Real-Time Motion Prediction via Heterogeneous Polyline Transformer with Relative Pose Encoding

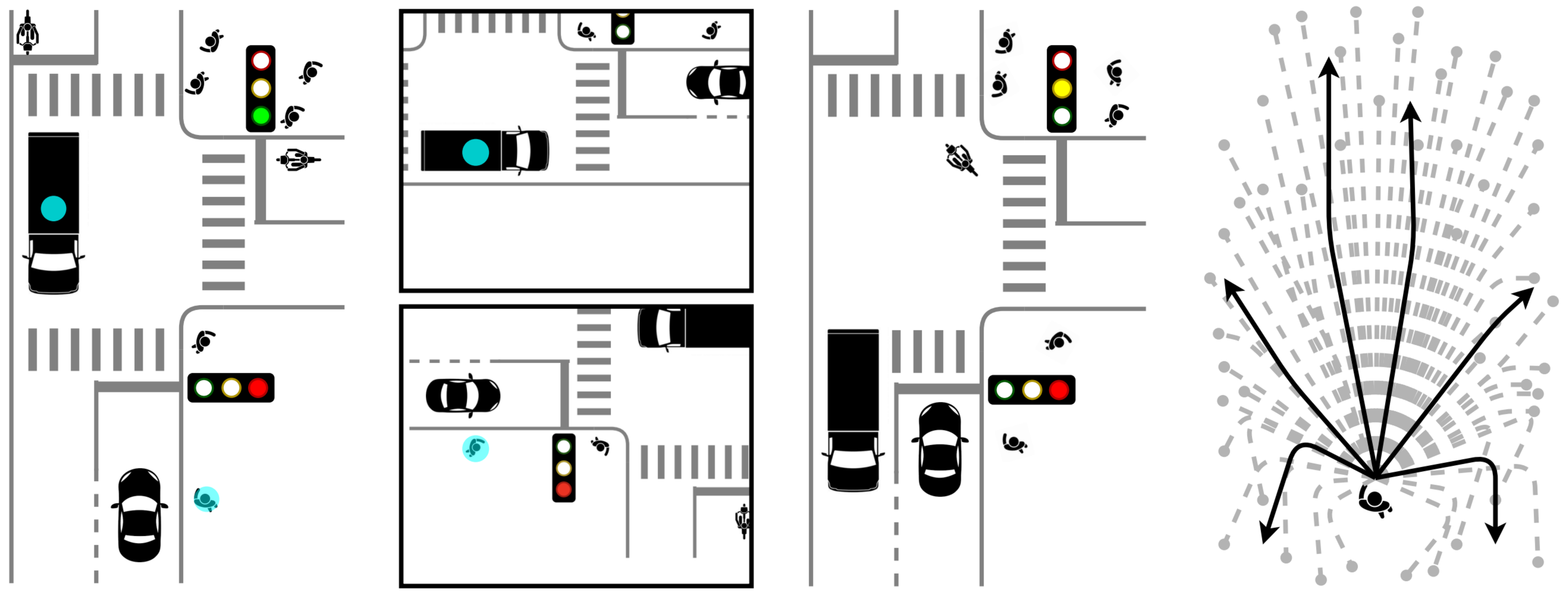
Zhejun Zhang[1], Alexander Liniger[1], Christos Sakaridis[1], Fisher Yu[1], and Luc Van Gool[1,2,3]

[1]Computer Vision Lab, ETH Zurich, CH. [2]PSI, KU Leuven, BE. [3]INSAIT, Un. Sofia, BU.

NEURAL INFORMATION PROCESSING SYSTEMS

ETH zürich CVL Computer Vision Lab

Motivation



(a) Dense traffic scenario (b) Agent-centric ROIs (c) Online inference (d) Trajectory aggregation

a) **Real-time and on-board** motion prediction in dense urban scenario.
b) Agent-centric: Good performance. Bad scalability.
c) **Online inference** with streaming inputs.
d) Expensive post-processing and ensembling.

KNARPE

K-nearest **N**eighbor **A**ttention with **R**elative **P**ose **E**ncoding.

