End-to-End Vectorized Map Construction and Planning

陈少宇

2024.06.10

地平线
Horizon Robotics

Outline

地平线
Horizon Robotics

MapTR v1

Paper: https://arxiv.org/abs/2208.14437

Project Page: https://github.com/hustvl/MapTR

MapTR v2

Paper: https://arxiv.org/abs/2308.05736

Project Page: https://github.com/hustvl/MapTR

LaneGAP

Paper: https://arxiv.org/abs/2303.08815

Project Page: https://github.com/hustvl/LaneGAP

VAD v1

Paper: https://arxiv.org/abs/2303.12077

Project Page: https://github.com/hustvl/VAD

VAD v2

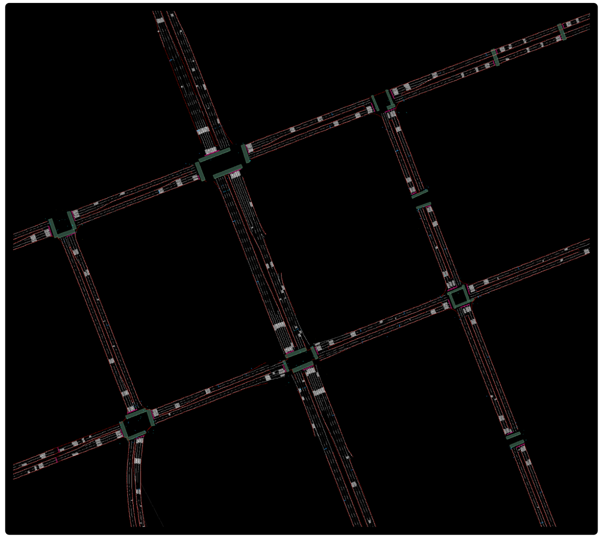
Paper: https://arxiv.org/abs/2402.13243

Project Page: https://hgao-cv.github.io/VADv2

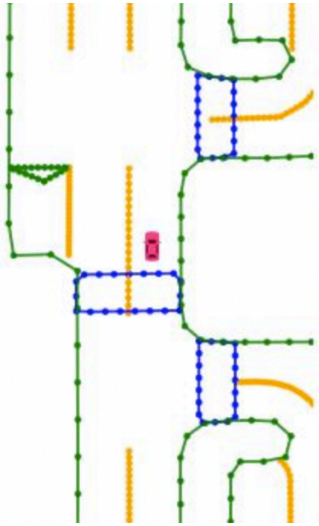
VMA

Paper: https://arxiv.org/abs/2304.09807

Project Page: https://github.com/hustvl/MapTR



Mapless



HD map

Navigation map + Online map construction

Limitations

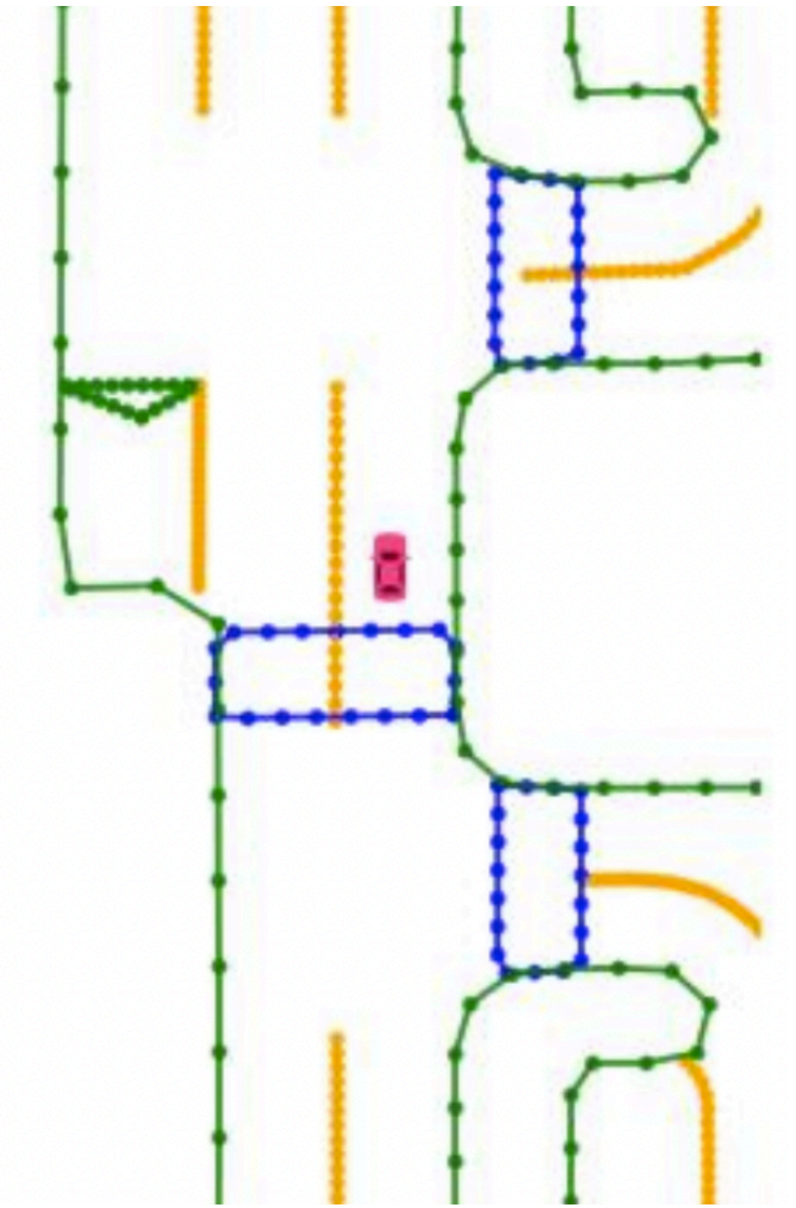
- High cost and complicated pipeline
- Scalability issue
- Freshness
- Limitation of law and regulation

Online Map Construction

Constructing map around ego-vehicle at runtime with onboard sensors

Methods

- BEV Segmentation + post-processing
  - Heavy engineering work
  - Corner case
- Lane detection
  - Anchor + regression
  - Rely on geometric prior
- Auto-regressive
  - Accumulated error
  - Efficiency



MapTRv1

GT Prediction Surrounding Views

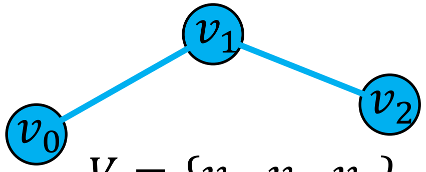
- End-to-end (no rule-based post-processing)
- Real-time
- Generalization ability (geometric shape and scenario)
- Fully data-driven and easy to scale up

Map Element Modeling

Open-shape map element

Discretization

Polyline

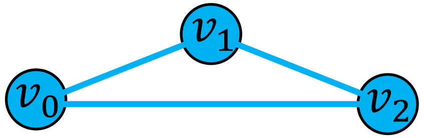


$V = \{v_0, v_1, v_2\}$

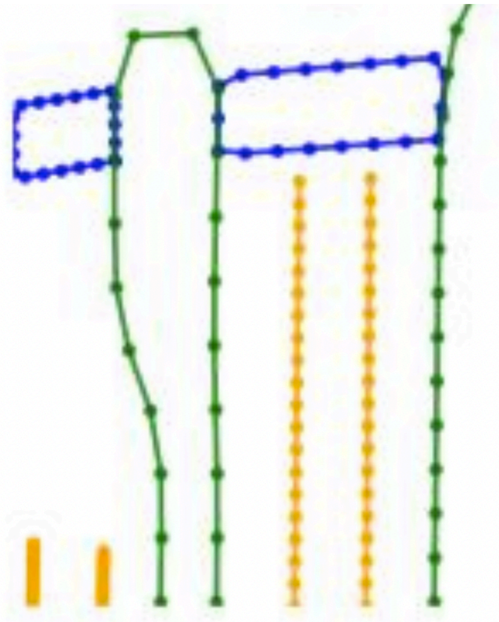
Closed-shape map element

Discretization

Polygon



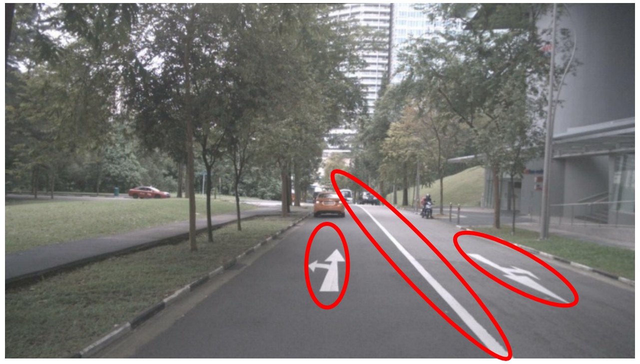
$V = \{v_0, v_1, v_2\}$

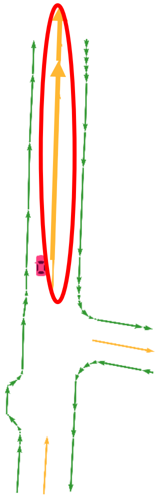


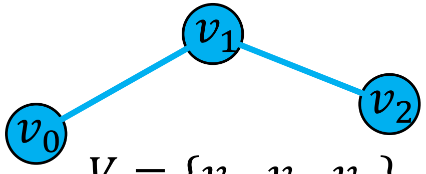
- No introducing geometric prior
- Well represent all kinds of geometric shapes

