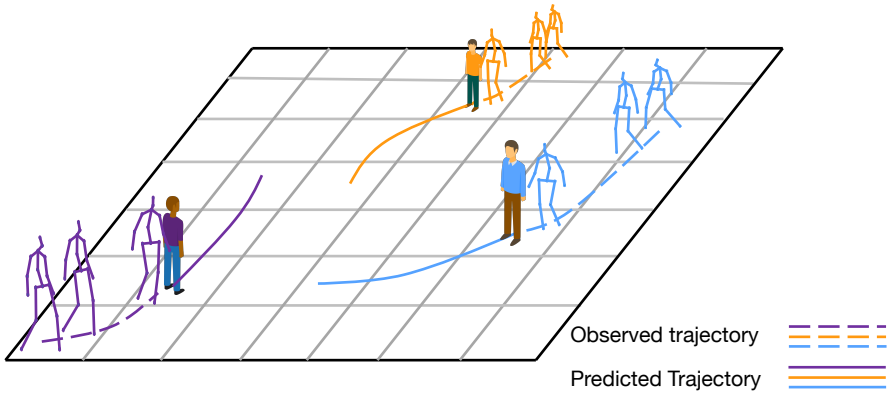
EPFL VITA

Social-Transmotion: Promptable Human Trajectory Prediction

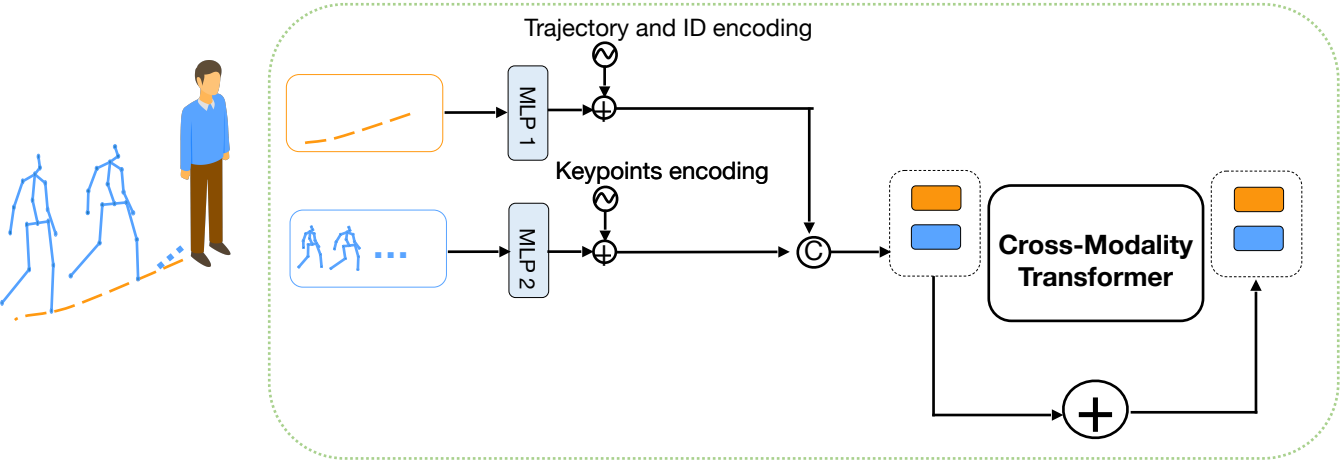
Saeed Saadatnejad*, Yang Gao*, Kaouther Messaoud, Alexandre Alahi

Problem

- Human trajectory prediction task
- Relying on trajectory input and ignoring other signals conveying mobility patterns
- How to integrate various types and quantities of visual cues in a general model, without excessive reliance on a single source of information?



