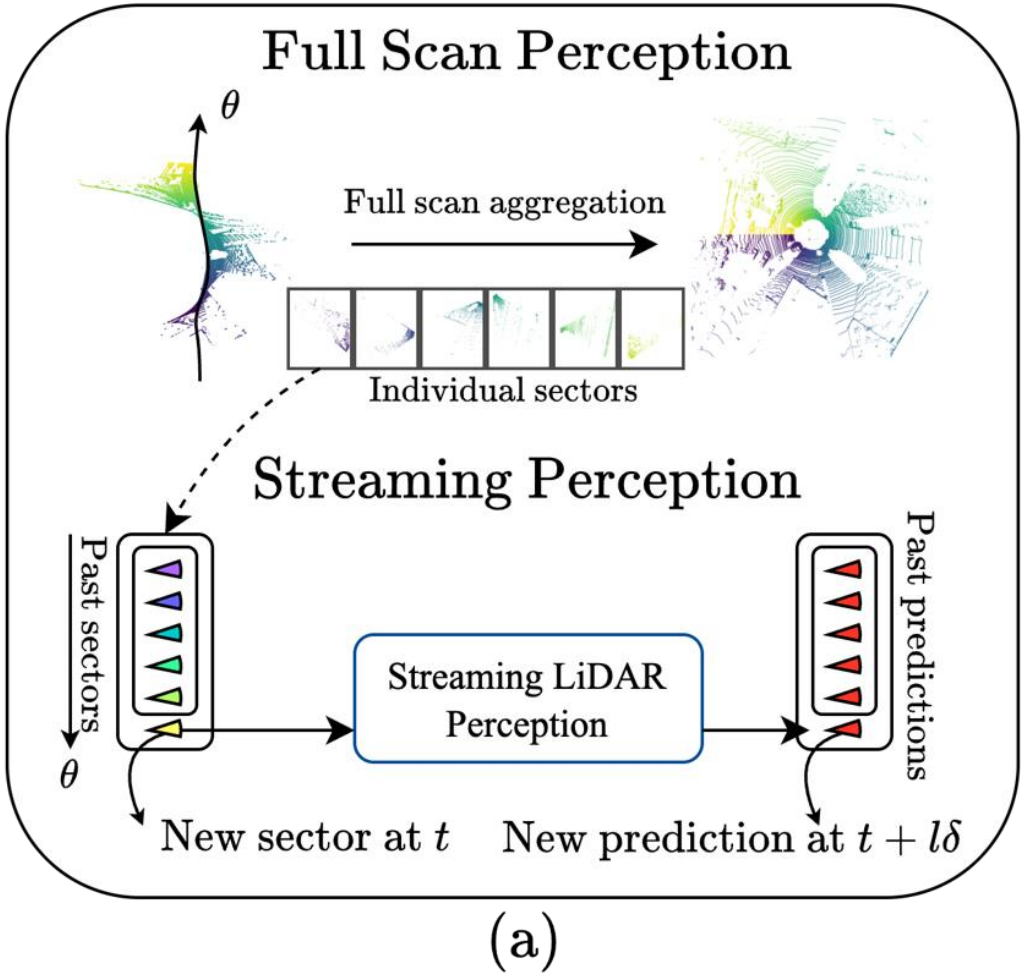
Towards Streaming LiDAR Object Detection with Point Clouds as Egocentric Sequences

Mellon M. Zhang, Glen Chou[1], Saibal Mukhopadhyay[2]
Trustworthy Robotics Lab[1], GREEN Lab[2]
Work available at https://www.arxiv.org/abs/2506.06944

The issue with full-scan methods

- Full scan methods are **slow**
  - Need to wait for full scan accumulation (~100 ms)
  - Processing entire point cloud at once leads to compute spikes
  - Artificial sensor latency leads to capped throughput and less reactiveness
- Streaming methods are **fast**
  - Process partial sectors of point clouds on-the-fly (~100/N ms)
  - Compute requirements are spread out along entire scanning process of the LiDAR sensor
  - Higher max throughput for more reactiveness



The issue with streaming methods

We identify two main gaps preventing streaming methods from approaching the performance of full-scan methods.