Equivalent Permutations

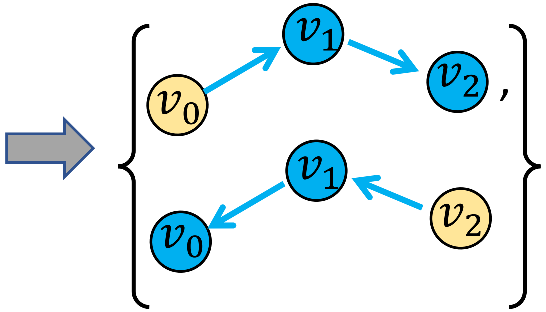
Open-shape map element



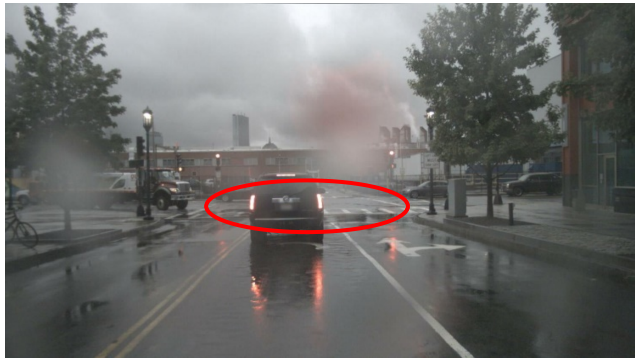


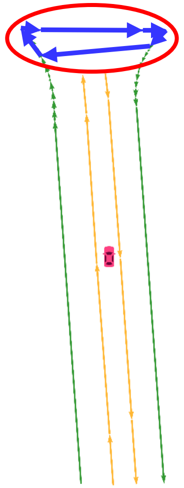


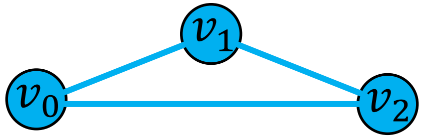
$V = \{v_0, v_1, v_2\}$



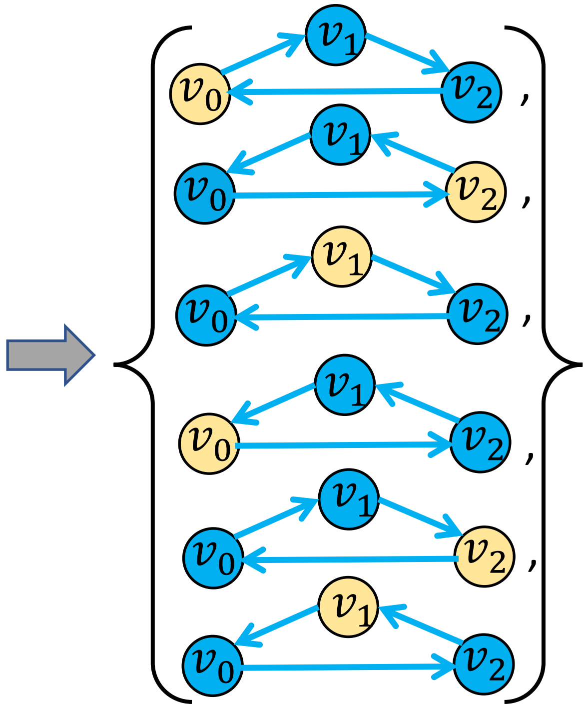
Closed-shape map element



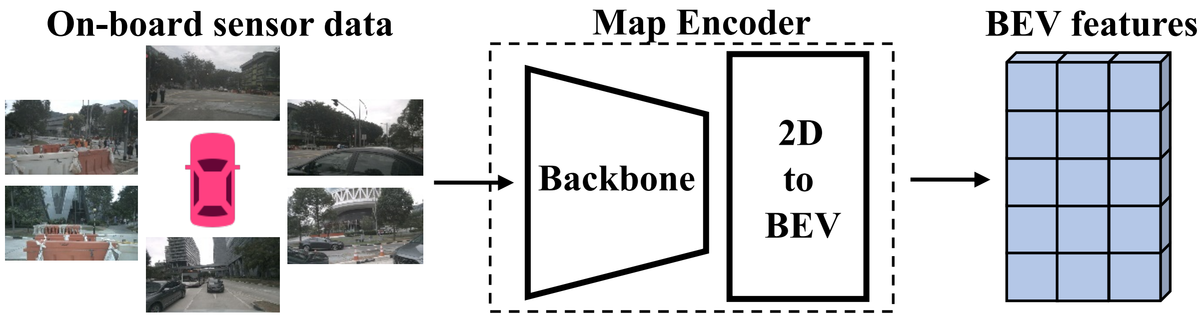




$V = \{v_0, v_1, v_2\}$



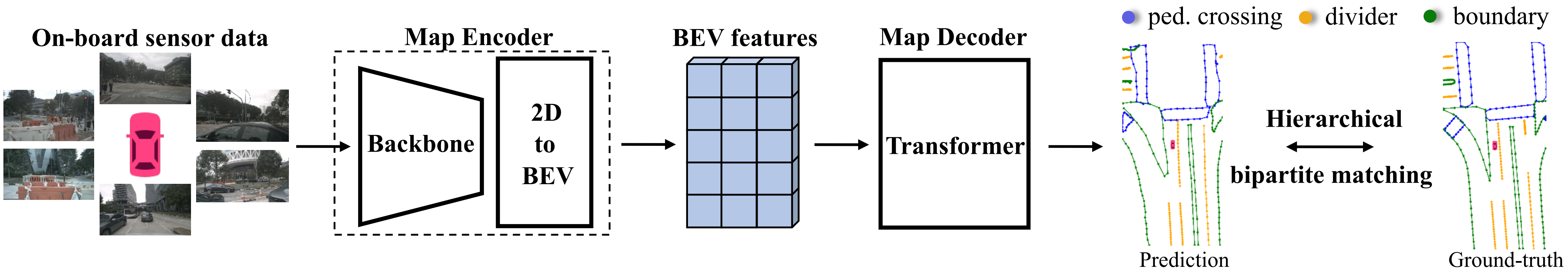
Framework



Map Encoder

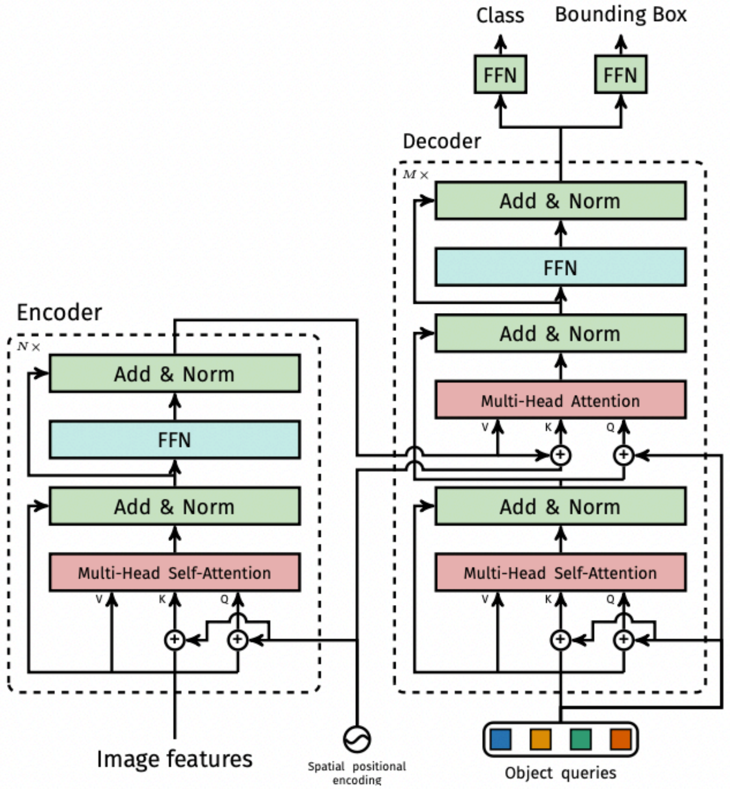
- Extracting features from original sensor data
- Transforming sensor features into a unified BEV representation

Framework

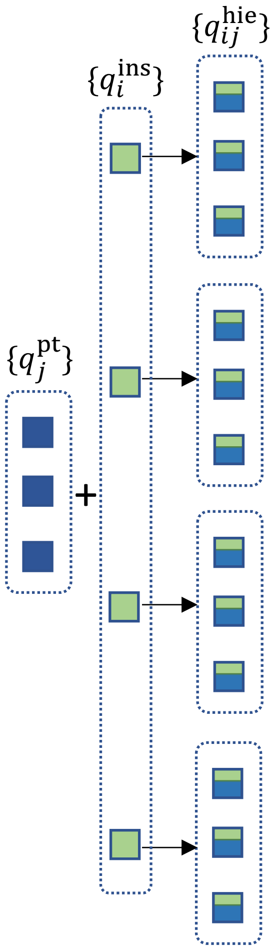


Map Decoder (DETR-like)

- Hierarchical query embeddings
- Parallel interaction and decoding
- Hierarchical matching



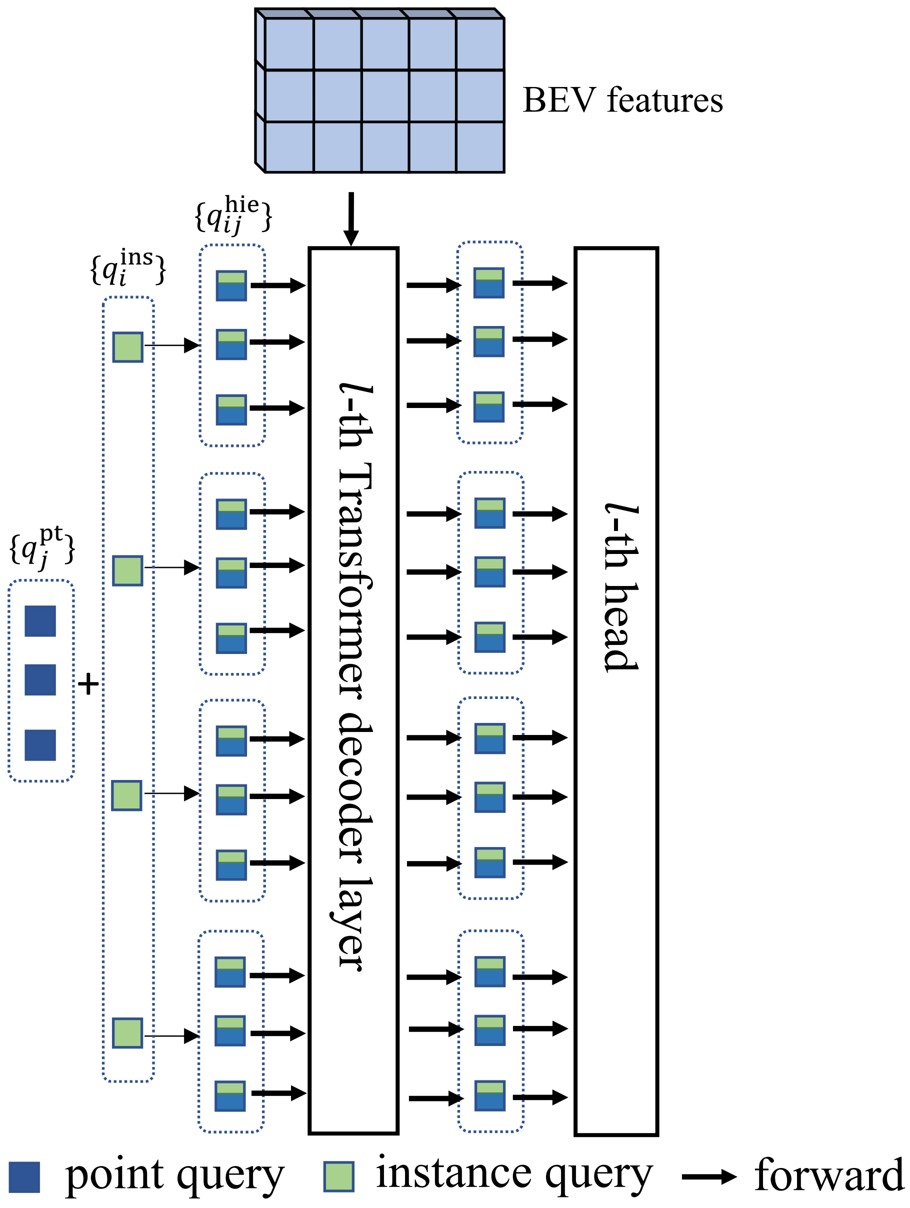
Hierarchical Query Embedding Scheme



Flexibly encoding each map element in a structured manner

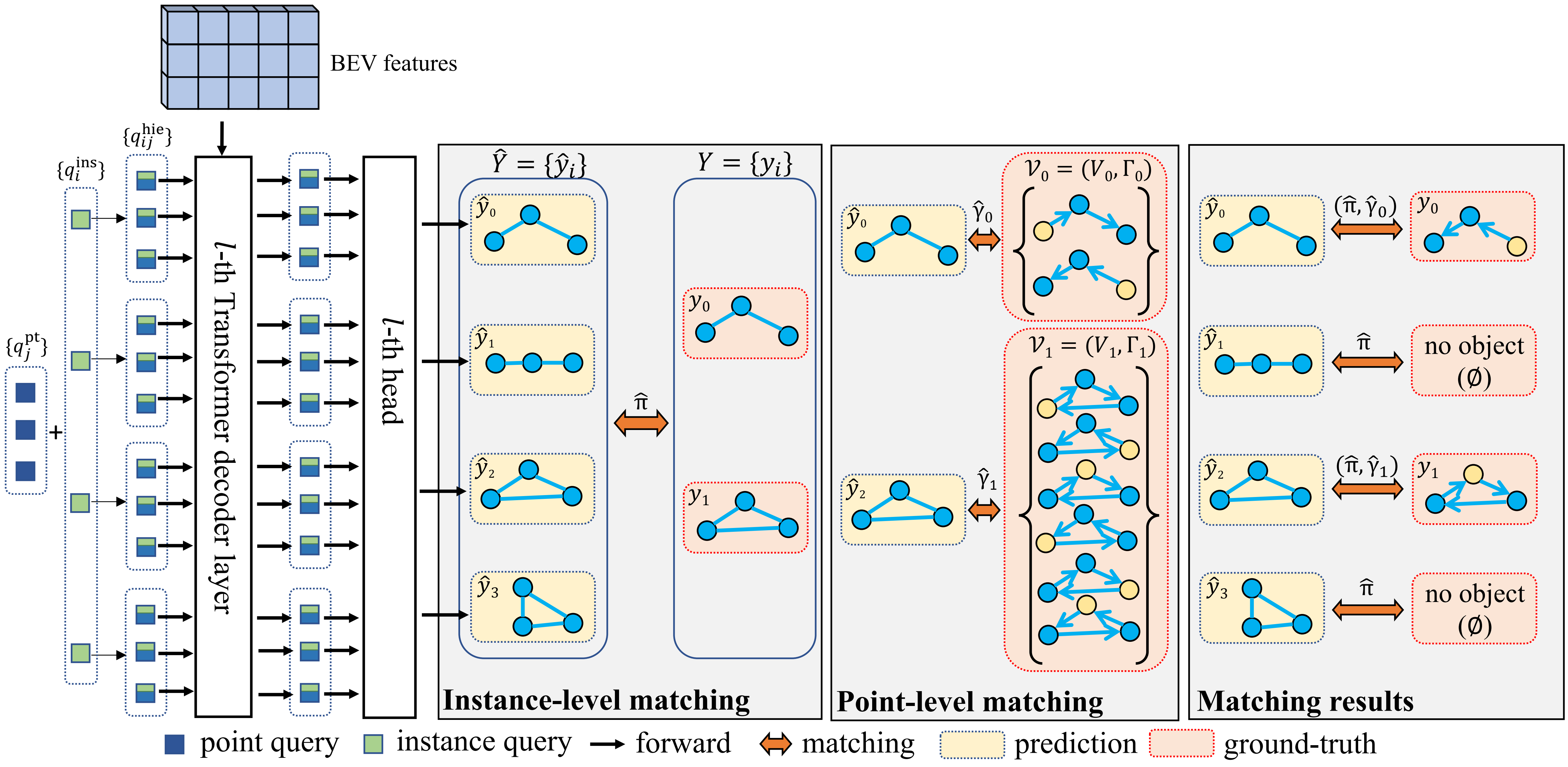
- point query: encode shared geometric info.
- instance query: encode element-specific info.

