Entry-Flipped Transformer for Inference and Prediction of Participant Behavior

Bo Hu[1,2] and Tat-Jen Cham[1,2]

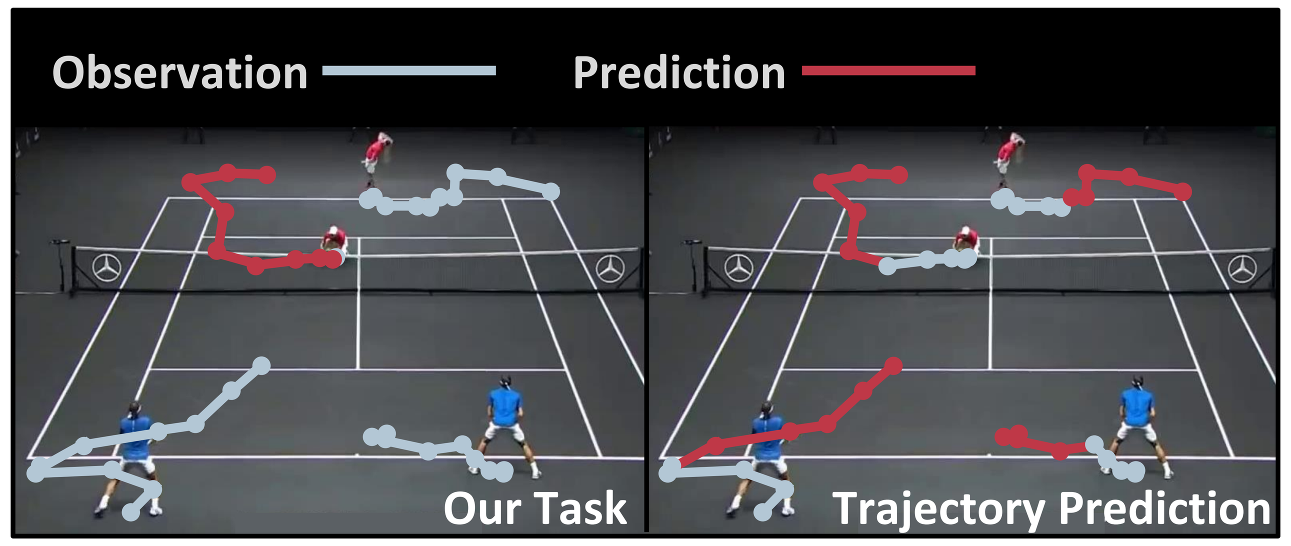
[1] Singtel Cognitive and Artificial Intelligence Lab (SCALE@NTU)

[2] School of Computer Science and Engineering, Nanyang Technological University

Singtel · Nanyang Technological University Singapore · ECCV Tel Aviv 2022

Problem Definition

Participants behavior inference and prediction are to estimate the behavior of a number of target participants in a group, based on information of other observed participants in the same group.



* It is different from trajectory prediction.