1. Information gap: full-scene dependencies/context is lost
   - Latent hidden states are not enough to maintain full context

2. Architecture gap: design choices have not evolved to support the polar space
   - No translation invariance assumption

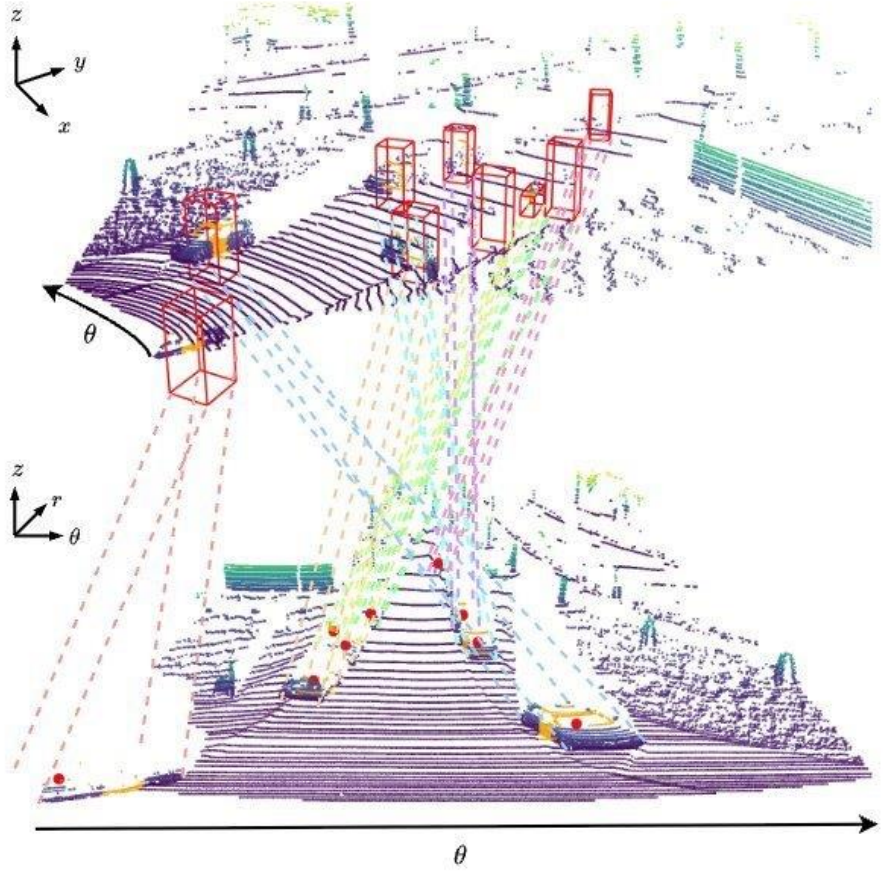


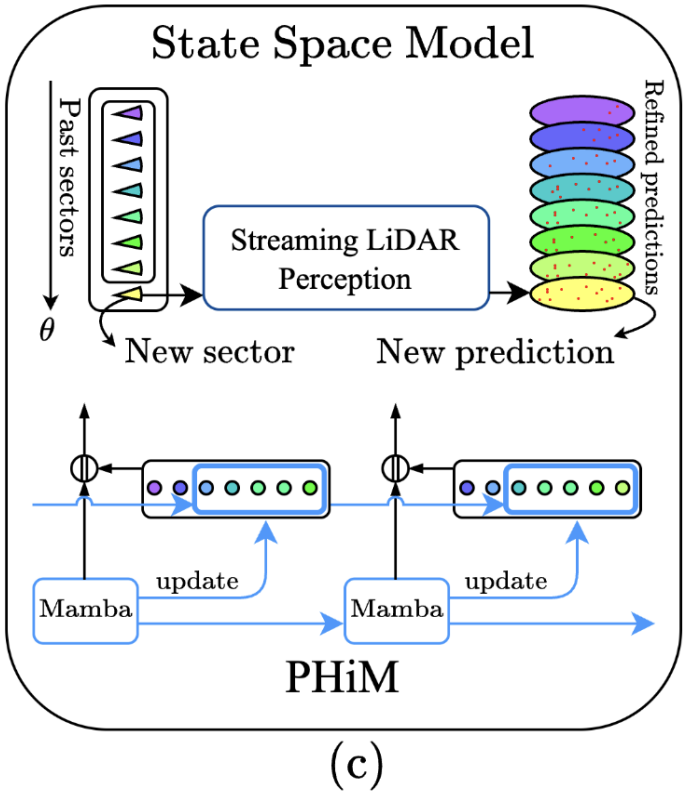
Figure 4: Cartesian BEV $(x-y)$ view of the point cloud with LiDAR scanning direction indicated by red lines. Distortion introduced by the polar coordinate system varies across the $r-\theta$ plane, hampering the learning of translation-invariant features with convolution kernels.