Parallel Interaction and Decoding



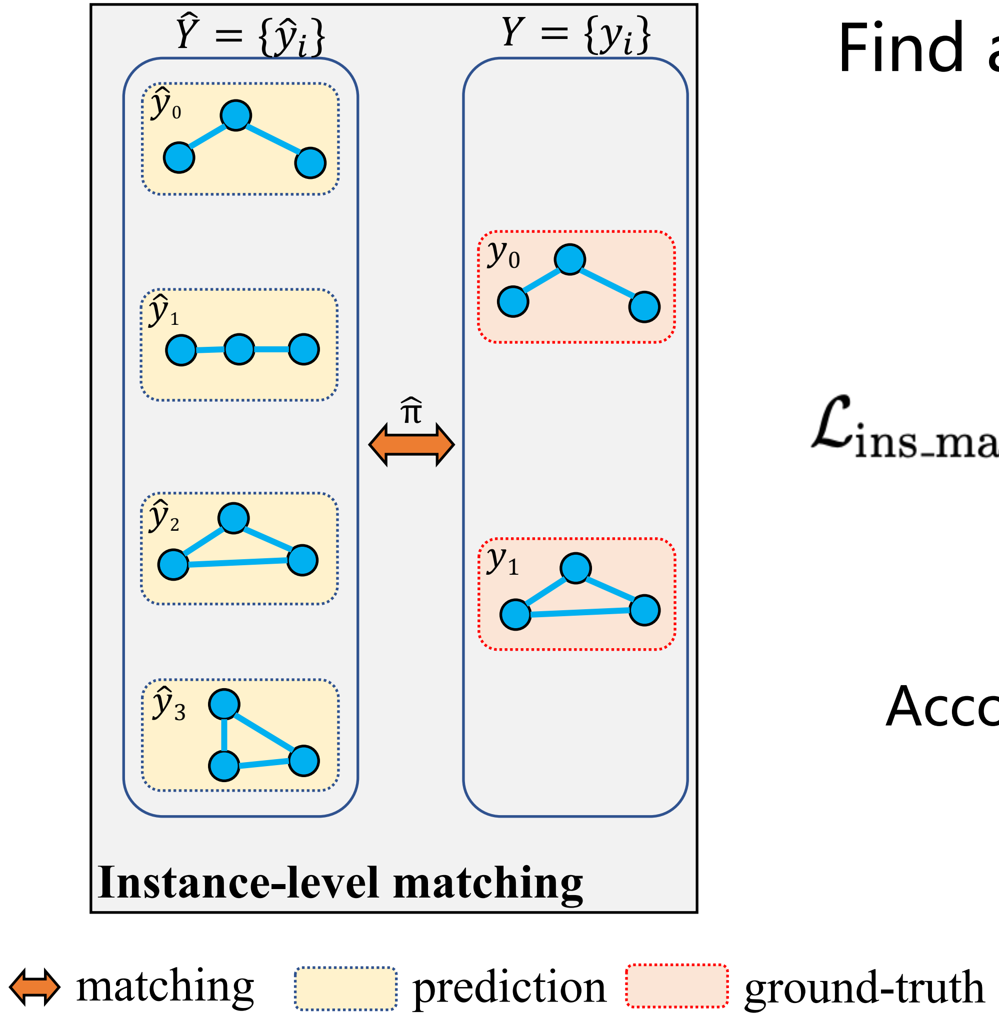
- Interaction with BEV features (cross-attention)
- Inter- and intra-instance interaction (self-attention)
- Parallelly output point sequence and class score

Hierarchical Matching



Instance-level Matching

地平线 Horizon Robotics



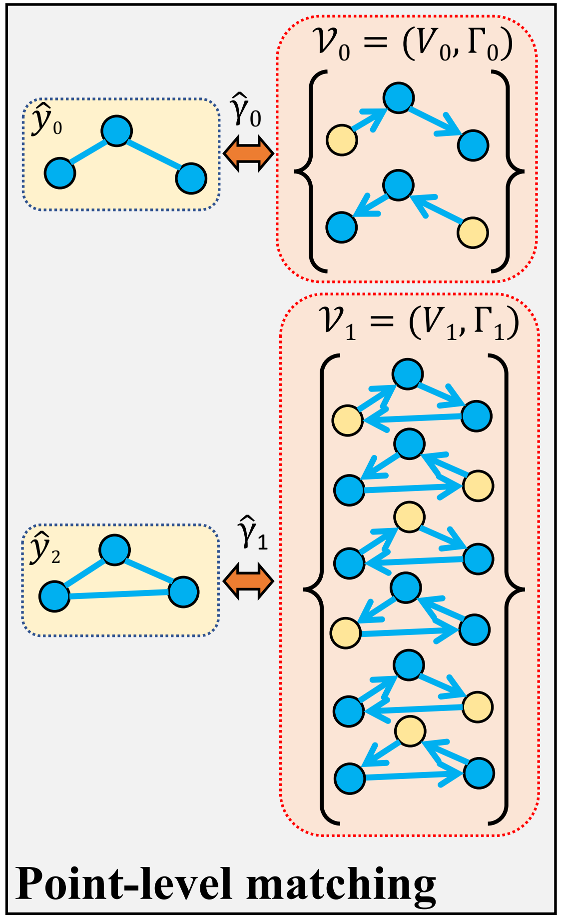
Find an optimal instance-level label assignment:

$$\hat{\pi} = \arg\min_{\pi \in \Pi_N} \sum_{i=0}^{N-1} \mathcal{L}_{\text{ins\_match}}(\hat{y}_{\pi(i)}, y_i).$$

$$\mathcal{L}_{\text{ins\_match}}(\hat{y}_{\pi(i)}, y_i) = \mathcal{L}_{\text{Focal}}(\hat{p}_{\pi(i)}, c_i) + \mathcal{L}_{\text{position}}(\hat{V}_{\pi(i)}, V_i).$$

According to class corelation and positional corelation

Point-level Matching



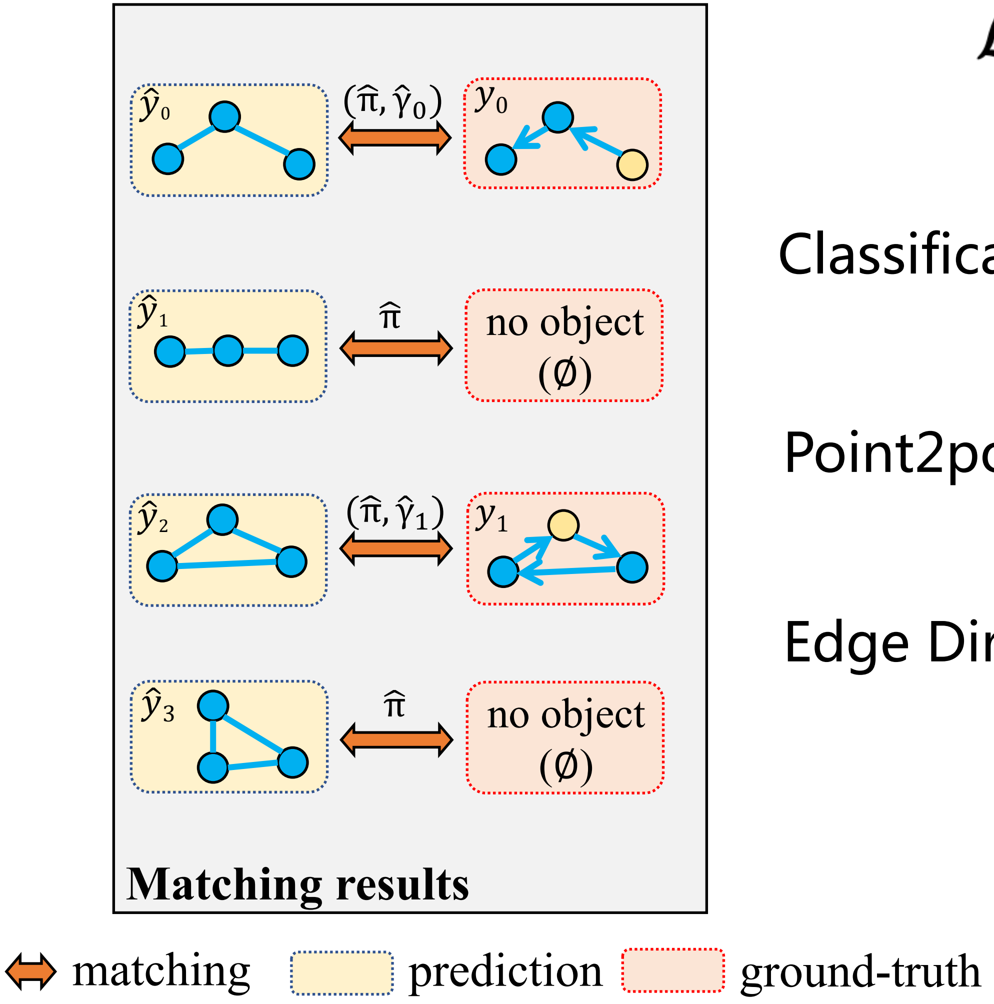
matching prediction ground-truth

Find an optimal point2point assignment among equivalent permutations:

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} \sum_{j=0}^{N_v - 1} D_{\mathrm{Manhattan}}(\hat{v}_j, v_{\gamma(j)}).$$

