Sensing-Aided and Neural Network-Based mmWave Beam Management

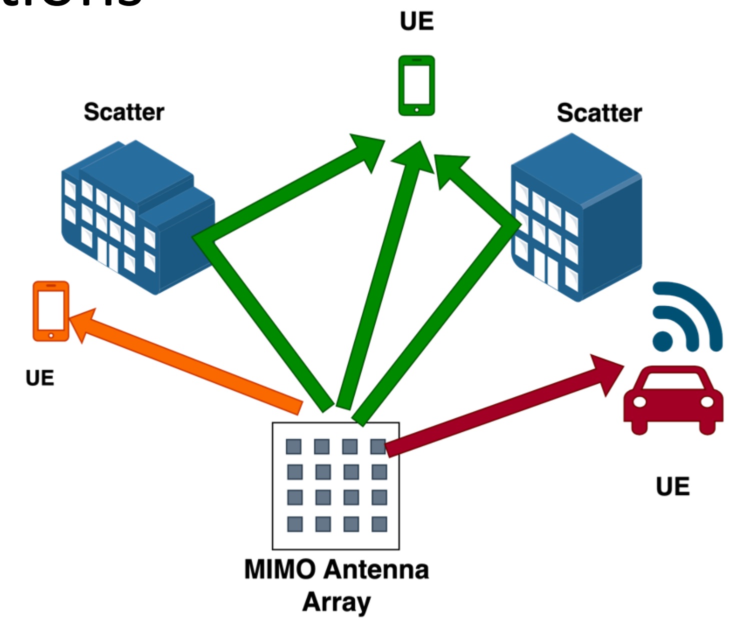
Kyle Guan

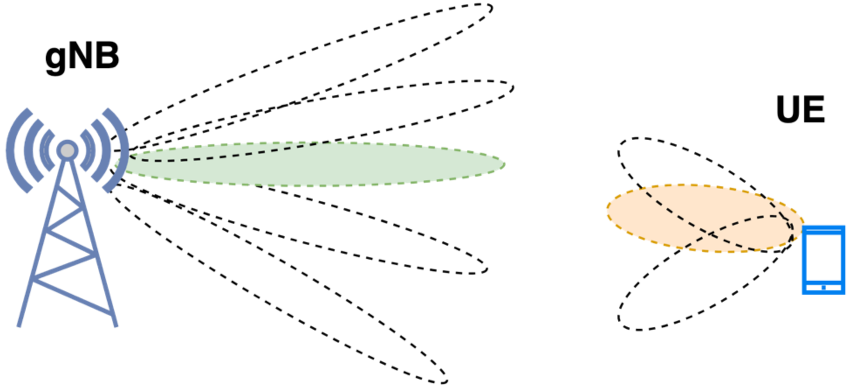
Outline

- Background
  - Beam management in 5G new radio (NR)
  - Sensing information from advance driver assistance system (ADAS) and roadside unit (RSU) sensors
- Problem statement: using sensing information to reduce the beam search overhead
- Neural network (NN) based approaches:
  - Sensing data collection and processing
  - NN model selection, training, and evaluation
    - Centralized training
    - Distributed training: federated learning
    - The trade-off between performance and bandwidth
- Summary and literature survey

