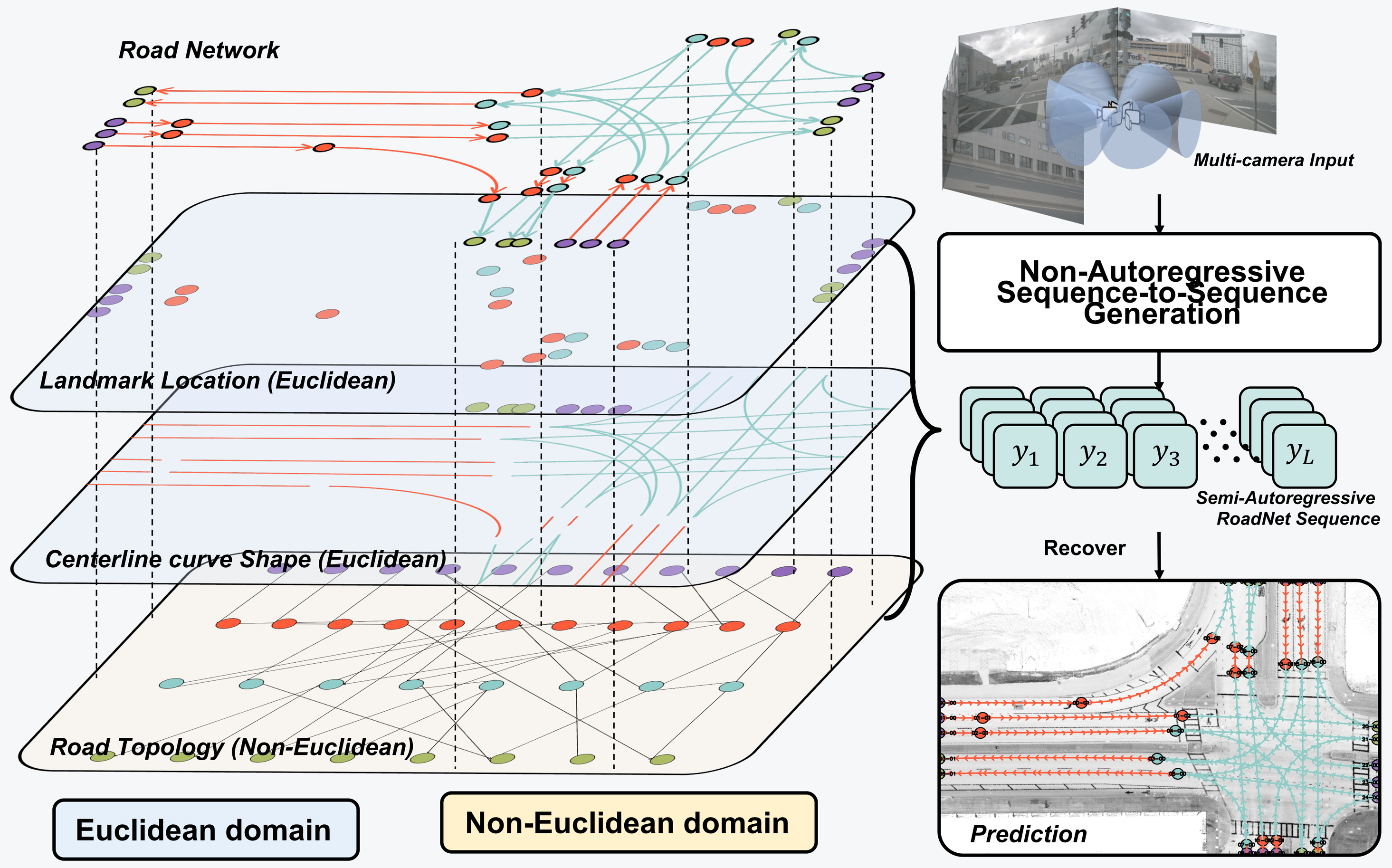
Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach

Jiachen Lu[1], Renyuan Peng[1], Xinyue Cai[3], Hang Xu[3], Hongyang Li[2], Feng Wen[3], Wei Zhang[3], Li Zhang[1]

