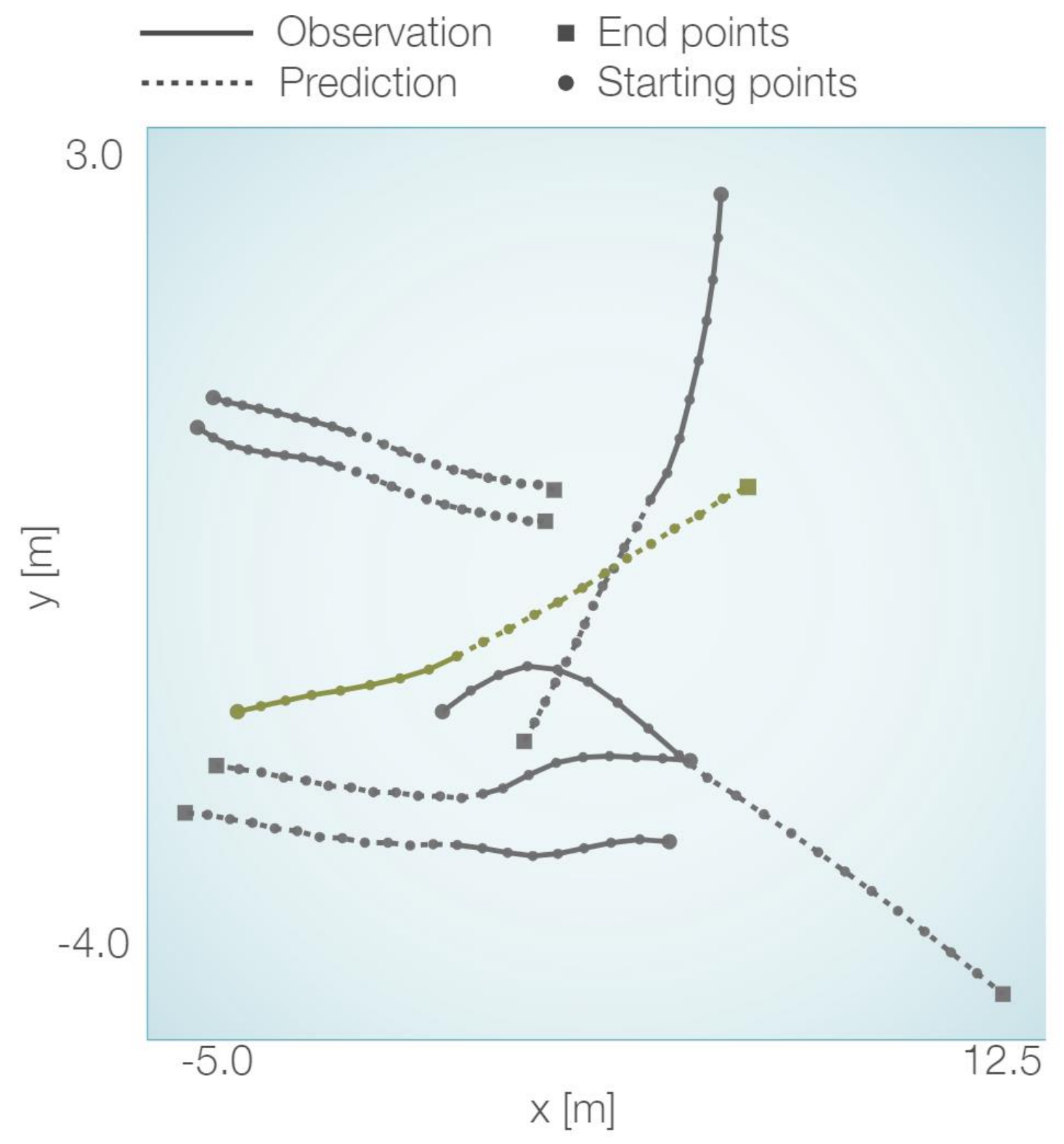
Robustness:

- Model is made robust leveraging the perturbations.