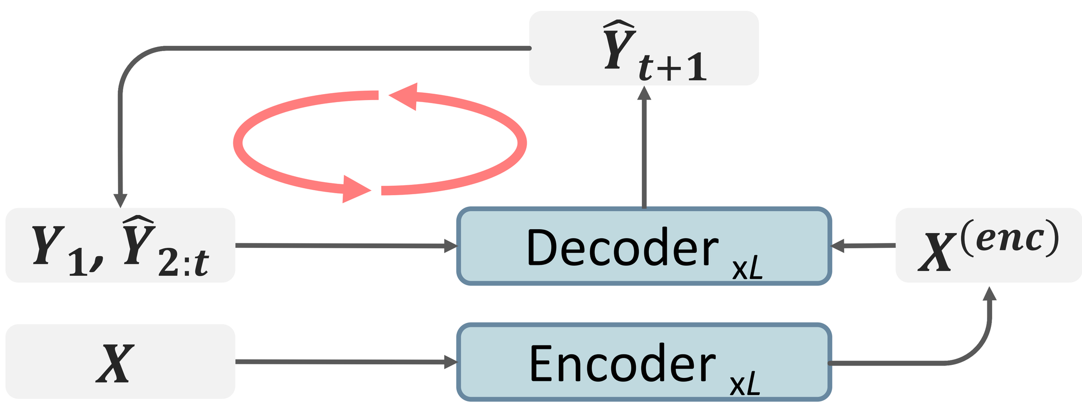
Challenges

1. Participants are highly correlated
2. Trajectories are highly non-smooth
3. Prediction relies on social interaction more than self intention

Motivation

- Tasks of participants behavior inference and prediction can be modeled as an **online frame-wise sequence estimation** problems
- which usually solved by **autoregressive** methods
- Current autoregressive frameworks are likely to introduce error accumulation
  - Seq2Seq (RNN-based)
  - Transformer (Attention-based)

Typical Transformer



* X : Observation participants
Y : Target participants $\hat{Y}$: Estimated target participants

Errors are accumulated in the autoregressive decoding procedure

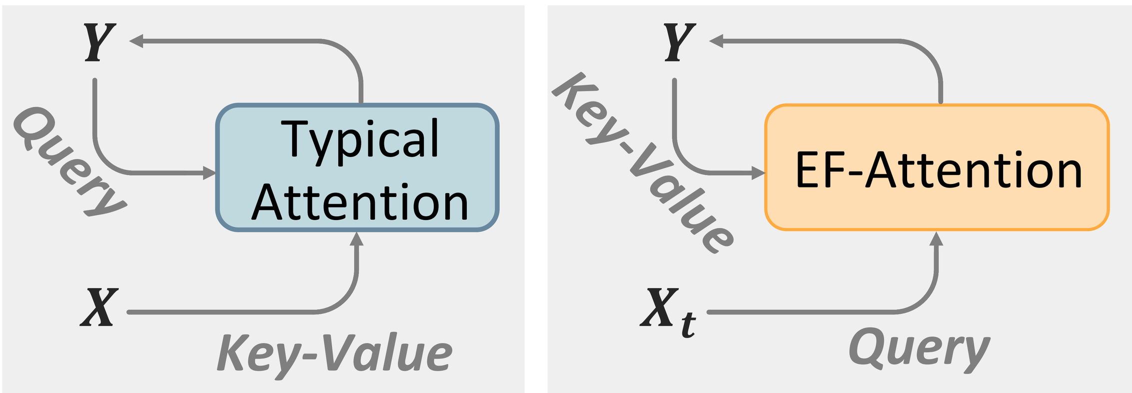
Entry-Flipping Mechanism

Attention function

$$X^{(\text{att})} = f_o\left[\frac{S(f_q(X_q)f_k(X_k)^T)}{\sqrt{d}} f_v(X_v)\right] + X_q$$

Attention is a summarization of query, key, and value entries, where

- Query is the **base**, where the error will be accumulated
- Key and value are **references**, where the error can be suppressed by low attention weight



Entry-Flipping in Decoders

1. Flip query and key-value entries
2. Send one-frame observation as query at one time

Analysis

Why is entry-flipping important?