Beam Management in 5G Wireless Communications

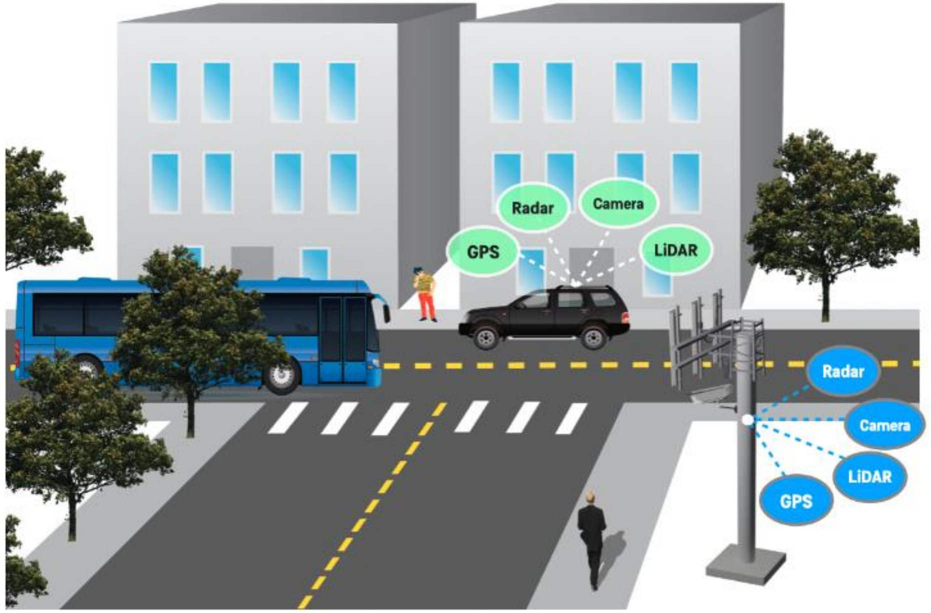
- 5G relies on millimeter wave (mmWave) to achieve high capacity
- Beam management techniques are used to mitigate the high pathloss and blockage
    - Goal: establish and maintain an optimal transmit-receive beam pair to achieve good connectivity
- Beam selection procedure between a user equipment (UE) and access network node (gNB):
    - Beam sweeping: exhaustively search over all the beams on the transmitter and the receiver sides,
    - Beam selection: choose the beam pair that offers the strongest reference signal received power (RSRP).
- Issues with beam search: exhaustive searches over all beams are computationally expensive and increases the initial access time (overhead).





Sensing Information

- The positions of gNBs and UEs, as well as environment factors (scatters and blockers) impact beam propagation
- Out-of-band data from ADAS and RSU sensors can provide understanding and awareness of environment
  - Sensor type: camera, radar, LiDAR, GPS, IMU, etc.
  - Data modality: image, point cloud, etc.
  - With NN models (e.g., object detections), contextual information of the environment can be extracted
    - Examples: UE locations and orientations, mobility pattern, and potential blockage
- The extracted information can be leveraged to improve the efficiency of the beam management:
  - Beam search
  - Beam blockage prediction
  - Proactive hand-off