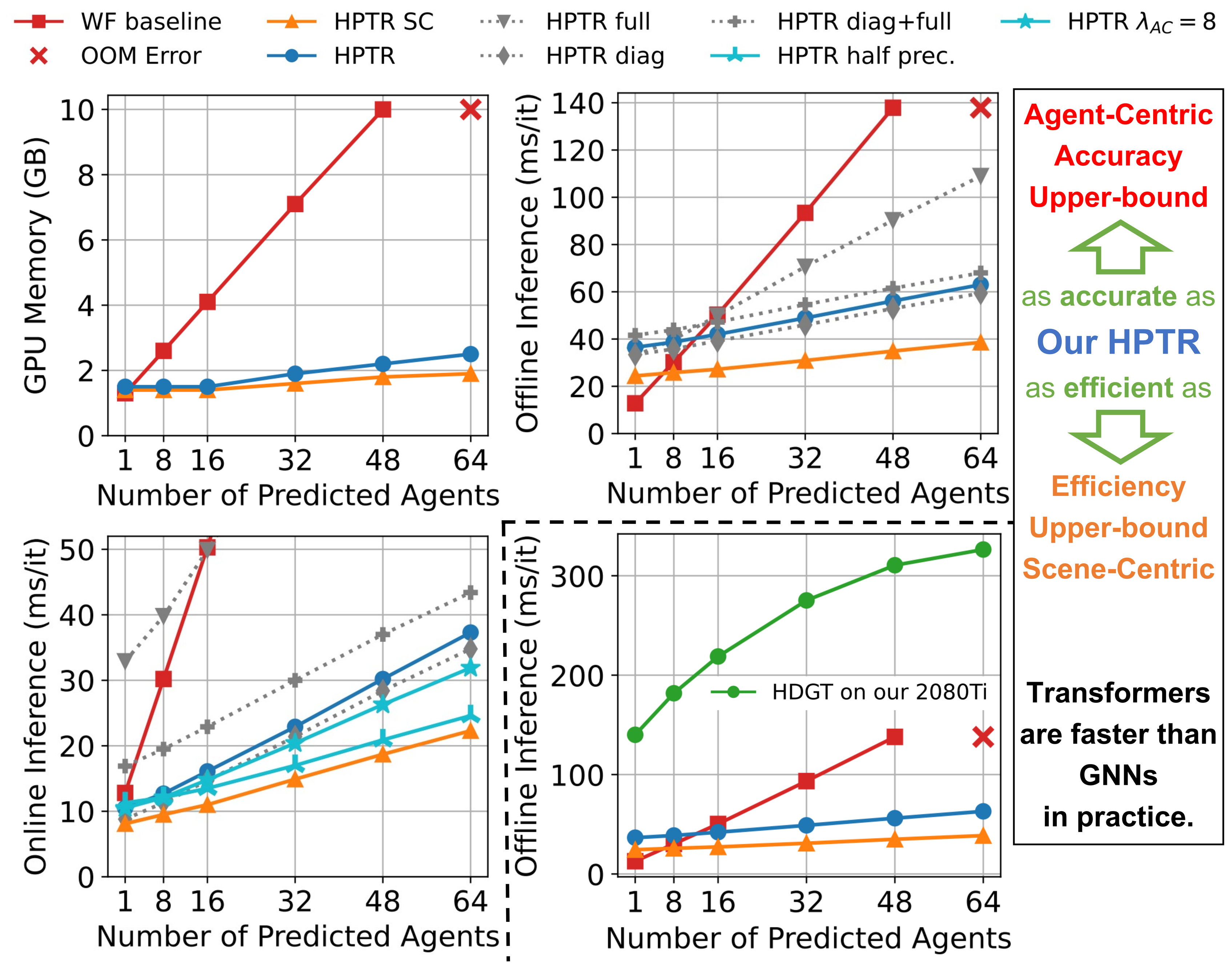
$$\mathbf{z}_i = \boxed{\textsc{Knarpe}}(\mathbf{u}_i, \mathbf{u}_j, \mathbf{r}_{ij} \mid j \in \kappa_i^K) = \sum_{\boxed{j \in \kappa_i^K}} \alpha_{ij} \left(\mathbf{u}_j \mathbf{W}^v + \mathbf{b}^v + \boxed{\mathrm{RPE}(\mathbf{r}_{ij}) \hat{\mathbf{W}}^v + \hat{\mathbf{b}}^v} \right)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \kappa_i^K} \exp(e_{ik})}, \quad e_{ij} = \frac{(\mathbf{u}_i \mathbf{W}^q + \mathbf{b}^q)(\mathbf{u}_j \mathbf{W}^k + \mathbf{b}^k + \boxed{\mathrm{RPE}(\mathbf{r}_{ij}) \hat{\mathbf{W}}^k + \hat{\mathbf{b}}^k})}{\sqrt{D}},$$