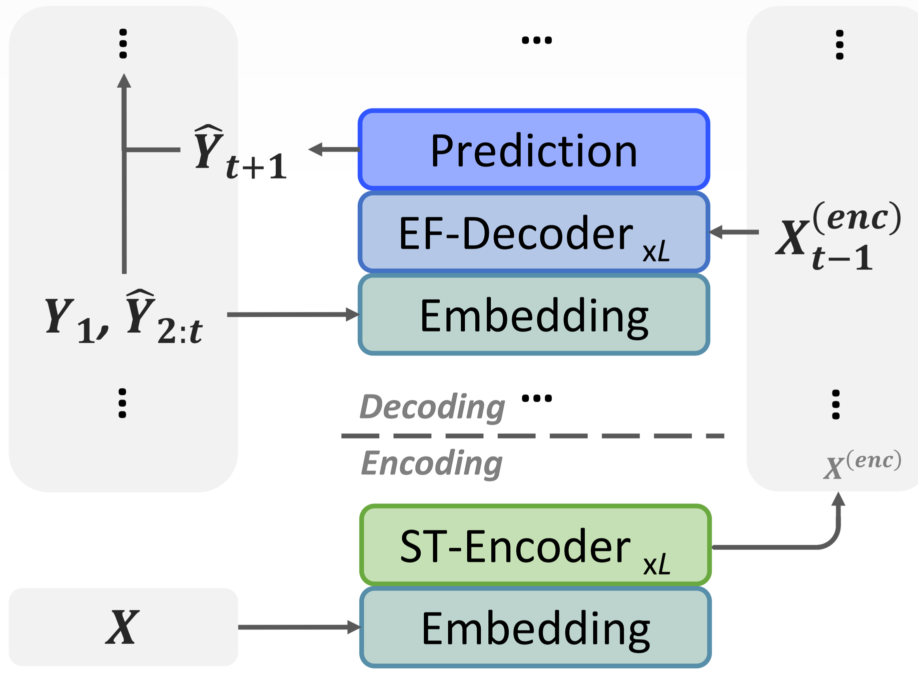
In scenarios where the behavior of participants are highly coupled and reactive, the most important clue for determining the behavior of a target participant in next frame would be the status of other observed participants in the current frame

Entry-flipping protects this clue by limiting the error level during autoregressive procedure

Trajectory Inference with manually injected noise

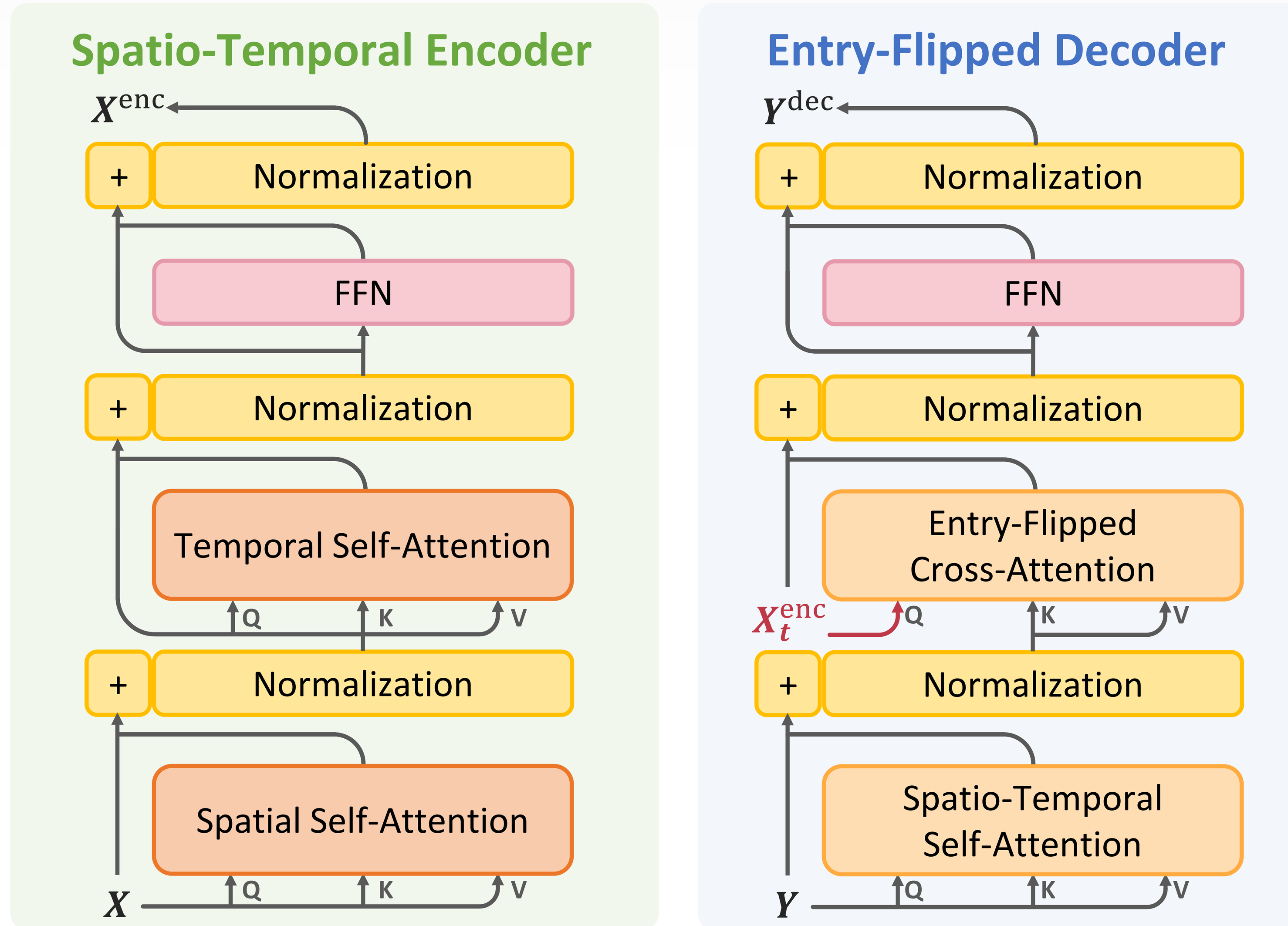
Noise Position	Methods	FAD Short	FAD Mid	FAD Long	FAD Avg
Transformer	No Noise	37.14	52.06	68.14	49.34
	Noise@t=3	75.99	103.24	141.06	99.67
	Noise@t=6	80.03	105.35	145.39	105.85
	Noise@t=9	131.76	161.07	205.26	157.81
EF-Transformer	No Noise	35.38	48.62	64.23	46.43
	Noise@t=3	37.23	56.37	84.65	54.15
	Noise@t=6	55.19	64.90	90.71	65.68
	Noise@t=9	115.93	123.30	145.31	124.29