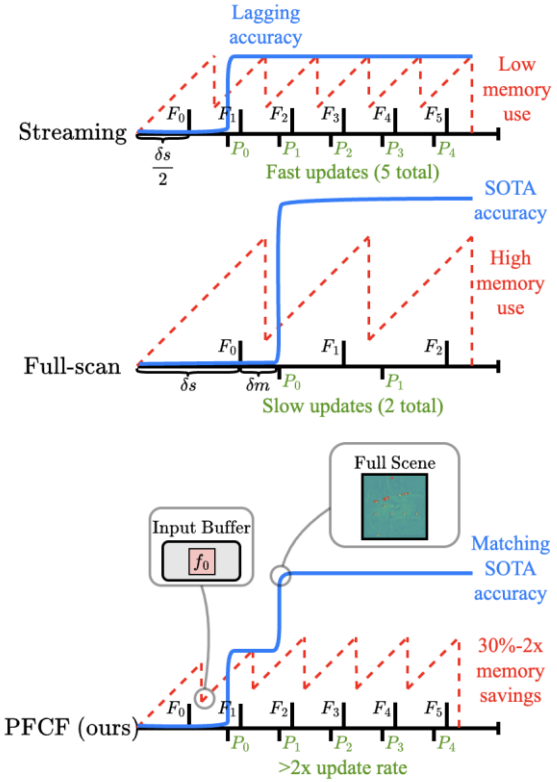
Motivation

*How can we bridge the full-scan and streaming paradigms to design perception algorithms that are **accurate, compute-efficient, and reactive**?*

We design a proof-of-concept approach that addresses both gaps:

- Architecture gap: Fast partial-sector processing with a streaming 3D backbone [**Polar Hierarchical Mamba (PHiM)**]
- Information gap: Lightweight sector buffer to store global context [**Polar-Fast-Cartesian-Full (PFCF)**]
- Full-context processing with full-scan 2D backbone