According to positional corelation

End-to-end Learning



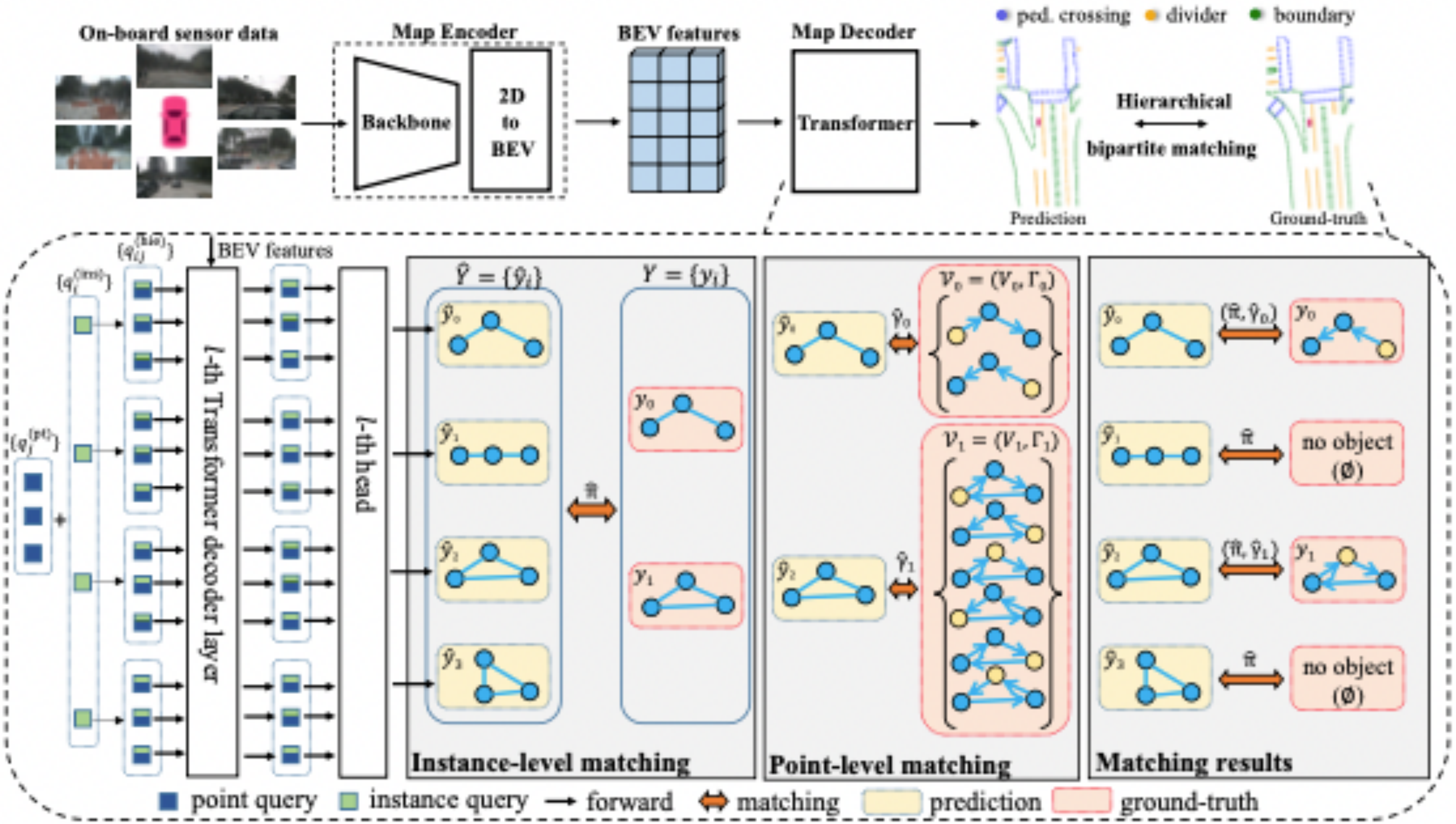
$$\mathcal{L} = \lambda \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{p2p}} + \beta \mathcal{L}_{\text{dir}}$$

Classification Loss: $$\mathcal{L}_{\text{cls}} = \sum_{i=0}^{N-1} \mathcal{L}_{\text{Focal}}(\hat{p}_{\hat{\pi}(i)}, c_i)$$

Point2point Loss: $$\mathcal{L}_{\text{p2p}} = \sum_{i=0}^{N-1} \mathbb{1}_{\{c_i \neq \varnothing\}} \sum_{j=0}^{N_v-1} D_{\text{Manhattan}}(\hat{v}_{\hat{\pi}(i),j}, v_{i,\hat{\gamma}_i(j)})$$

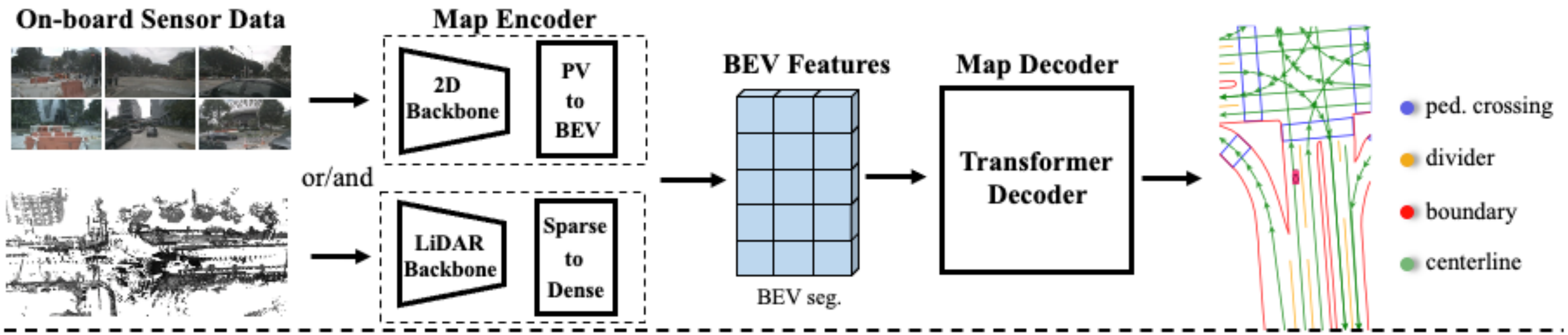
Edge Direction Loss: $$\mathcal{L}_{\text{dir}} = -\sum_{i=0}^{N-1} \mathbb{1}_{\{c_i \neq \varnothing\}} \sum_{j=0}^{N_v-1} \text{cosine\_similarity}(\hat{\boldsymbol{e}}_{\hat{\pi}(i),j}, \boldsymbol{e}_{i,\hat{\gamma}_i(j)})$$

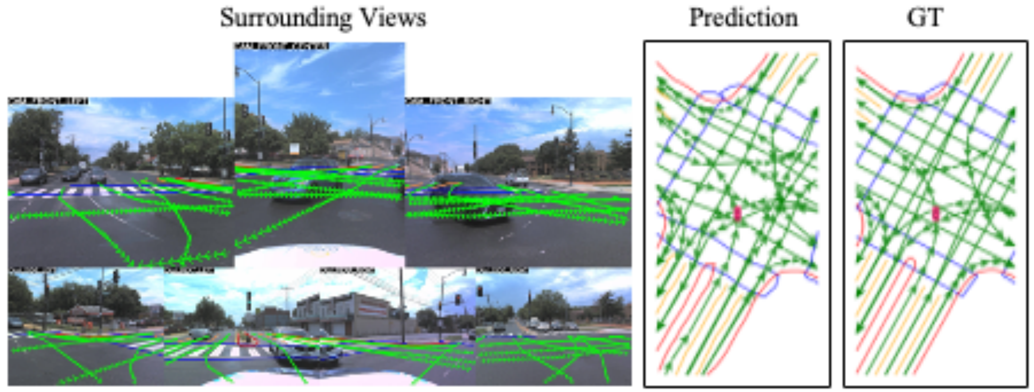
MapTRv1



- End-to-end (no rule-based post-processing)
- Parallel decoding and high efficiency

MapTR v2



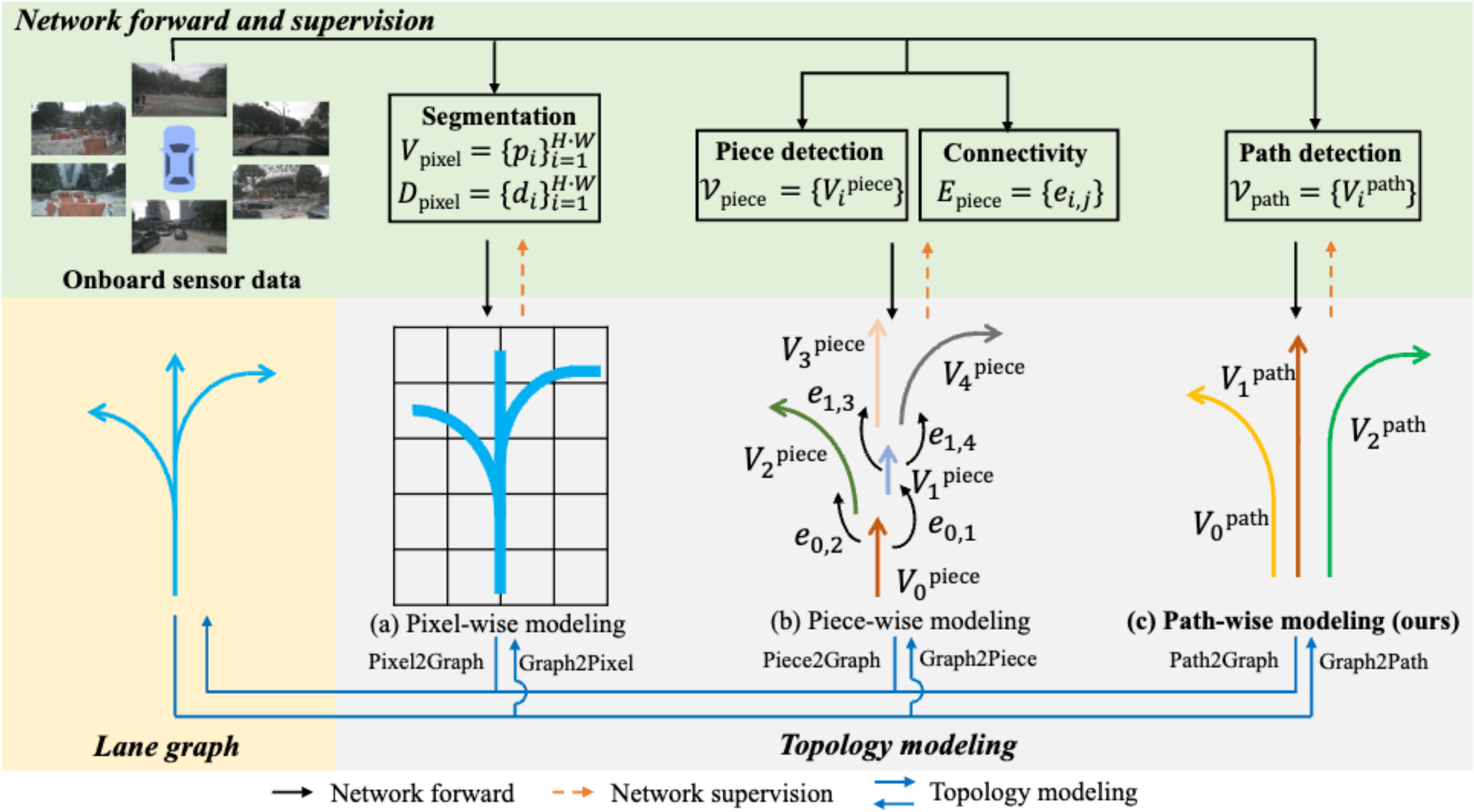


Qualitative result on Argoverse2 3D vectorized HD map, we render the predicted 3D vectorized map on the surrounding view images

Qualitative result on nuScenes 2D vectorized HD map

- Support lane topology modeling
- Support 3D mapping
- Improved model designs and training techniques