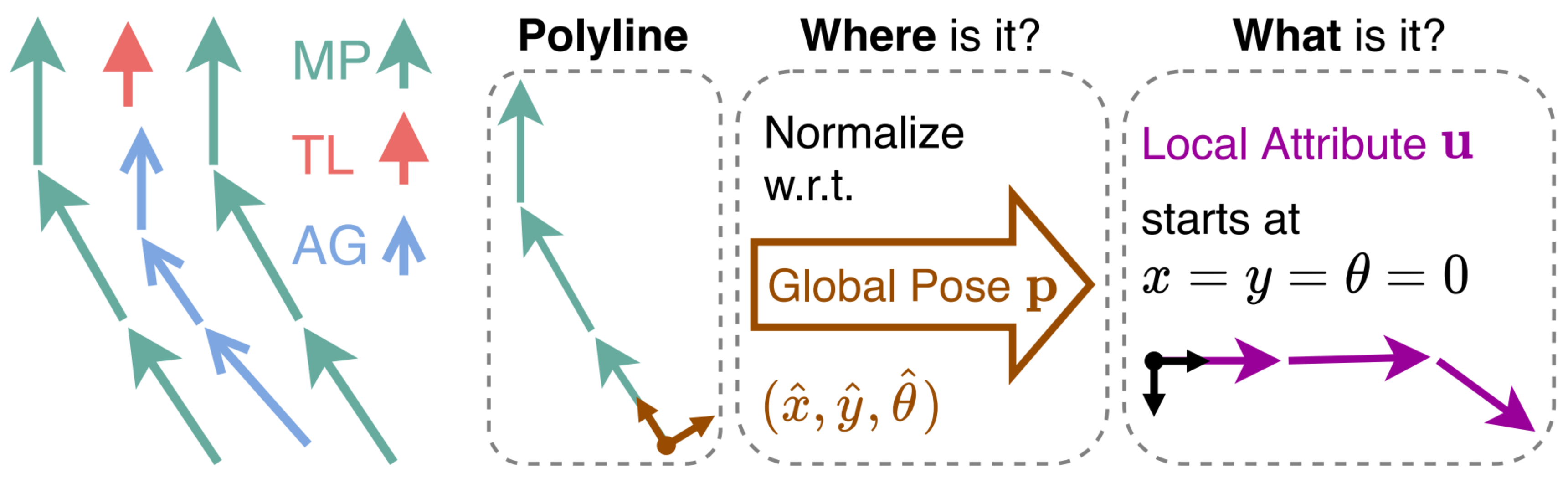
$$\mathrm{RPE}(\mathbf{r}_{ij}) = \mathrm{concat}(\mathrm{PE}(x_{ij}), \mathrm{PE}(y_{ij}), \mathrm{AE}(\theta_{ij})),$$

- Based on multi-head dot-product attention.
- Implemented with **basic matrix operations**: matrix indexing, summation and element-wise multiplication.
- Self-Attention: Local context aggregation like CNN.
- Cross-Attention: Rotated ROI alignment with CNN.