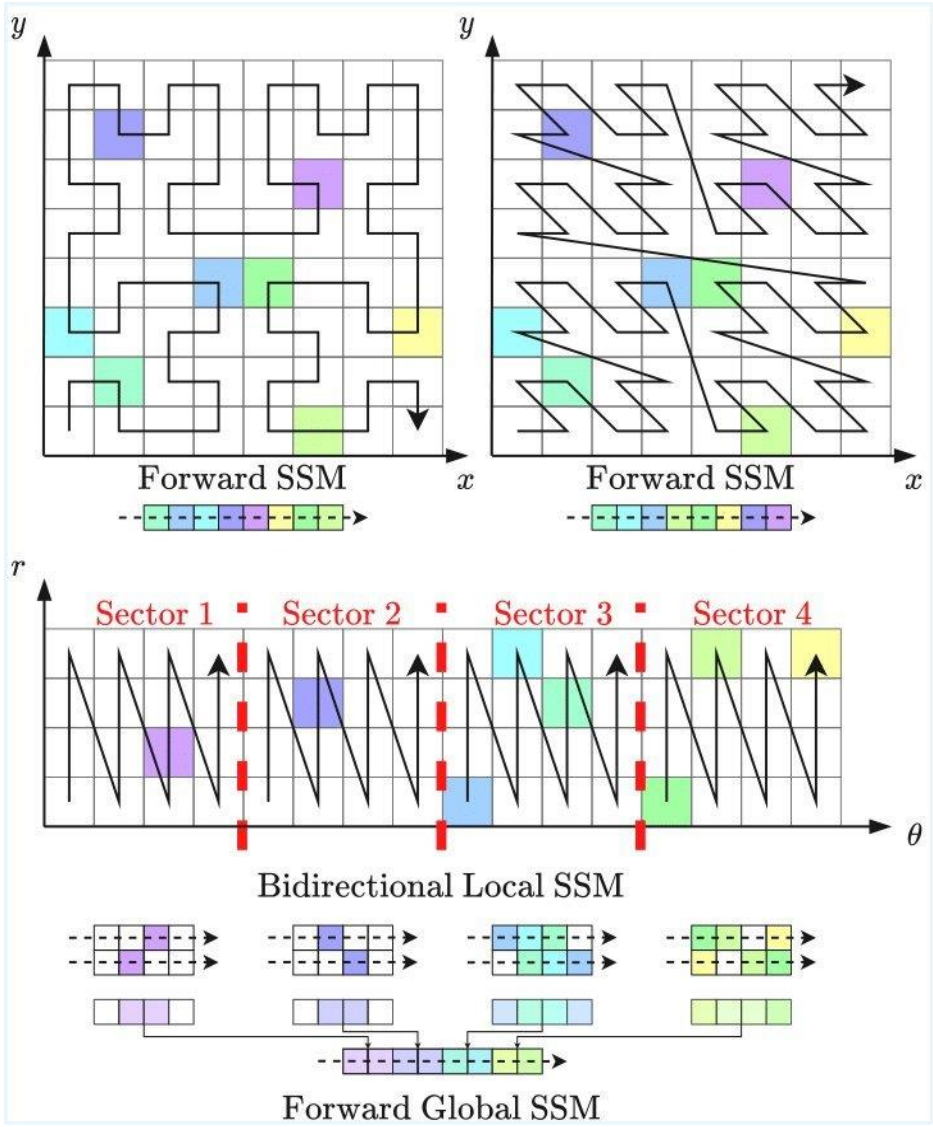


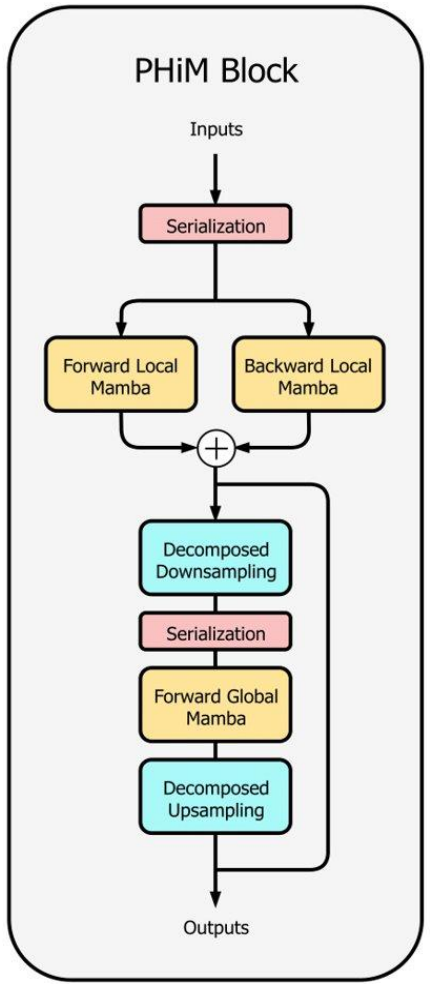
Polar Hierarchical Mamba (PHiM) Streaming Backbone

Input-time serialization (θ axis) saves memory:
- No stored memory patterns
- No need for recomputed patterns for different streaming granularities