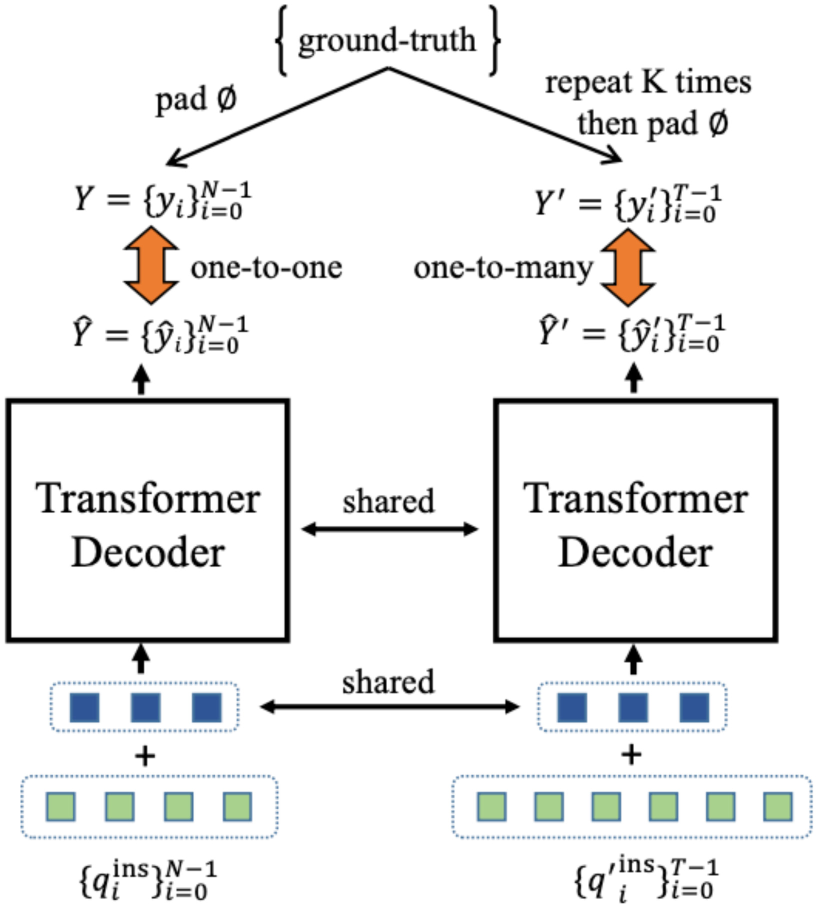
Lane Graph as Path (LaneGAP) for Lane Topology Modeling



- Convert irregular graph structure to structured path representation
- Merge / split points are less important and increase the convergence difficulty
- Path representation is compatible with downstream PnC task (as reference line)

Auxiliary Supervision

地平线 Horizon Robotics



- PV-based depth segmentation
- PV-based foreground segmentation
- BEV-based foreground segmentation

Auxiliary one2many branch

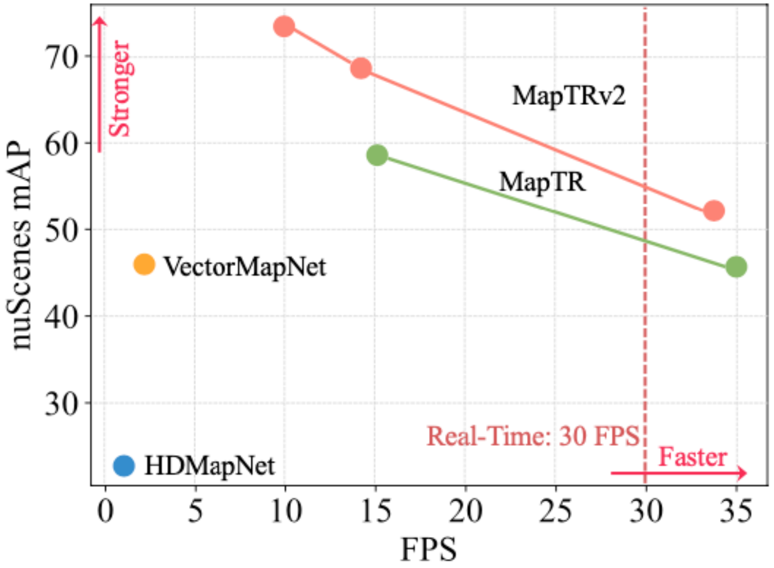
Auxiliary dense loss

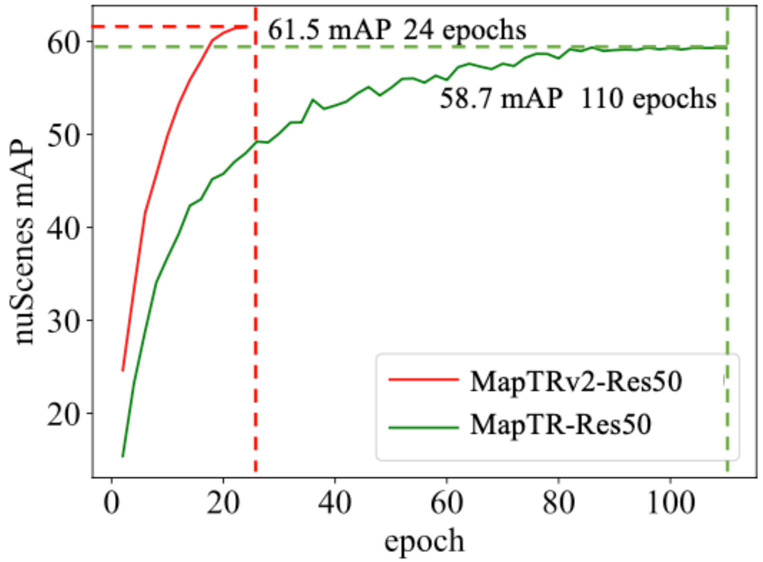
3D Mapping (Argoverse2)

2D point sequence to 3D point sequence

MapTR Series

Method	Modality	Backbone	Epoch	AP ped.	AP div.	AP bou.	AP mean	FPS
HDMapNet	C	Effi-B0	30	14.4	21.7	33.0	23.0	0.9
	L	PP	30	10.4	24.1	37.9	24.1	1.1
	C & L	Effi-B0 & PP	30	16.3	29.6	46.7	31.0	0.5
VectorMapNet	C	R50	110+ft	42.5	51.4	44.1	46.0	2.2
	L	PP	110	25.7	37.6	38.6	34.0	-
	C & L	R50 & PP	110+ft	48.2	60.1	53.0	53.7	-
MapTR	C	R18	110	39.6	49.9	48.2	45.9	**35.0**
	C	R50	110	56.2	59.8	60.1	58.7	15.1
	C	R50	**24**	46.3	51.5	53.1	50.3	15.1
	L	Sec	**24**	48.5	53.7	64.7	55.6	8.0
	C & L	R50 & Sec	**24**	55.9	62.3	69.3	62.5	6.0
MapTRv2	C	R18	110	46.9	55.1	54.9	52.3	33.7
	C	R50	110	68.1	68.3	69.7	68.7	14.1
	C	V2-99	110	**71.4**	**73.7**	**75.0**	**73.4**	9.9
	C	R50	**24**	59.8	62.4	62.4	61.5	14.1
	C	V2-99	**24**	63.6	67.1	69.2	66.6	9.9
	L	Sec	**24**	56.6	58.1	69.8	61.5	7.6
	C & L	R50 & Sec	**24**	65.6	66.5	74.8	69.0	5.8





- Real-time (up to 30 FPS)
- Convergence and performance

MapTR Series (nuScenes)

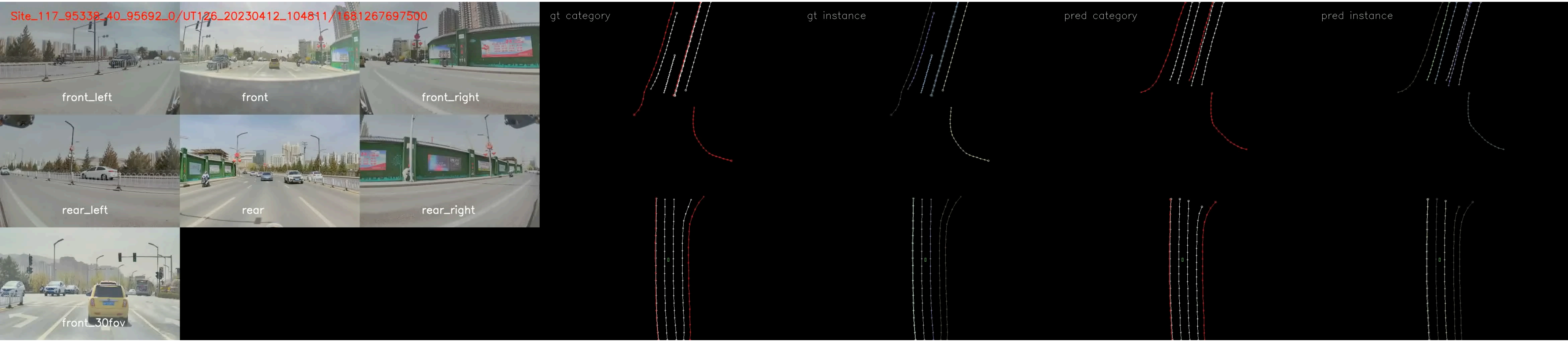
Sunny & cloudy

Rainy

Night

Surrounding Views Prediction GT

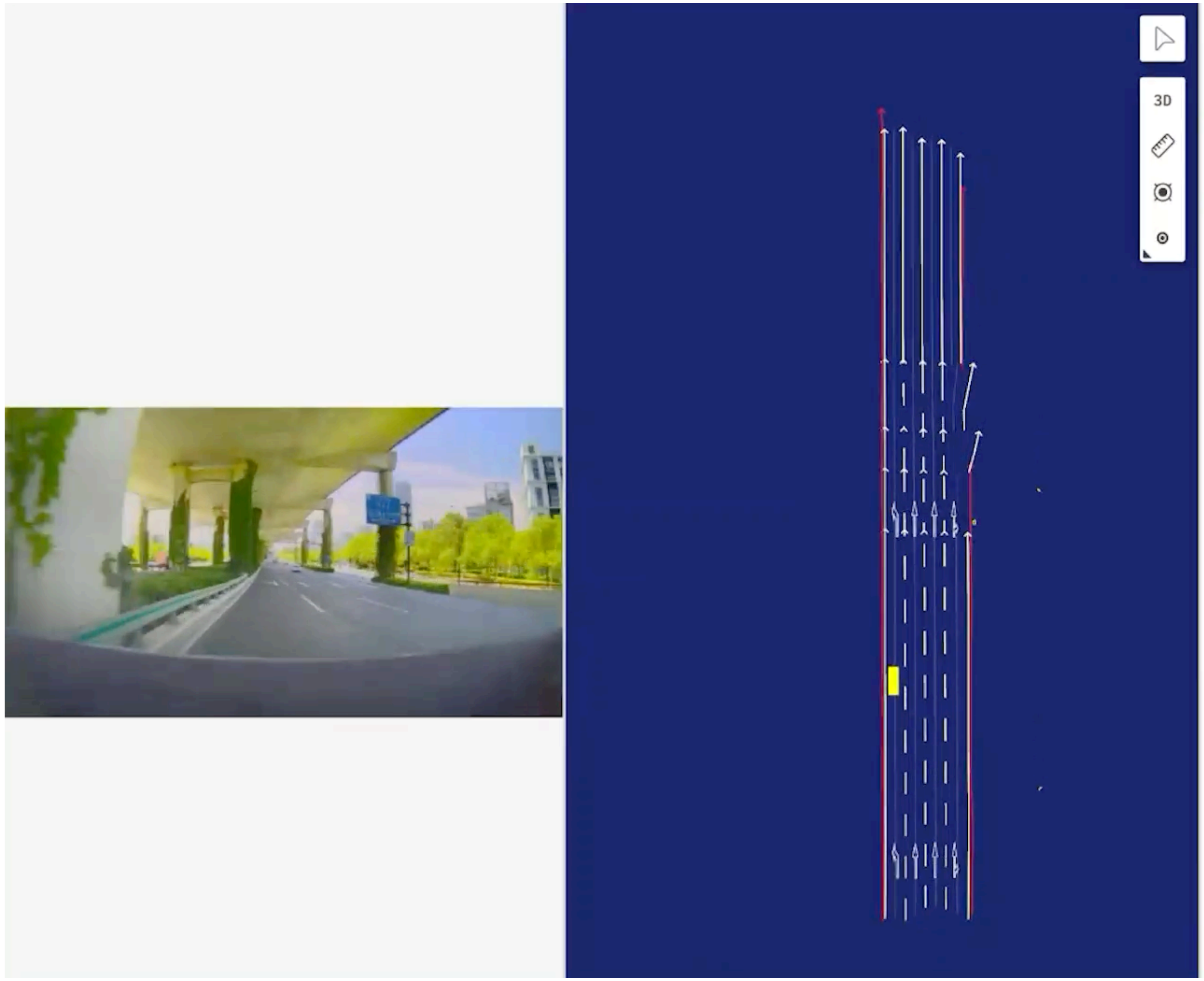
Long-range Mapping

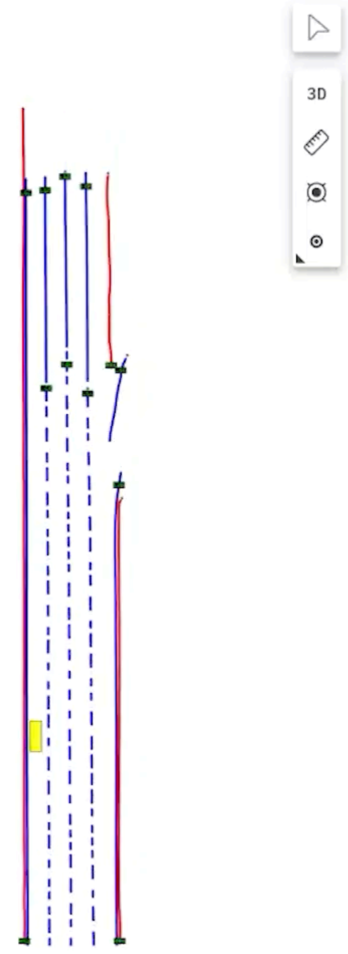


Map Perception Range: 120m

- Modeling in higher level and requiring more data
- Leveraging tens of millions of training samples
- Outperforming seg.-based methods with data scaling up