Approach



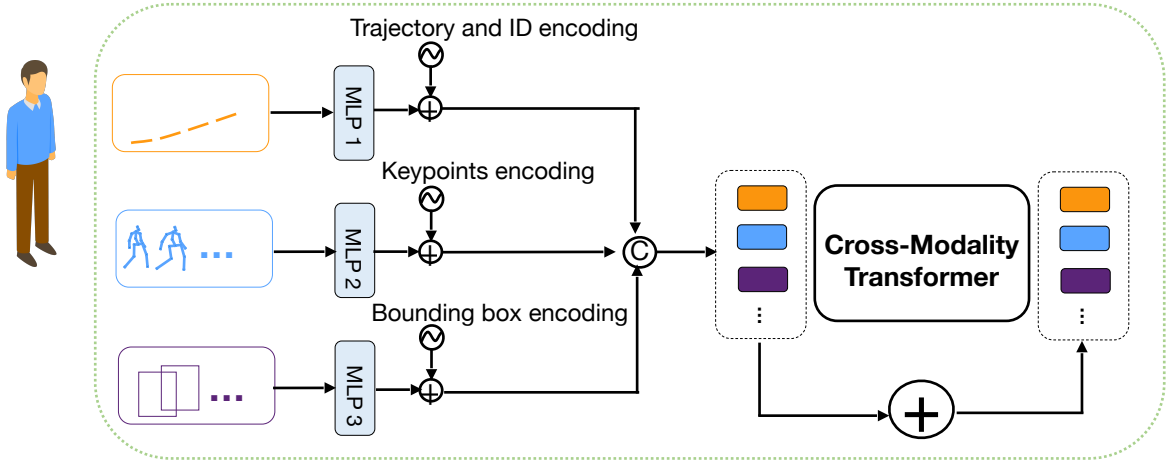
Agent past trajectory

Additional Visual Cues:
3D keypoints

Inputs Embedding +
Positional encoding

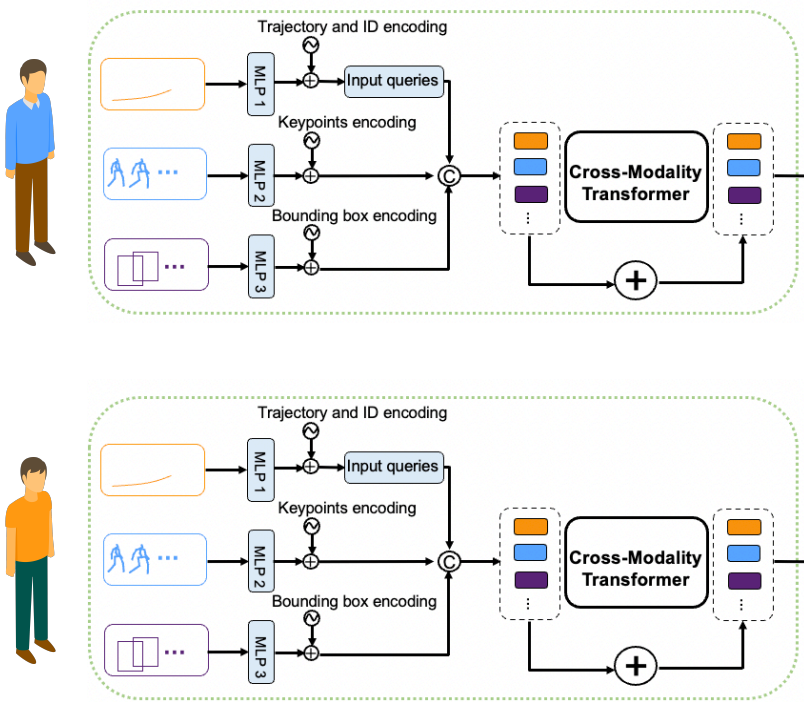
Cross modality motion encoding

Approach

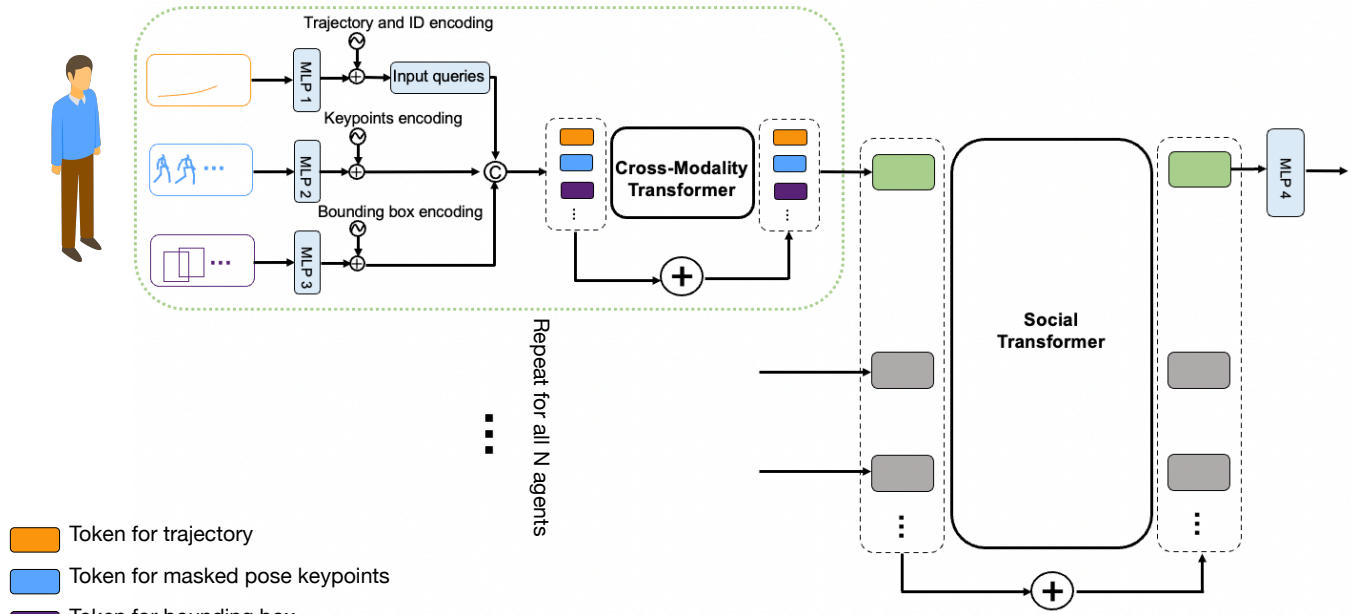


Flexible model