[1]Fudan University [2]Shanghai AI Lab [3]Huawei Noah's Ark Lab

Euclidean and Non-Euclidean Data for Road Network

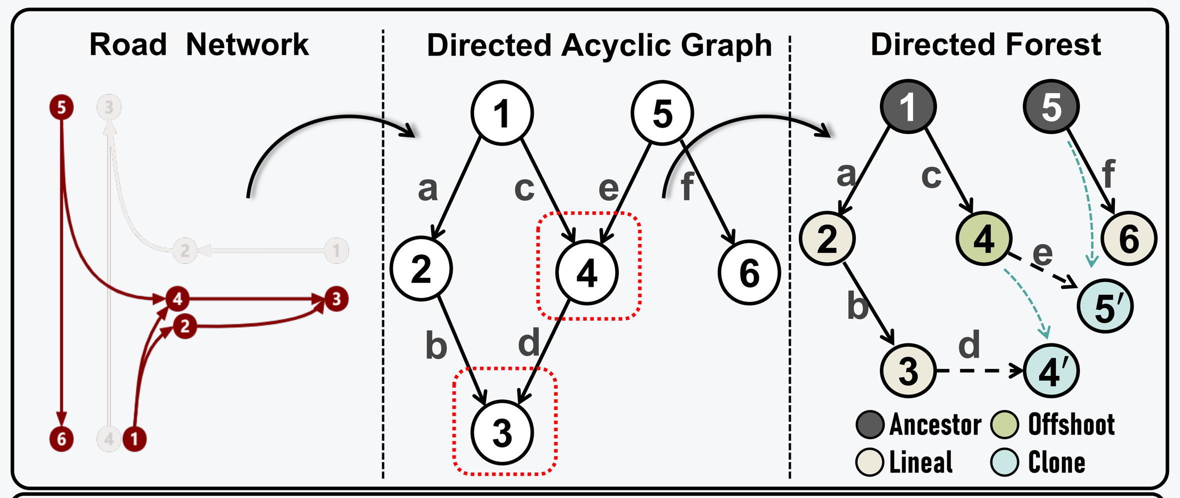


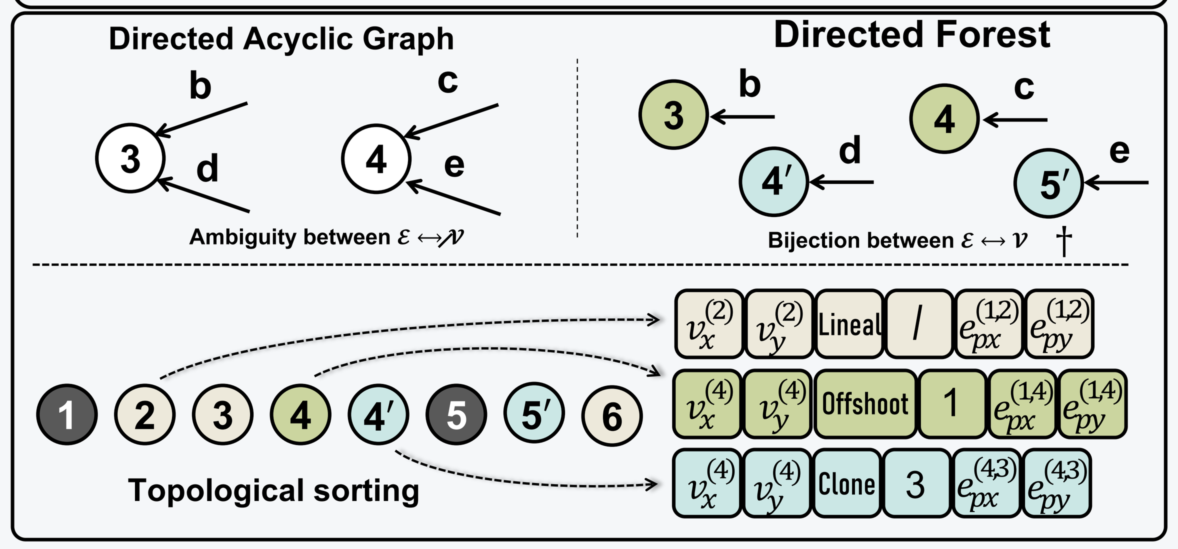
High-definition Road Network Topology contains:

1. **Euclidean data**: locations of landmarks and shapes of curves.
2. **Non-Euclidean data**: road topology.