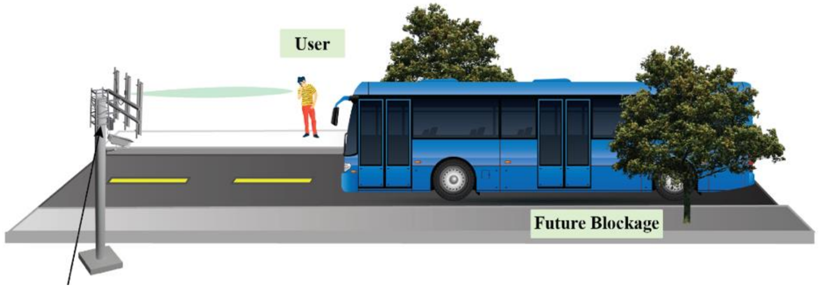
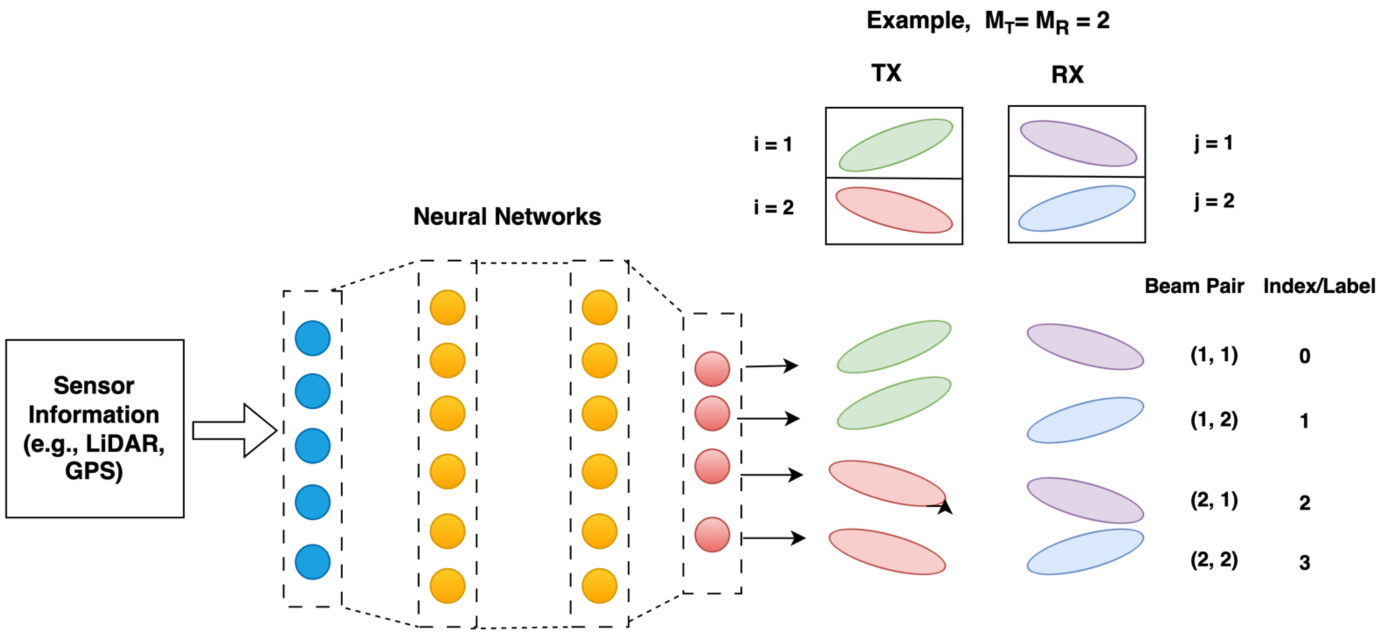
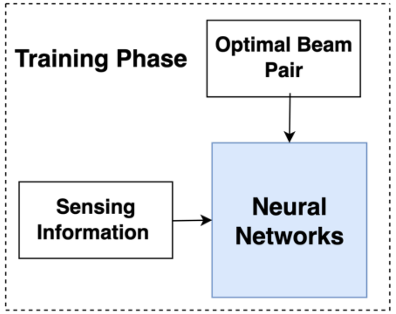


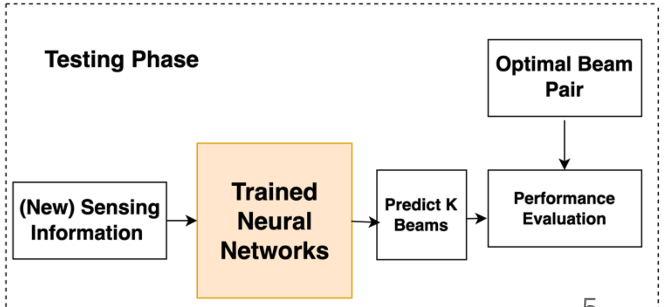
Image source: https://www.deepsense6g.net/

Problem Statement

- Issues: exhaustive searches over all beam pairs are computationally expensive
- Goal: leverage LiDAR and location information to reduce the overhead in beam search
- Approach: train a supervised NN model to recommend a set of K good beam pairs
  - Training samples: features extracted from LiDAR sensing
  - Training labels: true optimal beam pair index
  - Evaluation metric: top-N accuracy