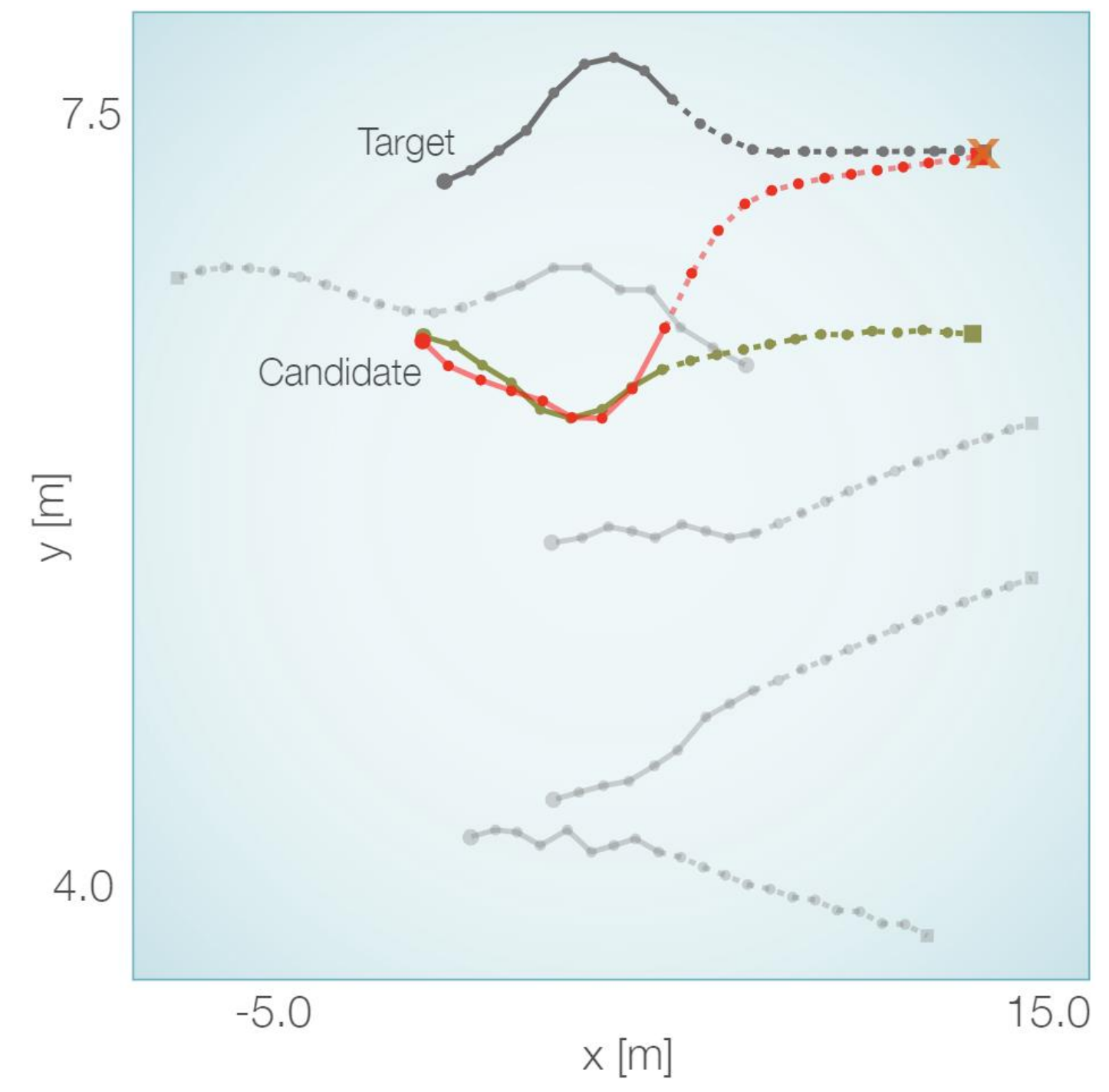
		Original		Attacked	
	ADE/FDE [m] ↓	CR [%] ↓	CR gain [%] ↑	CR [%] ↓	CR gain [%] ↑
D-Pool	0.57 / 1.23	7.3	-	37.3	-
D-Pool w/ rand noise	0.57 / 1.23	7.5	-2.7	36.1	+3.2
D-Pool w/ S-ATTack	0.60 / 1.28	**6.5**	**+10.4**	**14.7**	**+60**