Performance



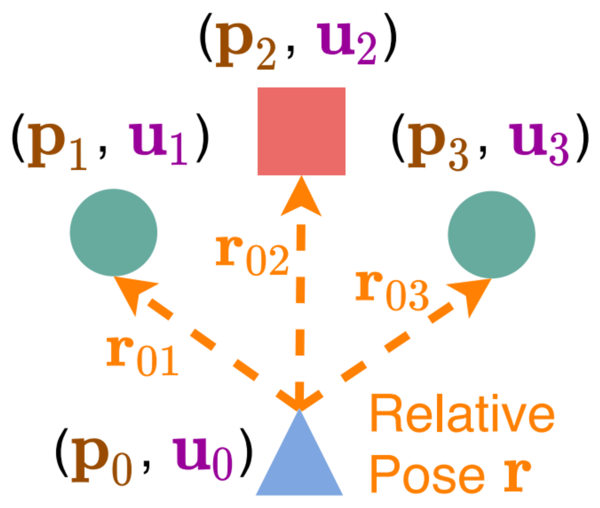
Pairwise-Relative Representation



(a) Polylines. (b) Local attribute represented in the global pose's frame.

Global Pose p: Where is the polyline?

Local Attribute u: What kind of polyline is it?



(c) Obtain relative poses.

Input to the Network:

3-dimensional relative pose $\mathbf{r}$.

High-dimensional local attribute $\mathbf{u}$.

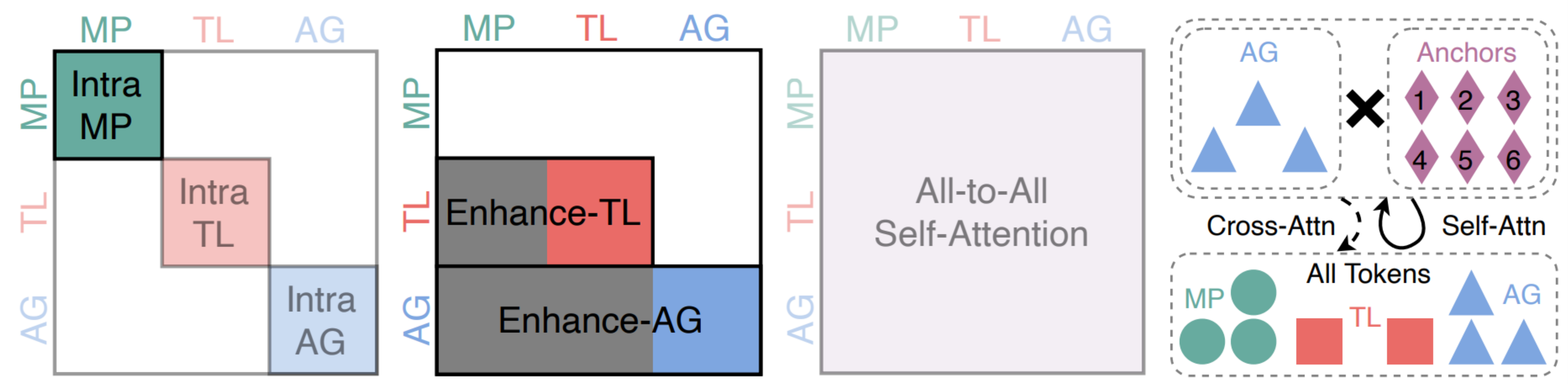
Rotation and translation invariance.

Good scalability by sharing $\mathbf{u}$.

So far only exploited by GNNs.

HPTR

Heterogeneous **P**olyline **T**ransformer with **R**elative pose encoding



(a) Intra-class TF encoder (b) Inter-class TF decoder (c) All-to-all TF encoder (d) AC-to-all TF decoder

- Based on KNARPE.
- Inspired by Wayformer.
- Remove redundant attentions.
- Transformers are organized in a hierarchical way.
- Intermediate results can be cached and reused.
- **Asynchronous token update** during online inference.

Summary

- **KNARPE** allows the pairwise-relative representation to be used by Transformers.
- **HPTR** uses hierarchical architecture to enable asynchronous token update.
- **SoTA** performance among E2E methods: WOMD and AV2 dataset (c.f. paper).
- **Good Performance**: As accurate as agent-centric methods.
- **Good Scalability**: As efficient as scene-centric methods.
- **Real-Time and On-Board** Motion Prediction: **40 FPS** during online inference. 80% reduction on online inference latency and GPU memory.
- **Code**: https://github.com/zhejz/HPTR