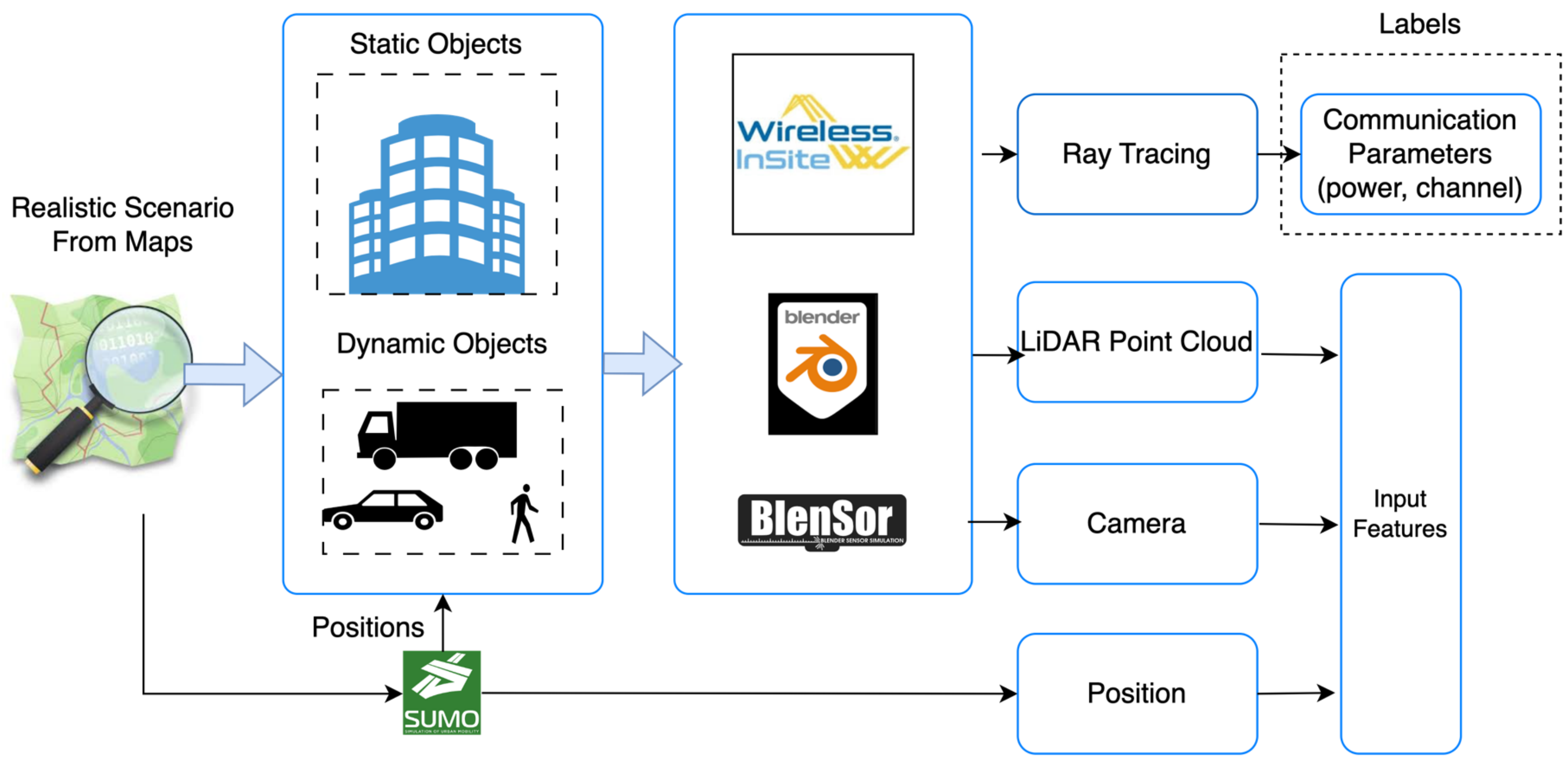






Training and Testing Data

- Use a suite of high-fidelity simulation tools to generate realistic sensing and wireless data