Pipeline: Entry-Flipping Transformer



Feature Embedding, Prediction, and FFN layers consists of MLPs

Layer normalization is applied in both ST-Encoders and EF-Decoders



Quantitative Results

Ablation study of decoders

Decoder	MAD Short	MAD Mid	MAD Long	MAD Avg	FAD Short	FAD Mid	FAD Long	FAD Avg
Typical	20.14	33.09	50.70	31.33	35.85	52.55	71.57	49.67
All Query	20.80	33.05	46.86	30.92	38.25	55.50	70.14	51.71
Limited Query	**19.24**	**30.71**	**41.98**	**28.44**	**34.97**	**50.36**	**62.60**	**46.83**
Hybrid(TP→EF)	19.49	31.01	45.71	29.29	35.31	50.96	69.81	48.43
Hybrid(EF→TP)	19.52	31.53	43.84	29.24	35.53	52.62	70.57	49.43

Trajectory prediction on Tennis dataset

Methods	MAD Short	MAD Mid	MAD Long	MAD Avg	FAD Short	FAD Mid	FAD Long	FAD Avg
CNN-based	22.58	41.81	71.57	39.80	38.84	70.35	105.26	64.76
RNN-based	23.84	41.99	78.97	41.57	41.34	68.29	110.63	65.58
Transformer	20.14	33.09	50.70	31.33	35.85	52.55	71.57	49.67
SR-LSTM [41]	20.43	43.86	85.88	42.37	39.11	75.36	117.43	69.25
STAR [36]	23.83	43.80	83.65	43.20	37.83	70.61	117.19	66.50
EF-Transformer	**19.24**	**30.71**	**41.98**	**28.44**	**34.97**	**50.36**	**62.60**	**46.83**

Qualitative Results