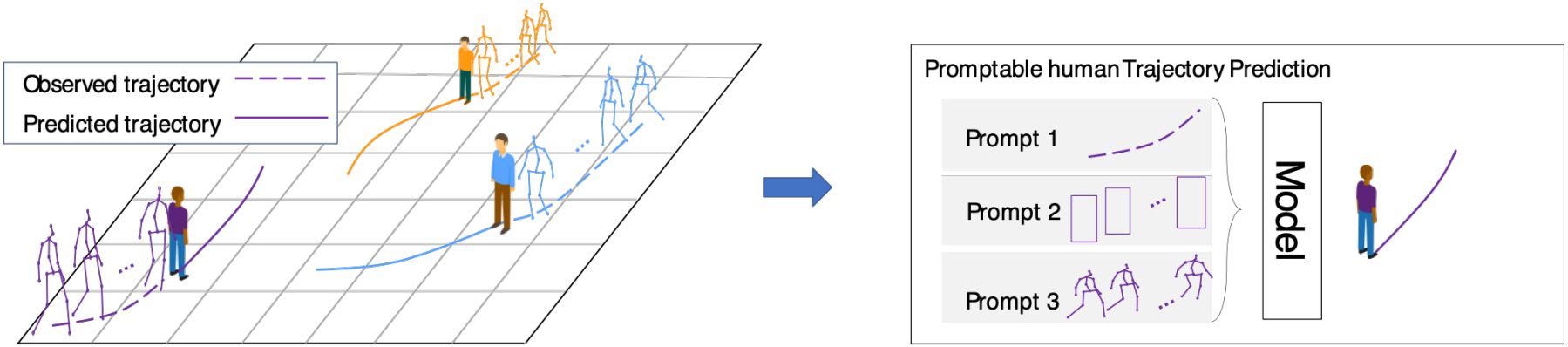
Approach


Trajectory and ID encoding
MLP 1
Input queries
Keypoints encoding
MLP 2
Bounding box encoding
MLP 3
Cross-Modality Transformer
Trajectory and ID encoding
MLP 1
Input queries
Keypoints encoding
MLP 2
Bounding box encoding
MLP 3
Cross-Modality Transformer