LiDAR and Beam Data Processing

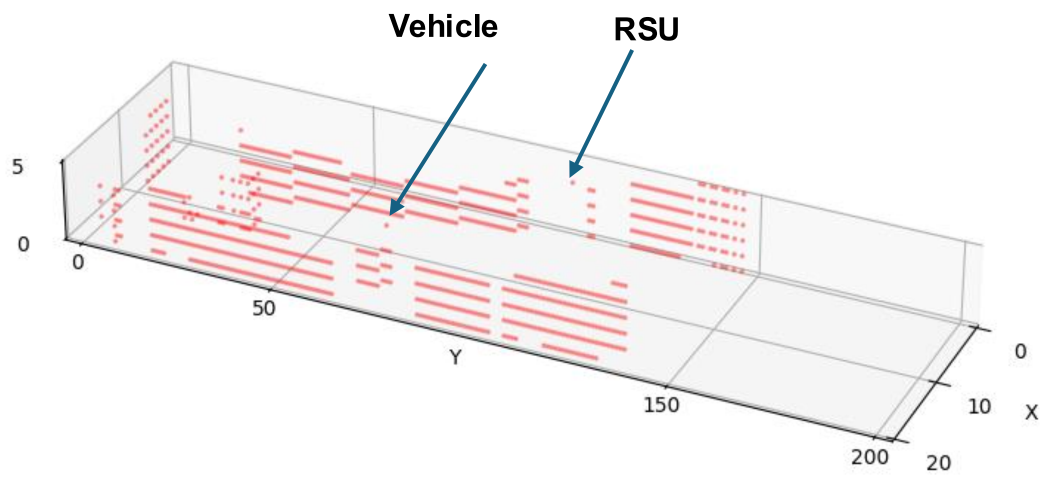
- Convert 3D point cloud into 2D feature representation
  - A 20 ╳ 200 matrix: each element is assigned a value from [0, 1, -1, -2]

Type	Value
Unoccupied	0
Vehicle (ego)	-2
RSU	-1
Other objects (obstacles)	1

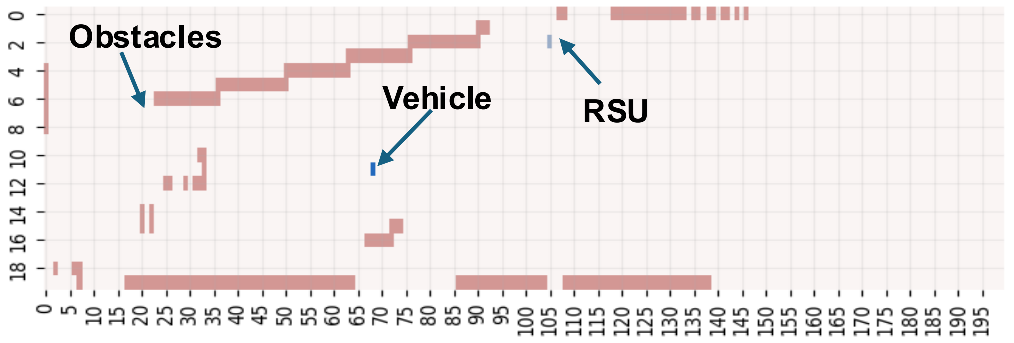
  - ~ 25 ╳ reduction in feature size compared to point cloud

Data Modality	Size (No. of Data Points)
Point clouds	~10^5
3D Grid	20x200x10
2D Grid	20x200