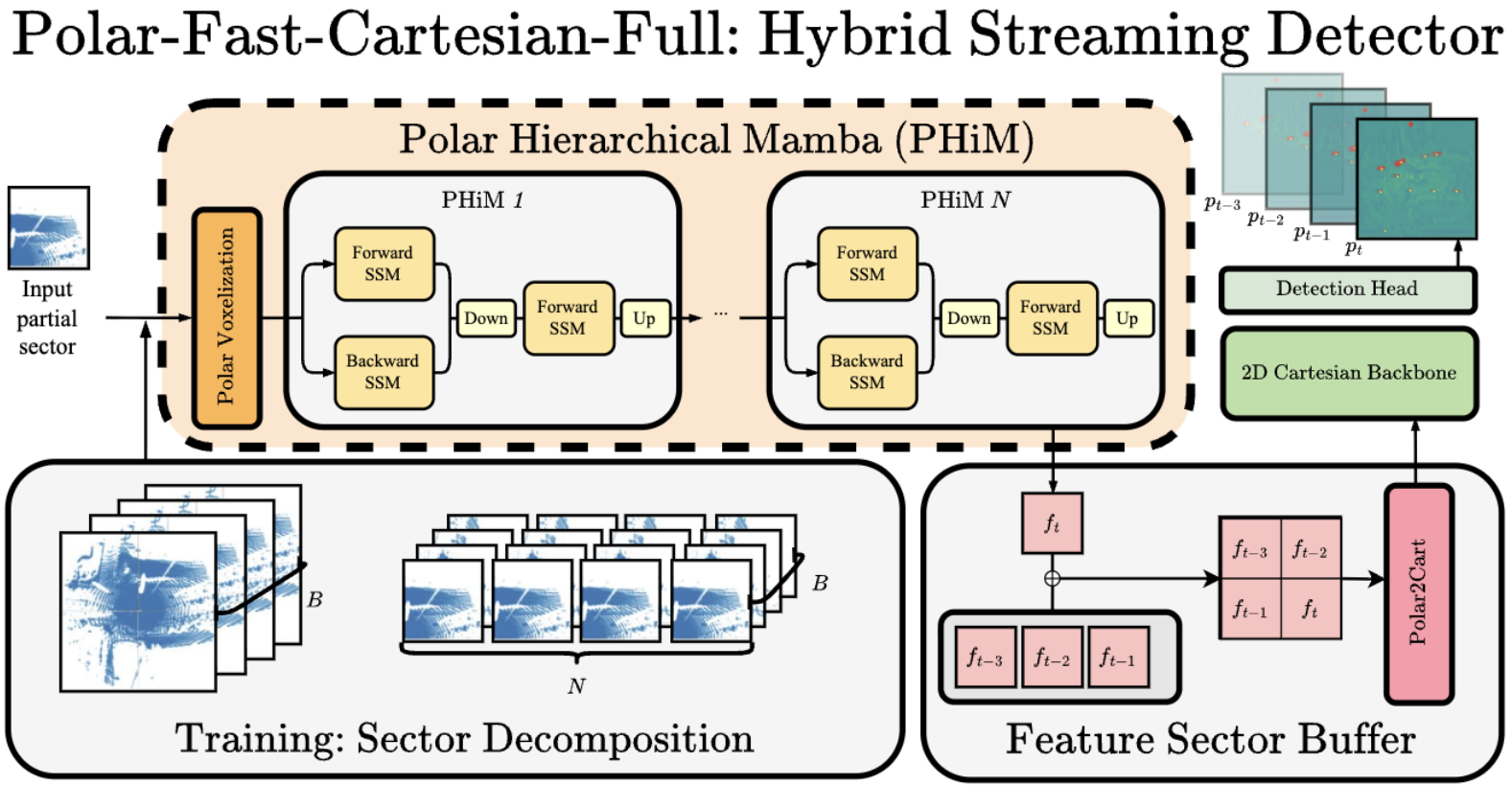
Bidirectional SSM fully captures local sector dependencies

Forward SSM captures global spatiotemporal relations while respecting causality





Polar-Fast-Cartesian-Full (PFCF)



- A general pipeline for training streaming methods on full-scan datasets.
- Feature sector buffer stores latent intermediate features, stitches together the features for downstream full-context perception