RoadNet Sequence

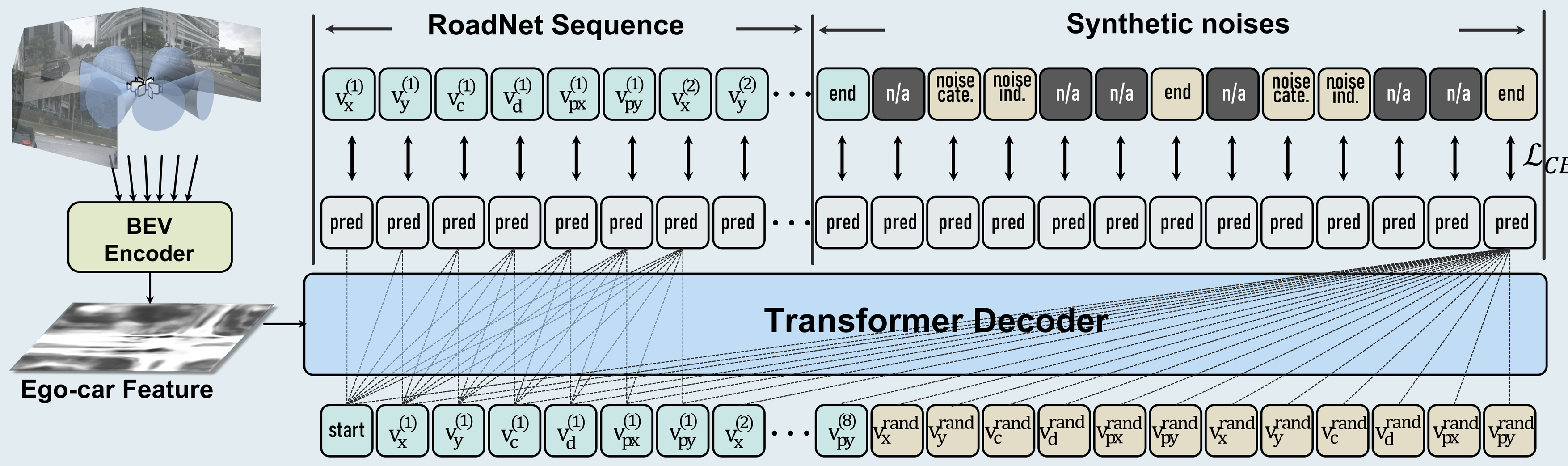
From DAG to Directed Forest → **Topological Sorting** → **Sequence Construction**





We introduce a Euclidean-nonEuclidean unified representation **RoadNet Sequence** with merits of **losslessness, efficiency and interaction**.