Complex Scenarios

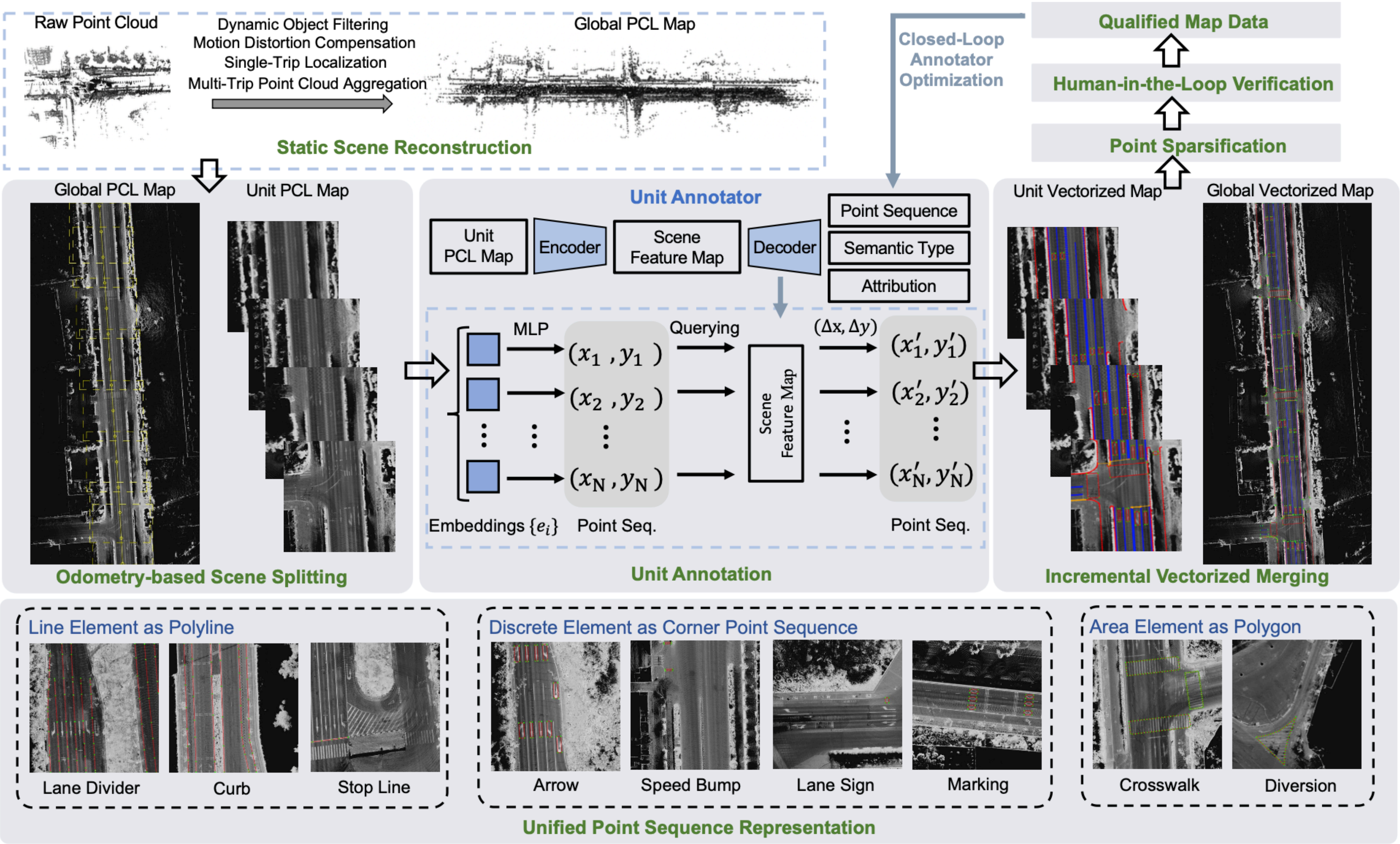




Shanghai Zhangjiang

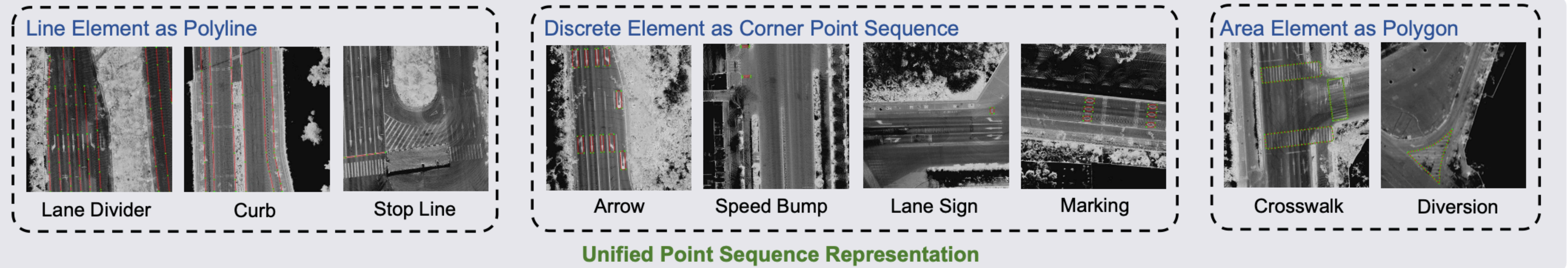
- Imaging ability and element completeness

Auto Labeling: VMA



extending MapTR to a general cloud-end map auto labeling framework

Extensibility

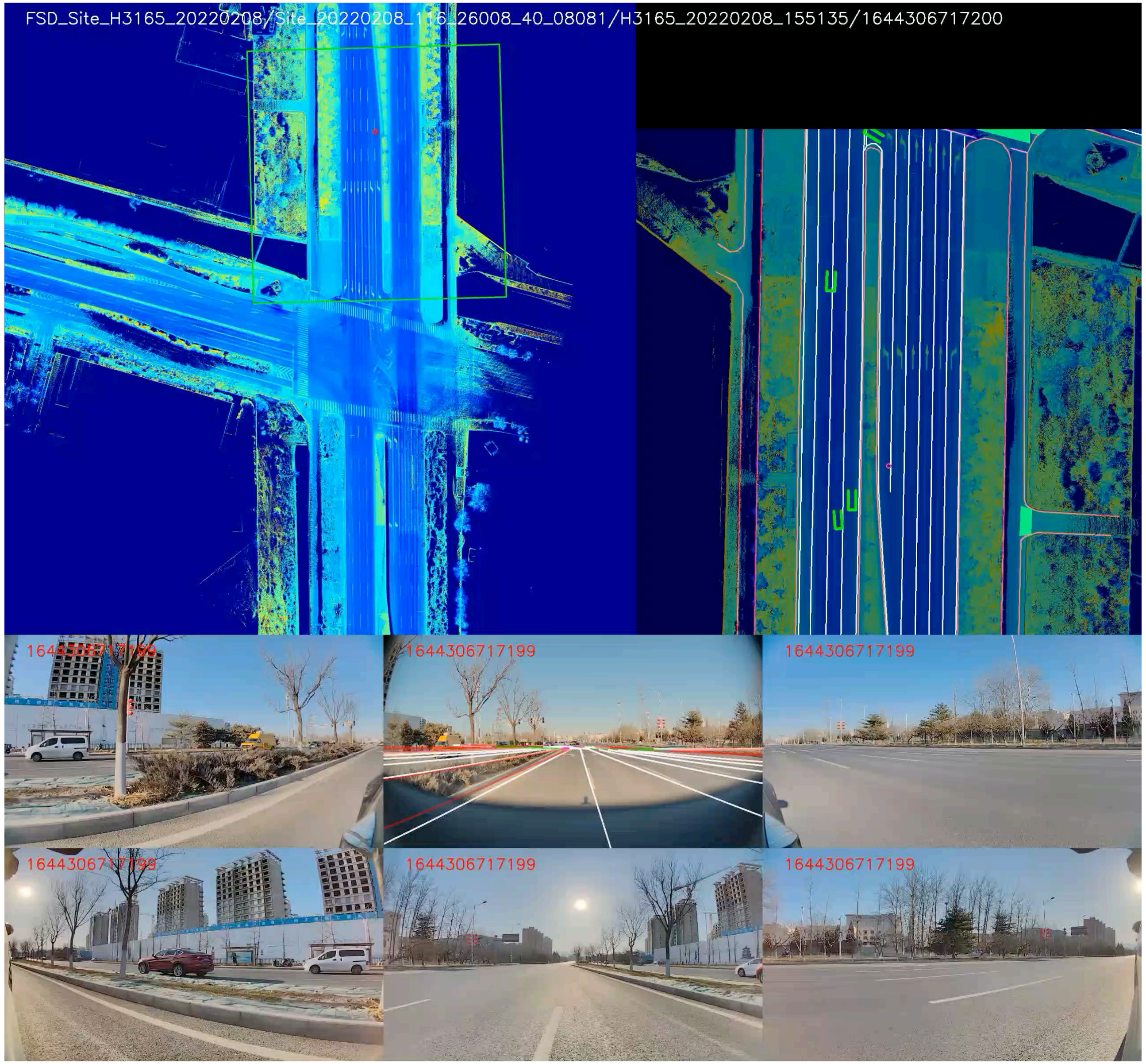


- All scenarios (highway / urban / parking)
- A wide range of elements (line / discrete / area)
- Attributions (color / direction / type)
- Driving scenarios:

Geometric Type	Vectorized Representation	Semantic Type	Attribution
Line Element	N-Point Sequence (Polyline)	Lane Divider Curb Stop Line ...	Direction: Unidirectional / Bidirectional; Line Type: Solid / Dotted / Fishbone; ... Curb Type: Ground Side / Road Side / Guardrail - ...
Discrete Element	Corner Point Sequence	Arrow Speed Bump Lane Sign Marking ...	Arrow Type: Straight / Turn Off / Merge Right / No Turn Left / ...; ... - Lane Sign Type: Bike Lane / Bus Lane Marking Type: Diamond Marking / Inverted Triangle Marking ...
Area Element	N-Point Sequence (Polygon)	Crosswalk Diversion ...	- - ...