Evaluation and Results

Type	Method	mAP/mAPH		Vehicle AP/APH		Pedestrian AP/APH		Cyclist AP/APH	
		L1	L2	L1	L2	L1	L2	L1	L2
Full	SECOND [43]	67.2/63.1	61.0/57.2	72.3/71.7	63.9/63.3	68.7/58.2	60.7/51.3	60.6/59.3	58.3/57.0
	PointPillar [23]	69.0/63.5	62.8/57.8	72.1/71.5	63.6/63.1	70.6/56.7	62.8/50.3	64.4/62.3	61.9/59.9
	CenterPoint [44]	75.9/73.5	69.8/67.6	76.6/76.0	68.9/68.4	79.0/73.4	71.0/65.8	72.1/71.0	69.5/68.5
	DSVT-Voxel [42]	80.3/78.2	74.0/72.1	79.7/79.3	71.4/71.0	83.7/78.9	76.1/71.5	77.5/76.5	74.6/73.7
	HEDNet [45]	81.4/79.4	75.3/73.4	81.1/80.6	73.2/72.7	84.4/80.0	76.8/72.6	78.7/77.7	75.8/74.9
	VoxelNeXt [7]	78.6/76.3	72.2/70.1	78.2/77.7	69.9/69.4	81.5/76.3	73.5/68.6	76.1/74.9	73.3/72.2
	Voxel Mamba [47]	-/79.6	-/73.6	80.8/80.3	72.6/72.2	85.0/80.8	77.7/73.6	78.6/77.6	75.7/74.8
	UniMamba [21]	-/-	76.1/74.1	80.6/80.1	72.3/71.8	86.0/81.3	78.7/74.1	80.3/79.3	77.5/76.5
Stream	PolarStream [6]*	-/-	-/60.8	72.4/71.8	64.6/64.0	-/-	-/-	-/-	-/-
	FPA-3DOD [4]	-/-	-/-	76.3/-	69.8/-	72.7/-	70.1/-	-/-	-/-
	PARTNER [31]	-/-	-/63.2	76.1/75.5	68.6/68.1	-/-	-/-	-/-	-/-
	PFCF (ours)	78.5/76.6	72.1/70.3	79.2/78.6	71.0/70.5	80.7/76.6	72.7/68.8	75.5/74.4	72.7/71.7