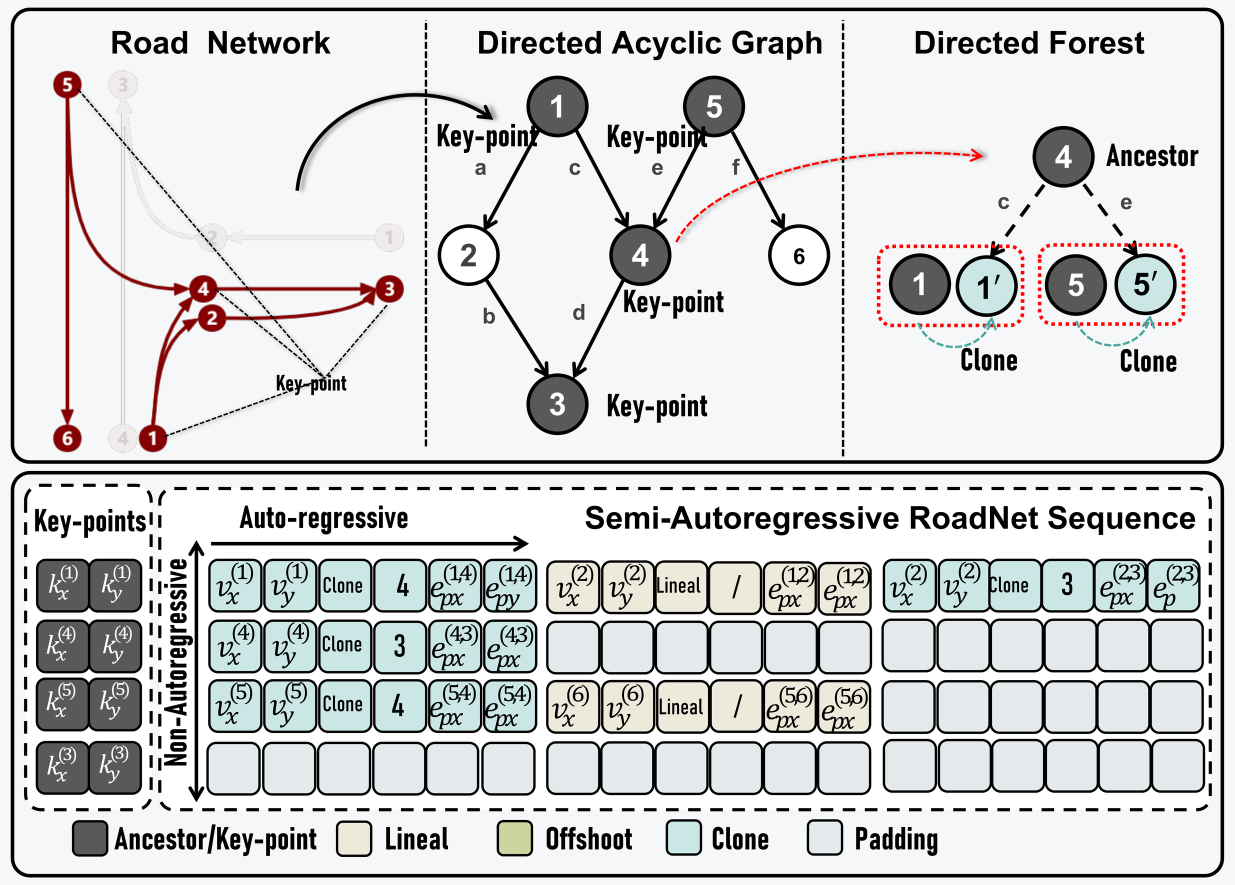
1. **Losslessness:** ensured by establishing a **bijection** from road network to RoadNet Sequence.
2. **Efficiency:** achieved by limiting RoadNet Sequence length to the shortest $\mathcal{O}(E)$ through a specially designed topological sorting rule.
3. **Interaction:** reveals the interdependence between Euclidean and non-Euclidean information within a single sequence.



Auto-Regressive RoadNetTransformer: We apply the encoder-decoder architecture.

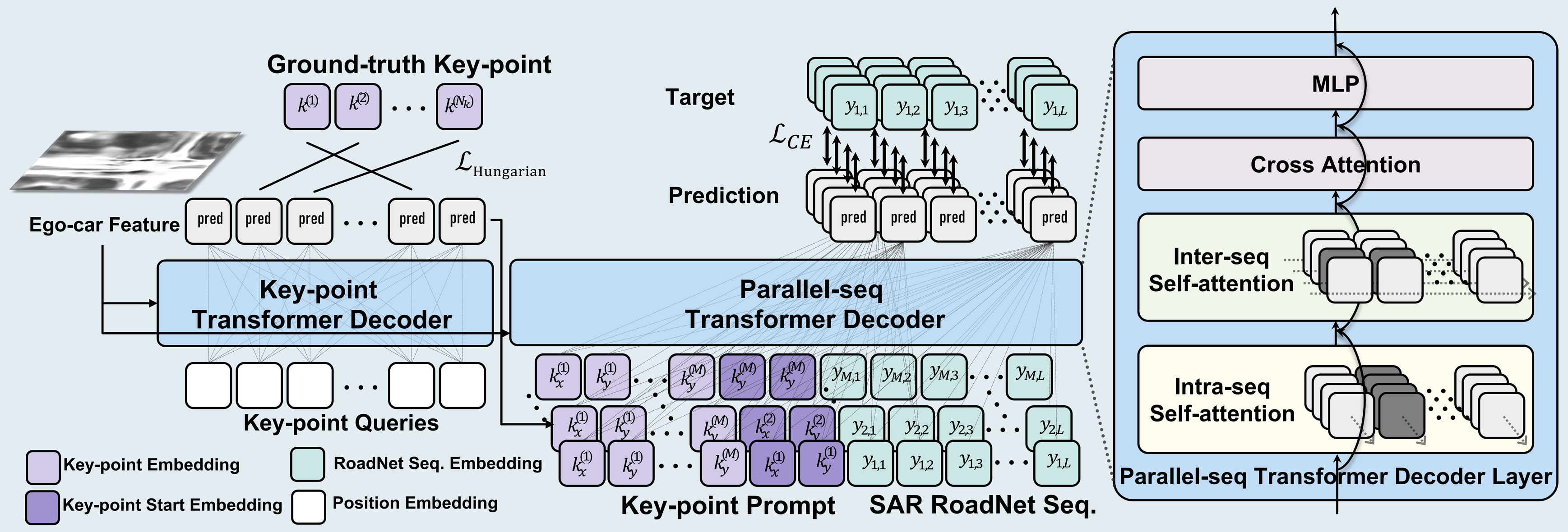
1. **Encoder** is responsible for extracting BEV feature from multiple onboard cameras such as Lift-Splat-Shoot.
2. **Decoder** includes a self-attention layer, a cross-attention layer and a MLP layer.