Approach

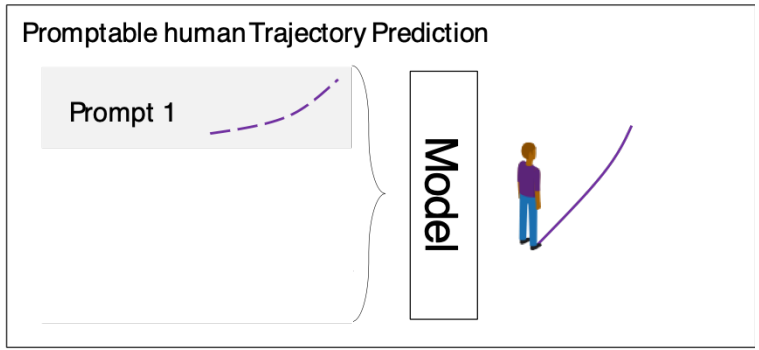

Trajectory and ID encoding
MLP 1
Input queries
Keypoints encoding
MLP 2
Bounding box encoding
MLP 3
Cross-Modality Transformer
Repeat for all N agents
Social Transformer
MLP 4
Token for trajectory
Token for masked pose keypoints
Token for bounding box
Token for primary agent
Token for neighbors

Approach

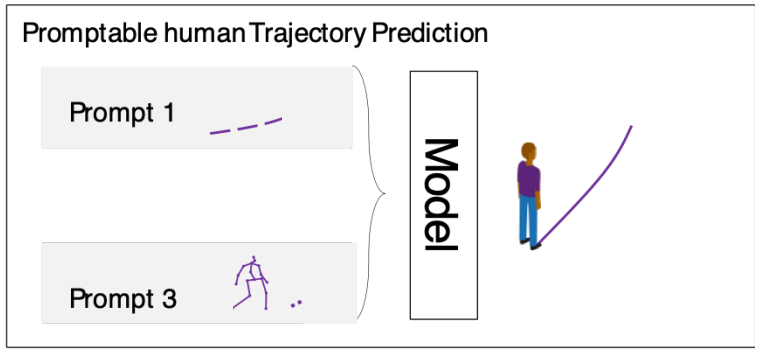


- We designed a promptable human trajectory predictor given available prompts such as past trajectories or body poses of agents
- How to make the model adaptable to any available visual cues?