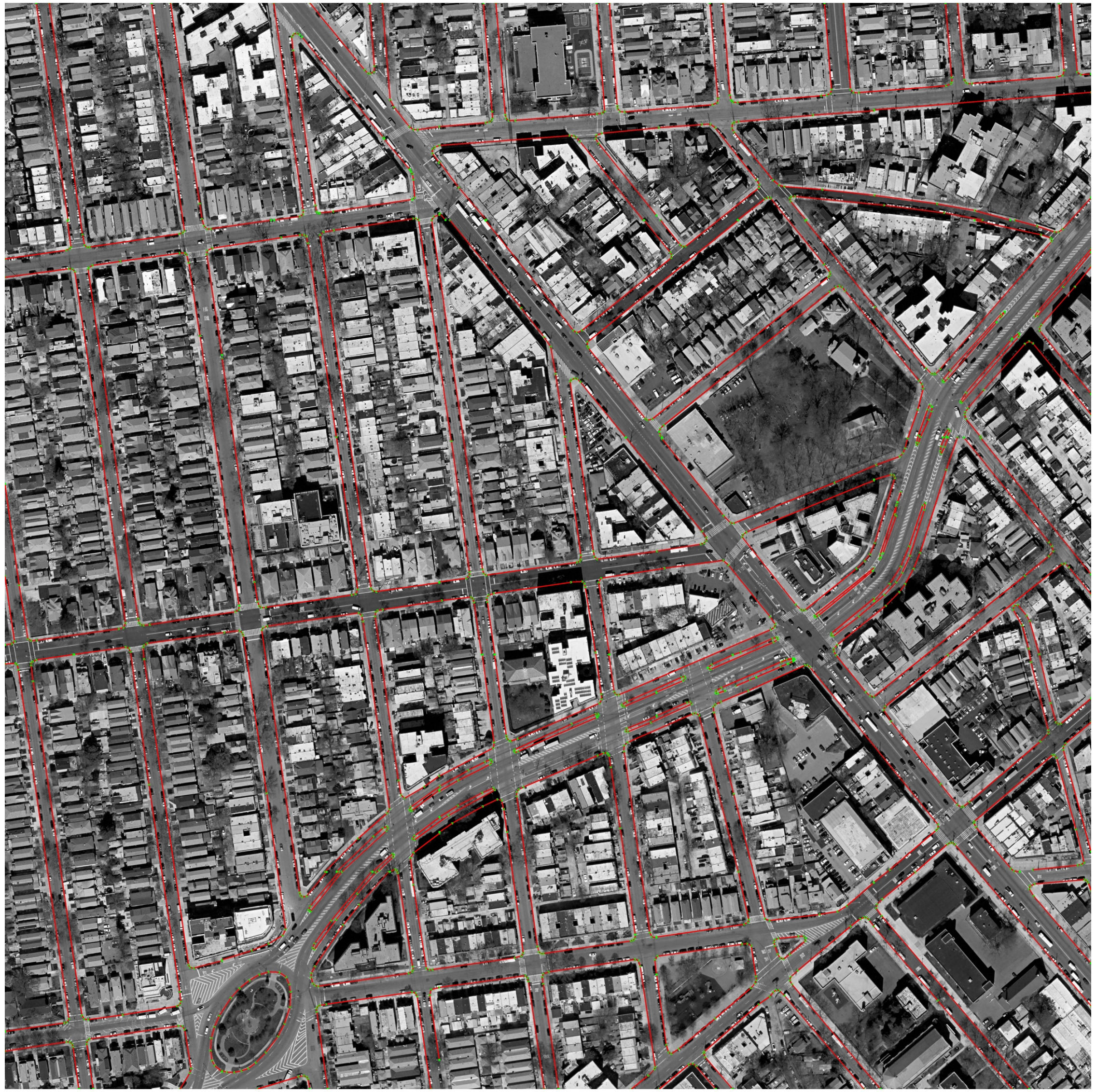
- Parking scenarios: Parking Lock / Cement Column / No Parking Line ...

Auto Labeling

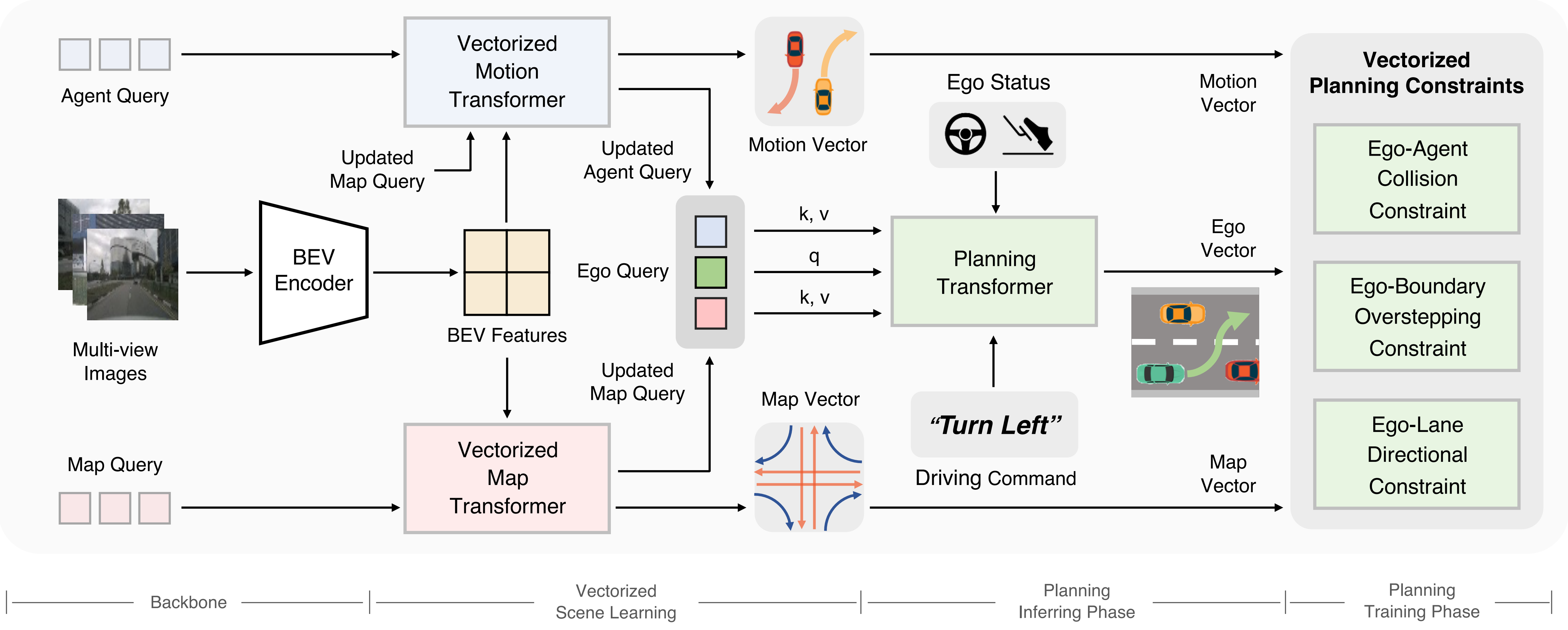


Beijing North 4th Ring Road

Auto Labeling: Remote Sensing

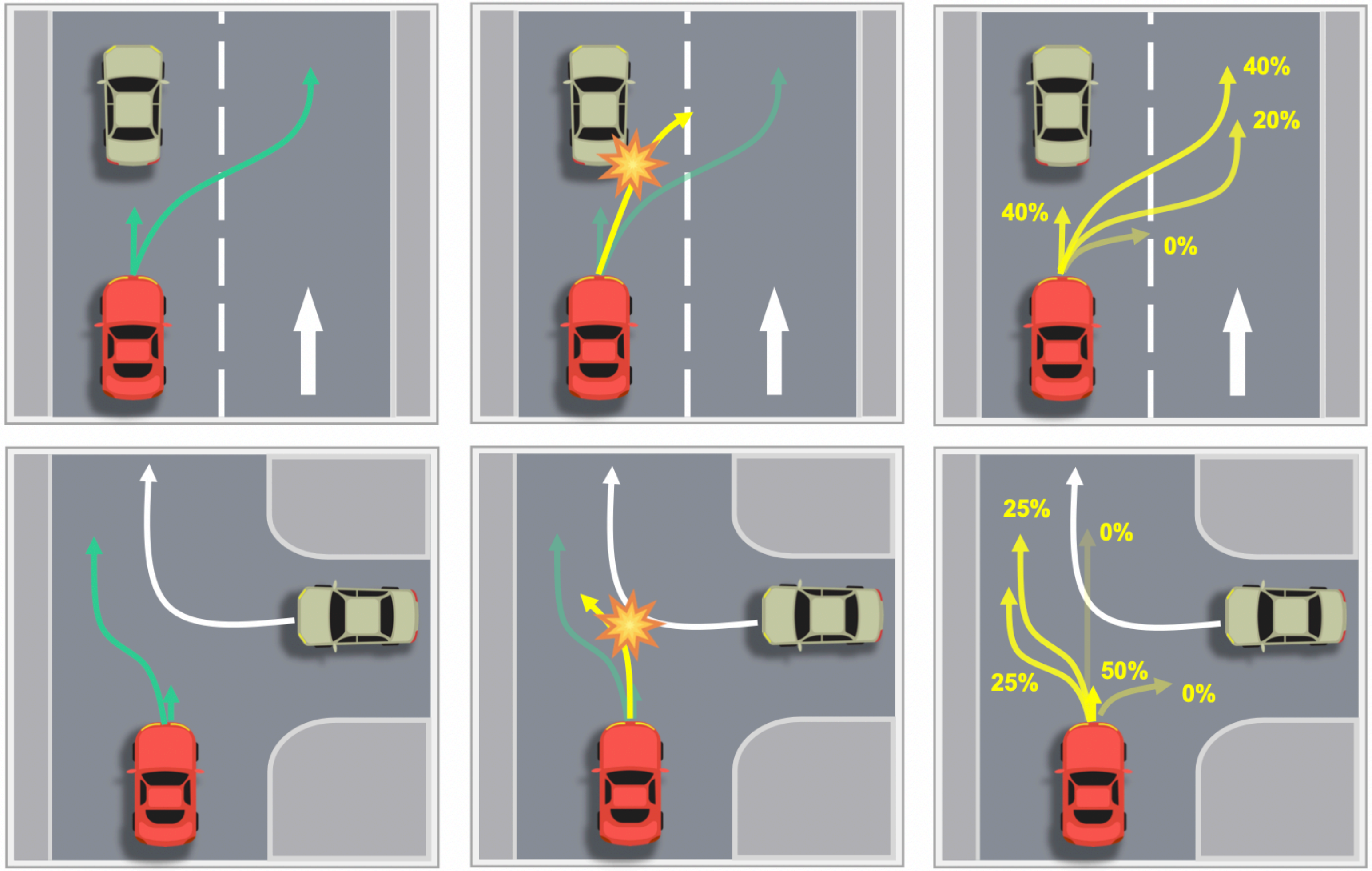


End-to-end Planning: VADv1



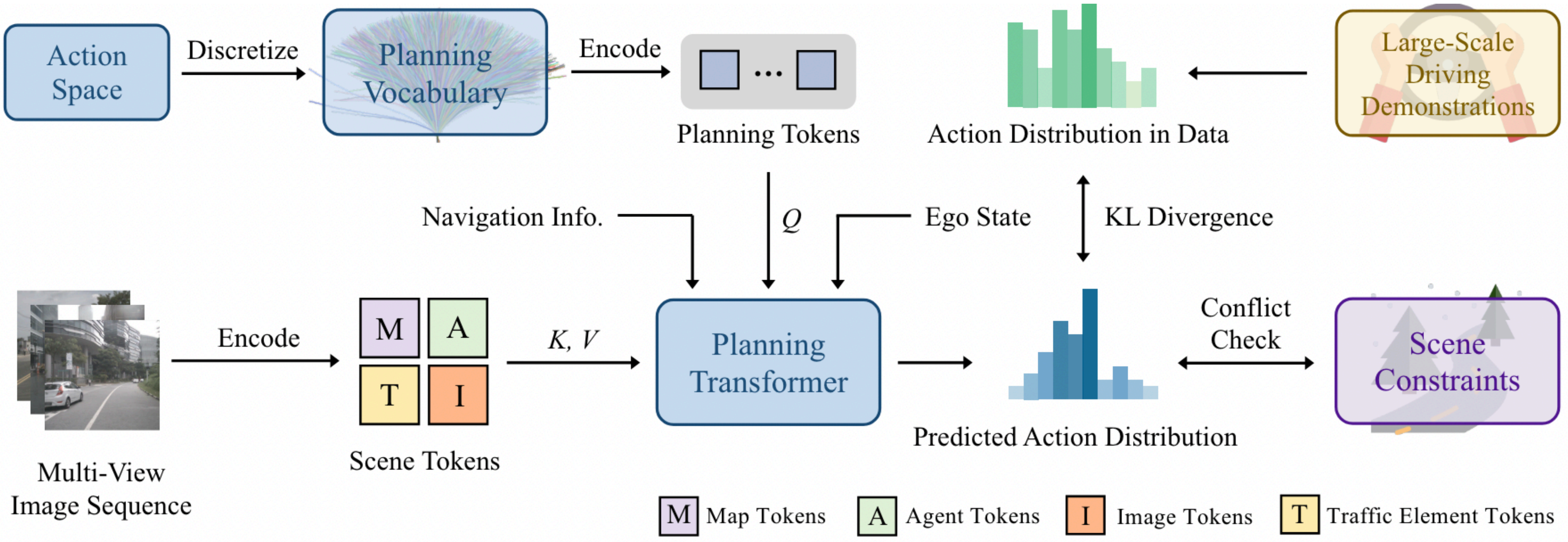
VADv1: extending MapTR to end-to-end planning