Semi-Autoregressive RoadNet Sequence



To parallelize the RoadNet Sequence, we have the following observations:

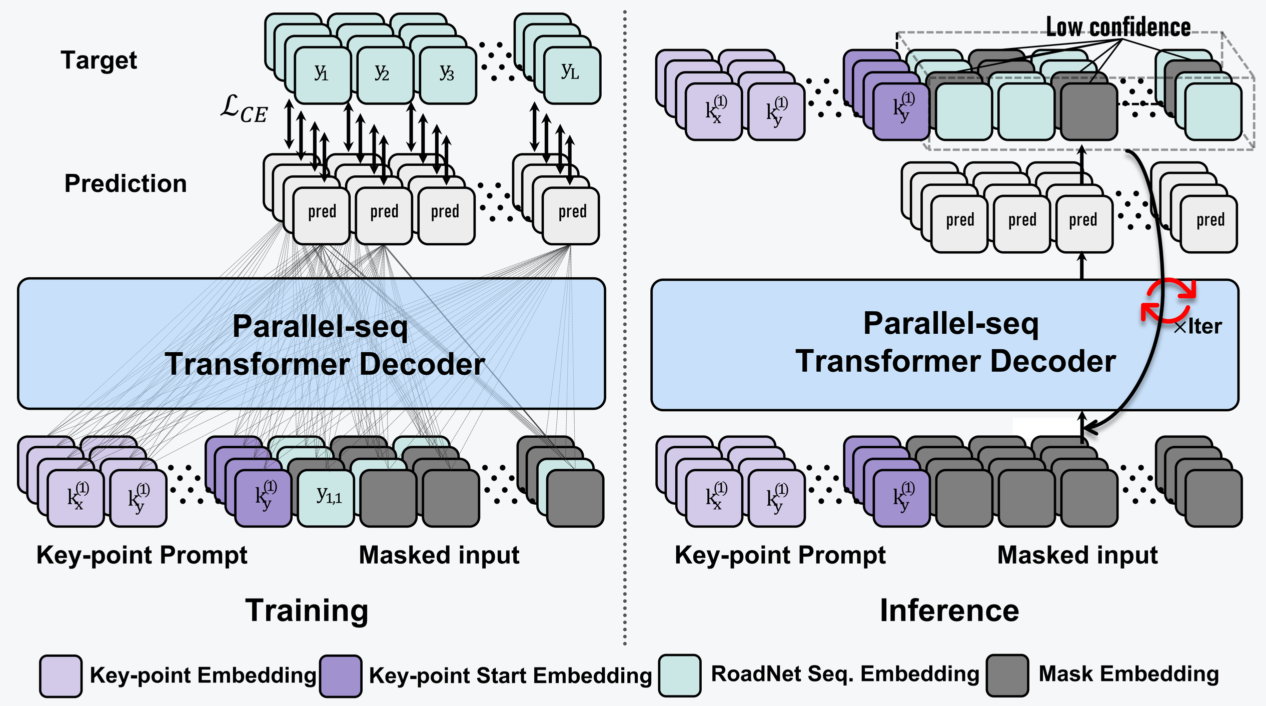
1. The locations of certain road points (start, fork or merge points) can be **independent** of previous vertices and instead depend solely on the BEV feature
2. Except for locations of these road points, other tokens are still **auto-regressive.**



Semi-autoregressive RoadNetTransformer can be divided into three parts: (i) Ego-car Feature Encoder, (ii) Key-point Transformer Decoder, (iii) Parallel-Seq Transformer Decoder.