Masking

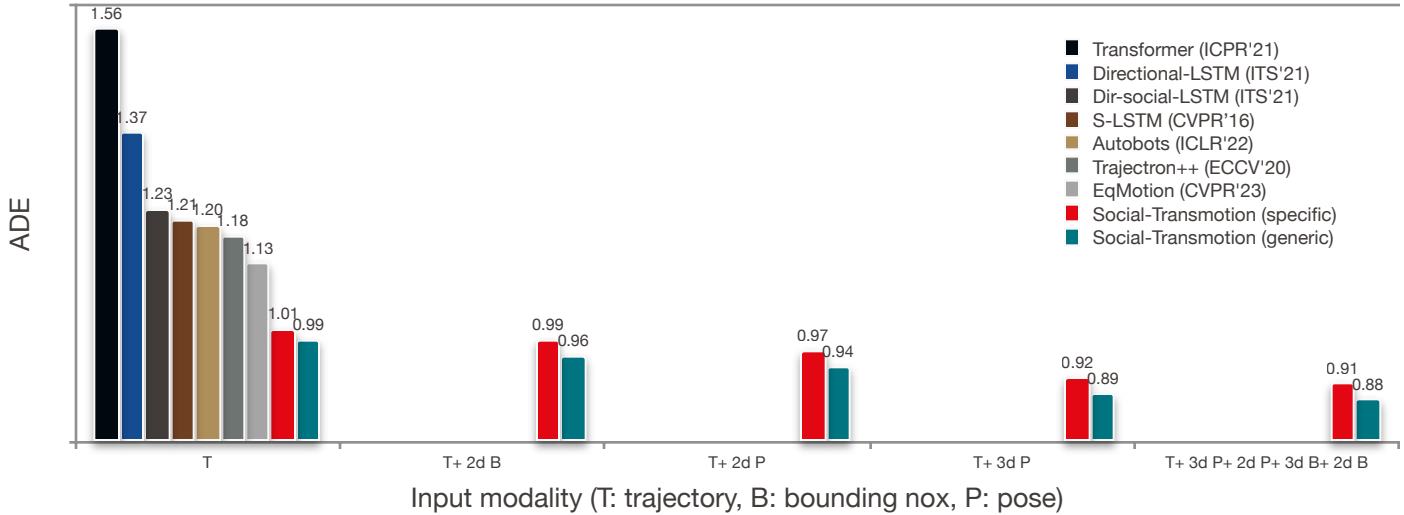


Masking


Promptable human Trajectory Prediction
Prompt 1
Prompt 3
Model

Results

- Generic model performs better than training specific models for each modality
- All visual cues help. 3D pose the most.



Dataset: JTA

Results


Social Transmotion + 3D Pose
Social Transmotion
Ground truth
Observation

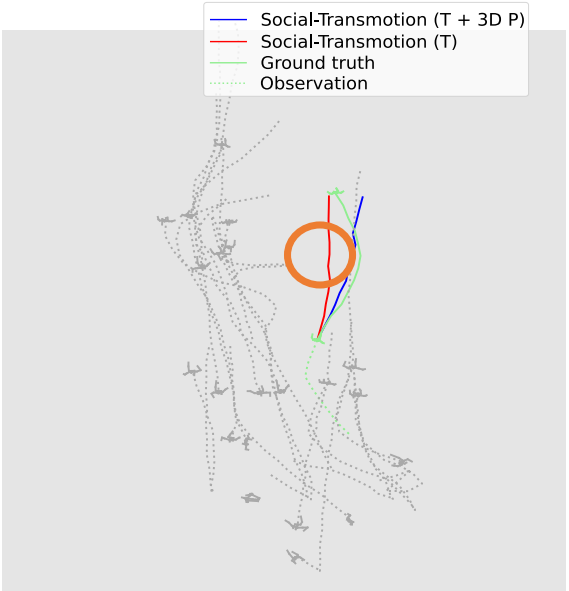
Discussions

- Robustness against imperfect input data

Input Modality at Inference	ADE / FDE
T + 3D P	0.89 / 1.81
90% T + 90% 3D P	0.89 / 1.81
50% T + 50% 3D P	1.01 / 2.00
50% T + 10% 3D P	1.10 / 2.16
T + 3D P w/ Gaussian Noise (std=25)	0.98 / 1.94
T + 3D P w/ Gaussian Noise (Std=50)	1.05 / 2.05
T + Random Leg and Arm Occlusion	0.90 / 1.83
T + Structured Right Leg Occlusion	0.90 . 1.82
T + 50% Complete Frame Missing	0.93 / 1.89
T + 90% Complete Frame Missing	0.99 / 1.98

Discussions

- A failure case due to missing context



Summary

- **Visual cues**, such as human body pose, are helpful **augmentation** for enhanced trajectory prediction
- Social-Transmotion is a **generic** Transformer-based model able to leverage **multiple** visual cues in **varying** quantities powered by the **masking** strategy

github.com/vita-epfl/social-transmotion