Uncertainty of Planning



Driving Demonstrations	Deterministic Planning	Probabilistic Planning
• Uncertainty of scenario human behavior	• Modeling deterministic relation between environment and action	• Modeling environment-conditioned probabilistic distribution of action

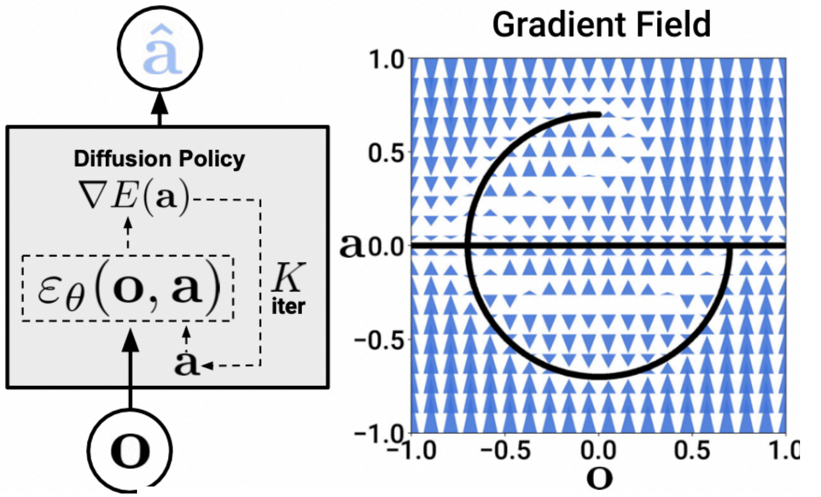
VADv2



- Output the probabilistic distribution of trajectories, easy to combine with rule-based and optimization-based PnC (as post-solver)
- Satisfying kinematic constraints & consistency with ego state (compared with linear regression)
- Get the confidence score of each action (how confidence the e2e model is)

VADv2 vs. Diffusion Policy

- Both aiming at modeling the multimodal action distribution,
- solving the uncertainty of planning
- Action space of autonomous driving (only spatiotemporal trajectory) is relatively small than robotics (tens of degrees of freedom)
- Discretizing and scoring is feasible
- When the granularity is small enough, discretization error is negligible

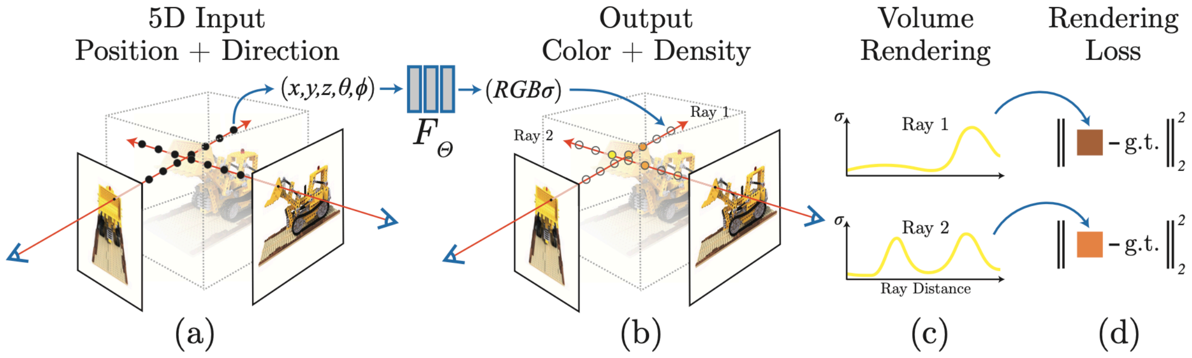


Diffusion Policy

Chi et.al. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. RSS 2023.

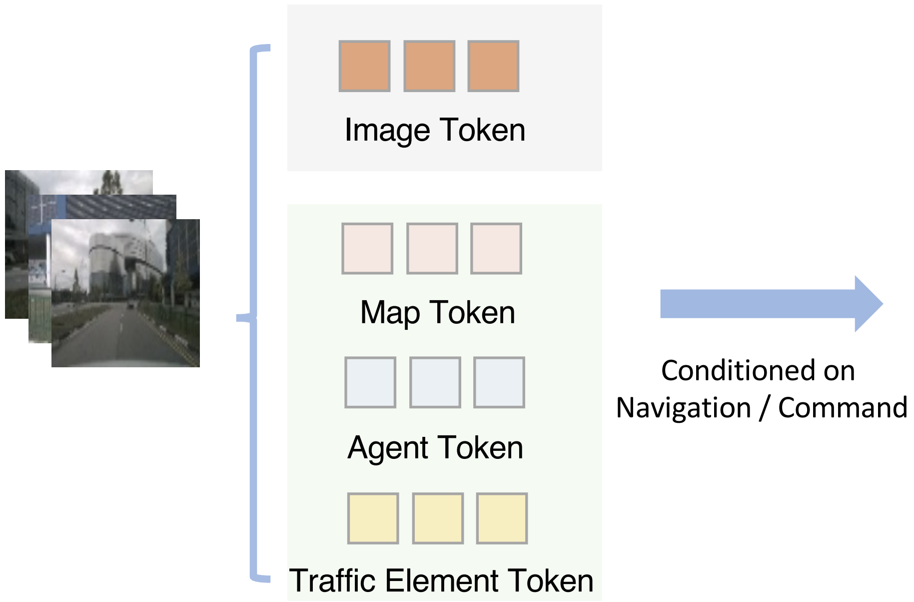
Spatially-continuous Tokenization and Planning Probabilistic Field

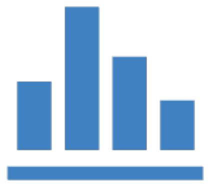
NeRF



$$r, g, b, \sigma = f(x, y, z, \theta, \phi)$$

Planning Probabilistic Field





$$p(\boldsymbol{a}) = \sigma(\mathrm{MLP}(\phi(E(\boldsymbol{a}), E_{\mathrm{scene}}) + E_{\mathrm{navi}} + E_{\mathrm{state}})).$$

$$E(\boldsymbol{a}) = \big(\Gamma(x_1), \Gamma(y_1), ..., \Gamma(x_{\mathrm{T}}), \Gamma(y_{\mathrm{T}})\big),$$

$$\Gamma(\mathrm{pos}) = \big(\gamma(\mathrm{pos}, 0), \gamma(\mathrm{pos}, 1), ..., \gamma(\mathrm{pos}, \mathrm{L}-1)\big),$$

$$\gamma(\mathrm{pos}, \mathrm{j}) = \big(\cos(\mathrm{pos}/1\mathrm{e}4^{2\pi\mathrm{j}/\mathrm{L}}), \sin(\mathrm{pos}/1\mathrm{e}4^{2\pi\mathrm{j}/\mathrm{L}})\big).$$

CARLA Closed-Loop Simulation

CARLA Town05 10 miles long route

Sampling top1 action w/o post-processing

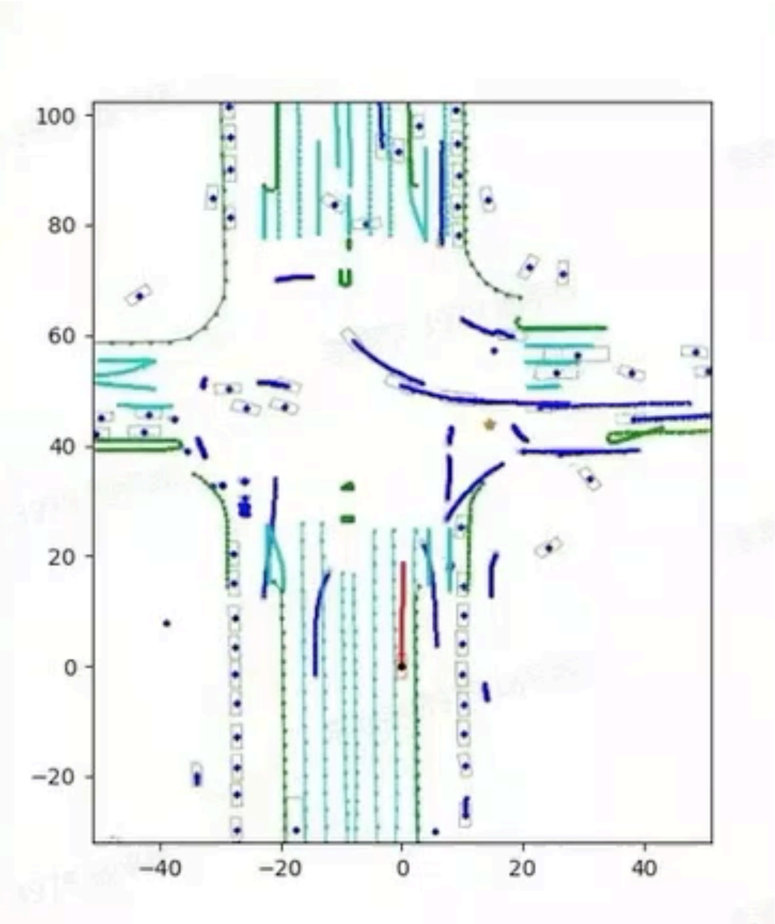
CARLA Benchmark

地平线 Horizon Robotics

Method	Modality	Reference	Driving Score ↑	Route Completion ↑	Infraction Score ↑
CILRS [9]	C	CVPR 19	7.8	10.3	0.75
LBC [6]	C	CoRL 20	12.3	31.9	0.66
Roach [54]	C	ICCV 21	41.6	96.4	0.43
Transfuser† [40]	C+L	TPAMI 22	31.0	47.5	0.77
ST-P3 [18]	C	ECCV 22	11.5	83.2	-
VAD [23]	C	ICCV 23	30.3	75.2	-
ThinkTwice [21]	C+L	CVPR 23	70.9	95.5	0.75
MILE [16]	C	NeurlIPS 22	61.1	97.4	0.63
Interfuser [45]	C	CoRL 22	68.3	95.0	-
DriveAdapter+TCP [20]	C+L	ICCV 23	71.9	97.3	0.74
DriveMLM [49]	C+L	arXiv	76.1	98.1	0.78
VADv2	C	Ours	85.1	98.4	0.87

Town05 Long

Real-World Scenarios



Shanghai Zhangjiang

Thanks

MapTR v1

Paper: https://arxiv.org/abs/2208.14437

Project Page: https://github.com/hustvl/MapTR

MapTR v2

Paper: https://arxiv.org/abs/2308.05736

Project Page: https://github.com/hustvl/MapTR

LaneGAP

Paper: https://arxiv.org/abs/2303.08815

Project Page: https://github.com/hustvl/LaneGAP

VAD v1

Paper: https://arxiv.org/abs/2303.12077

Project Page: https://github.com/hustvl/VAD

VAD v2

Paper: https://arxiv.org/abs/2402.13243

Project Page: https://hgao-cv.github.io/VADv2

VMA

Paper: https://arxiv.org/abs/2304.09807

Project Page: https://github.com/hustvl/MapTR