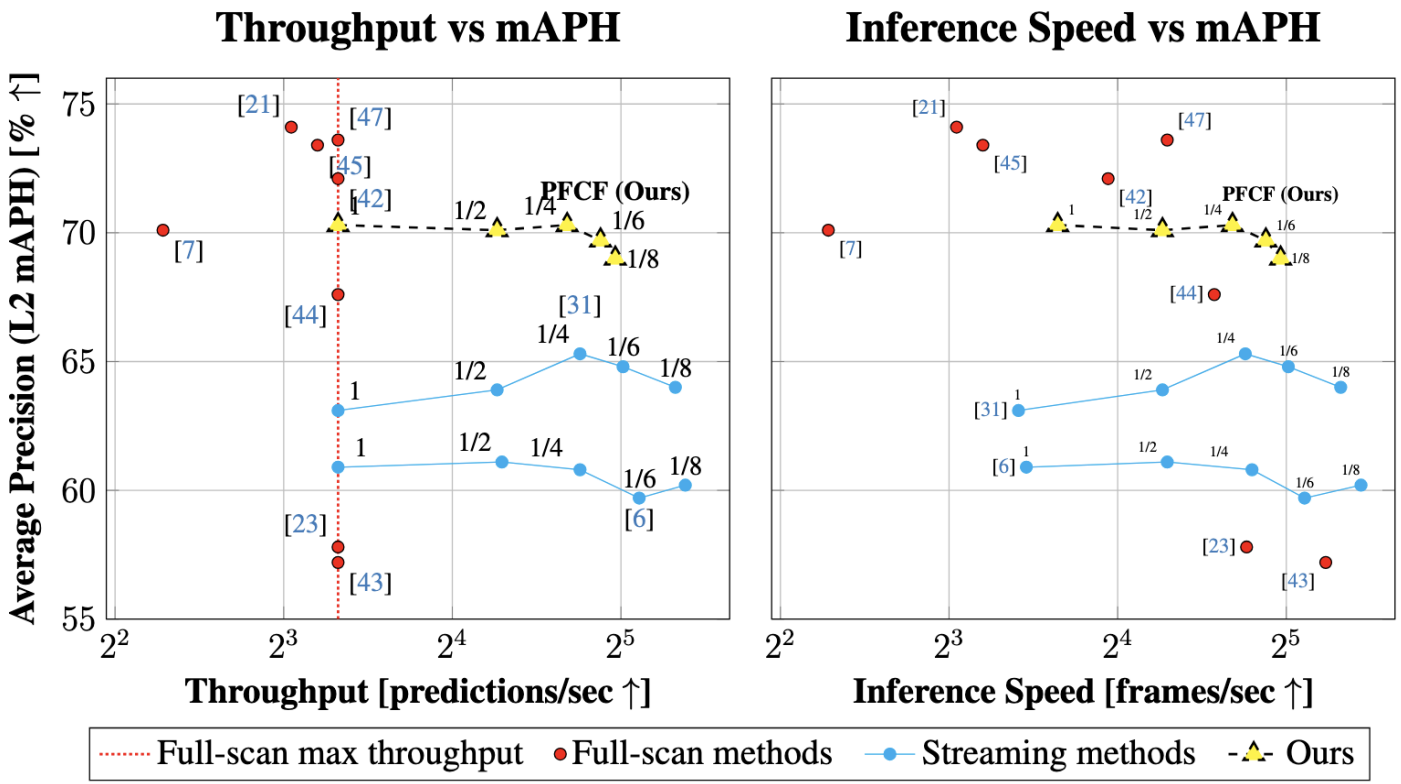
Table 1. **Full scan results on Waymo Open validation set.** Metrics: mAP/mAPH(%)↑ for overall results, AP/APH (%)↑ for each category. * denotes reimplementation results. mAP denotes mean average precision and mAPH denotes mean average precision with heading. AP and APH are per-category average precision and average precision with heading.

Method	mAP/mAPH		Vehicle AP/APH		Pedestrian AP/APH		Cyclist AP/APH	
	L1	L2	L1	L2	L1	L2	L1	L2
Voxel Mamba 1/4 [47]	24.6/23.0	22.2/20.8	24.9/24.6	21.6/21.3	23.7/20.3	20.4/17.4	25.2/24.3	24.6/23.7
PHiM 1/4 (ours)	40.5/36.5	36.4/32.8	46.2/45.5	39.9/39.3	35.1/25.9	30.1/22.2	40.2/38.0	39.2/37.1

Method	Rep.	Number of streaming sectors				
		1	2	4	6	8
STROBE [12]	Cartesian	60.5/59.8	59.5/58.9	58.8/58.3	58.3/57.6	58.0/57.3
Han [17]	Cartesian	61.8/61.4	61.7/61.1	60.7/60.2	60.0/59.3	59.9/59.3
PolarStream [6]	Polar	61.4/60.8	61.8/61.2	61.2/60.7	60.3/59.7	60.7/60.2
PARTNER [31]	Polar	63.8/63.2	64.3/63.8	66.0/65.5	65.3/64.7	64.5/64.0
PFCF (Ours)	Hybrid	**72.1/70.3**	**71.2/70.1**	**70.6/70.3**	**70.1/69.7**	**69.5/69.0**

Streaming	DDC	PHiM	SFB	BSSM	L1 mAP
Incompatible	✗	✗	✗	✓	58.02
Incompatible	✓	✗	✗	✓	62.61
Incompatible	✓	✗	✓	✓	67.45
Compatible	✗	✓	✗	✓	66.31
Compatible	✗	✓	✓	✓	69.71
Compatible	✓	✓	✗	✓	69.24
Compatible	✓	✓	✓	✗	74.92
Compatible	✓	✓	✓	✓	**78.50**

Throughput-Accuracy Pareto Frontier



Prediction Refinement Over Time