1. **Key-point Transformer Decoder** is a parallel Transformer decoder, which predict locations of key points based on set prediction.
2. **Parallel-Seq Transformer Decoder** is proposed for solving mixture of auto-regressive and non-autoregressive problem.

Non-Autoregressive RoadNetTransformer

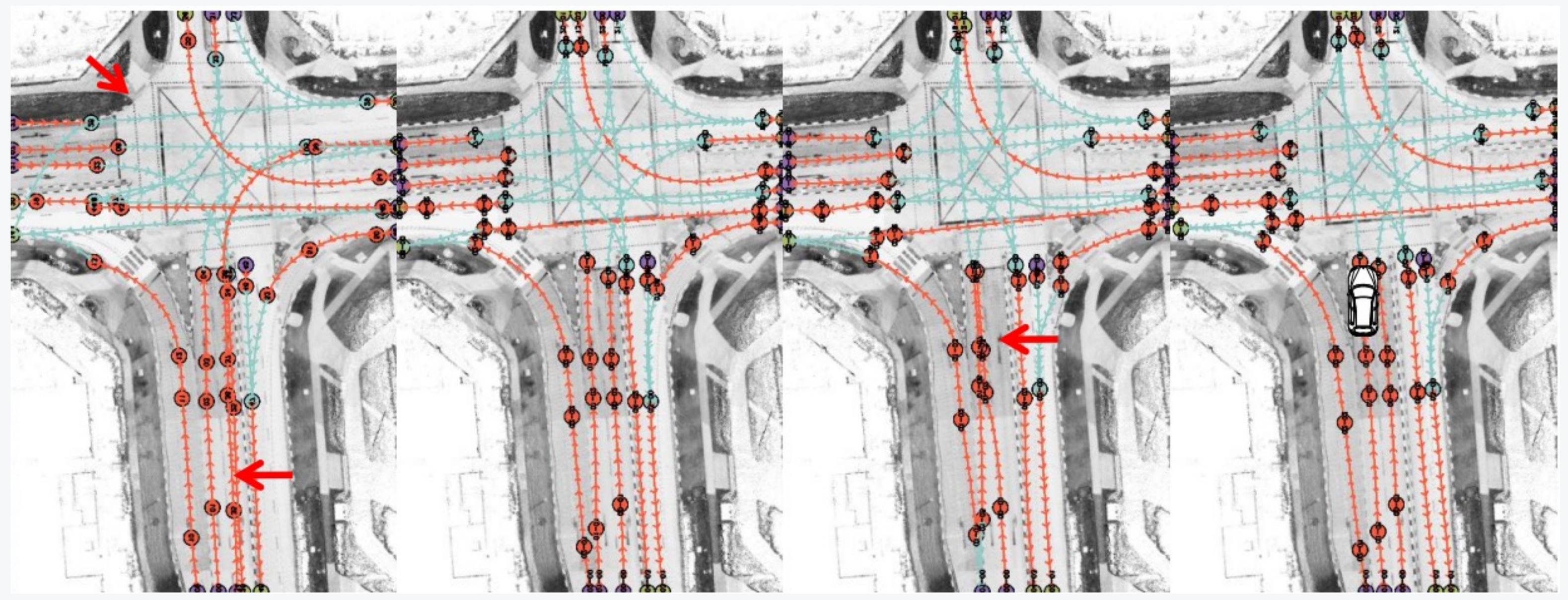


We propose a fully non-autoregressive generation model by utilizing a masked language modeling strategy that involves masking a high percentage of the input ground-truth sequence. During inference, with each iteration, the results will be gradually refined.

Results on the nuScenes validation set

Methods	Landmark			Reachability			FPS
	L-P	L-R	L-F	R-P	R-R	R-F	
NAR-RNTR (ResNet)	57.1	42.7	48.9	63.7	45.2	52.8	5.5
AR-RNTR (VovNet)	62.6	47.9	54.3	73.2	52.9	61.4	0.1 (1.0×)
SAR-RNTR (VovNet)	**66.0**	**55.9**	**60.5**	**74.5**	**61.1**	**67.1**	0.6 (6.0×)
NAR-RNTR (VovNet)	65.6	55.7	60.2	73.4	60.0	66.0	4.7 (47×)

Qualitative results on nuScenes dataset



Auto-Regressive Semi-Autoregressive Non-Autoregressive GT