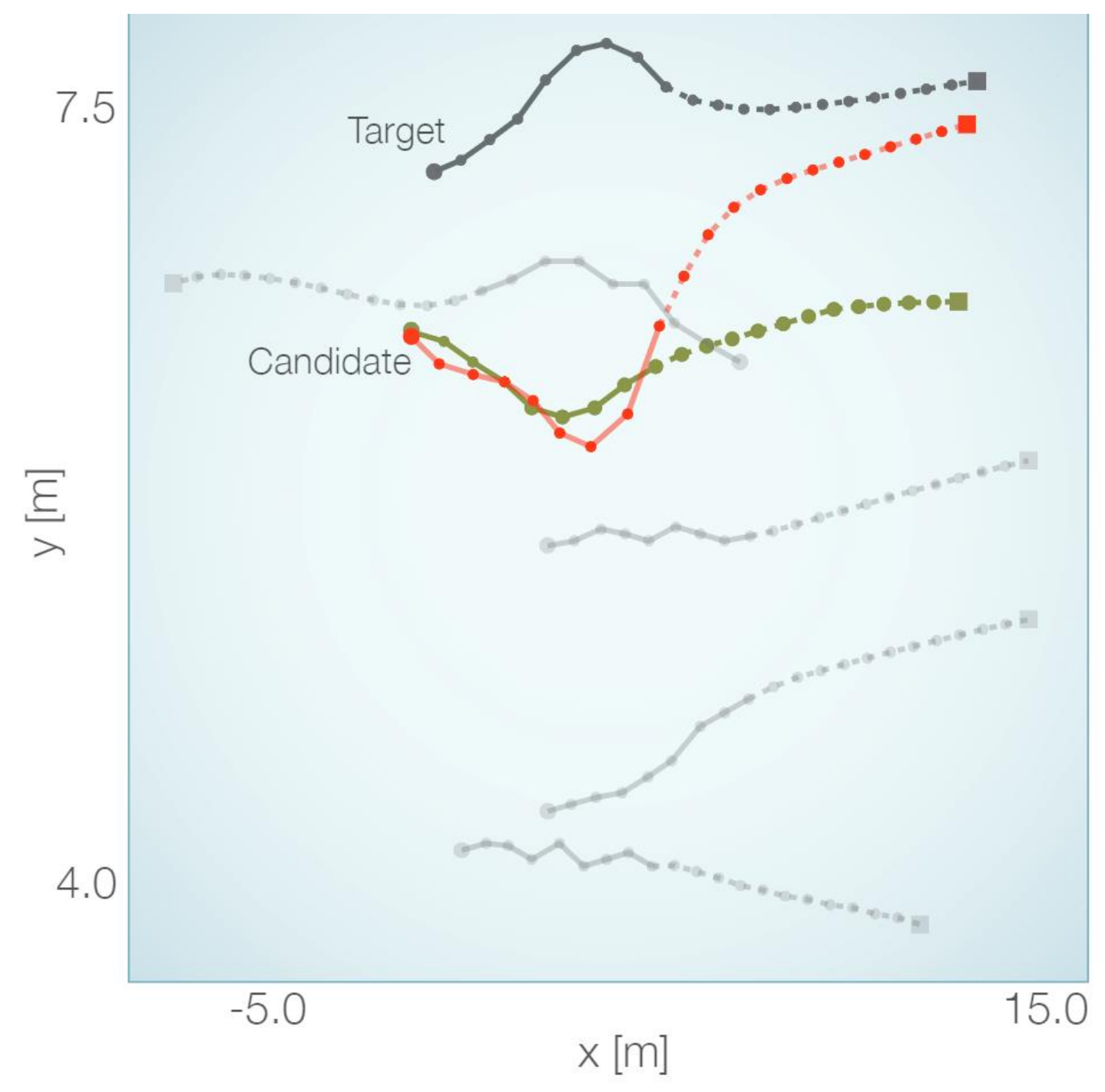
(e) D-Pool (collision)



(f) Robust D-Pool (no collision)



(g) Attack on D-Pool (collision)



(h) Attack on robust D-Pool (no collision)

Code @ https://s-attack.github.io/