3D-Grid



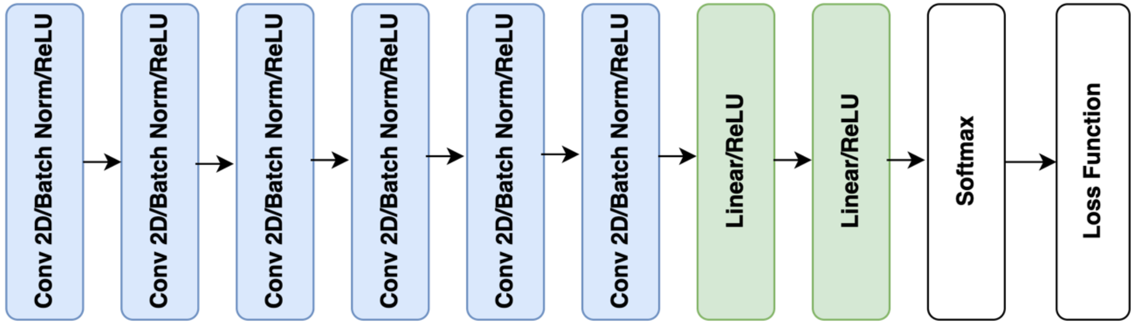
2D-Grid



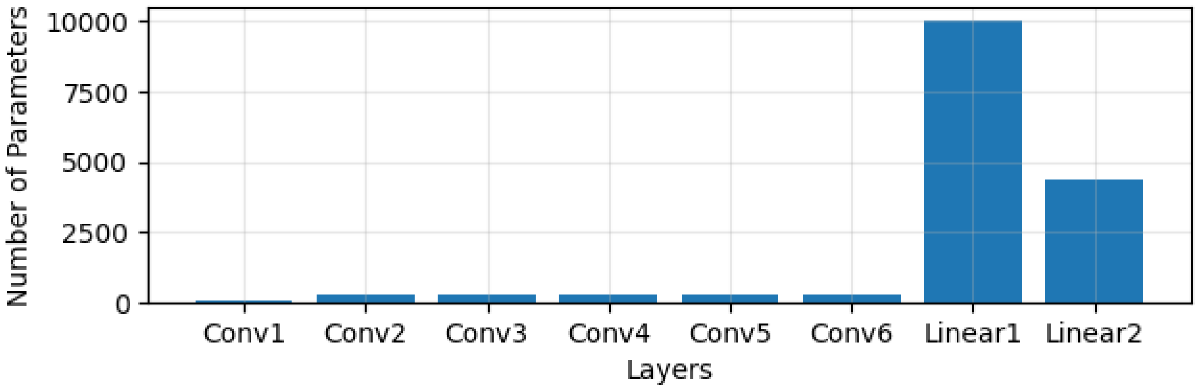
Model Selection: Baseline Model

- CNN-based baseline model
  - 6 convolutional layers with batch normalizations and ReLU activations
  - 2 linear layers
  - Softmax layer provides probability for each beam pair
- Linear layers accounts for ~90% of the total parameters

Baseline Model Architecture



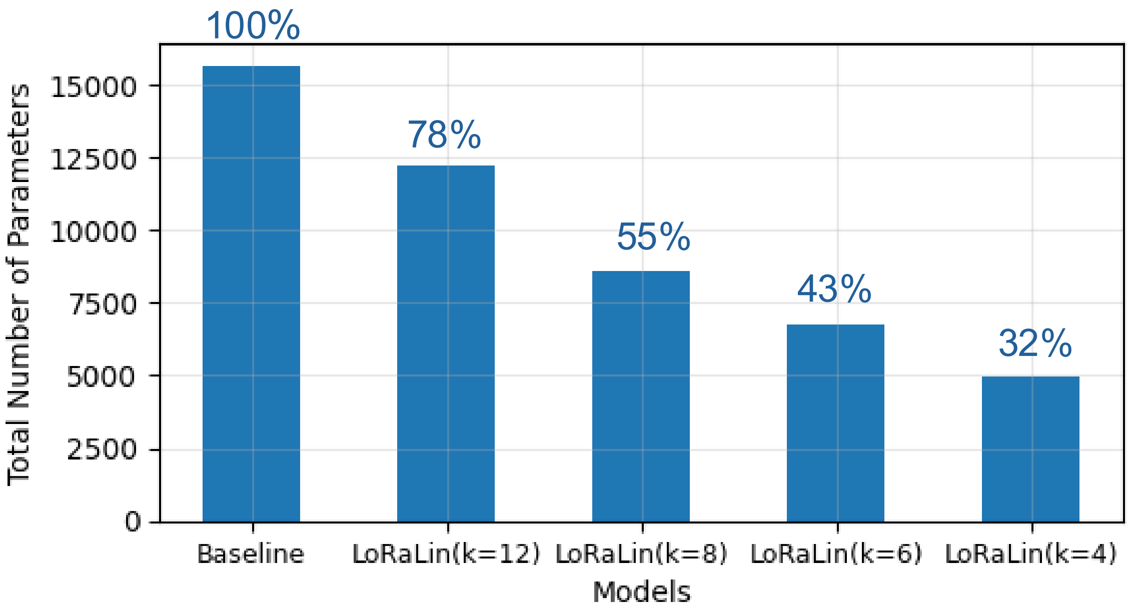
Number of Parameters

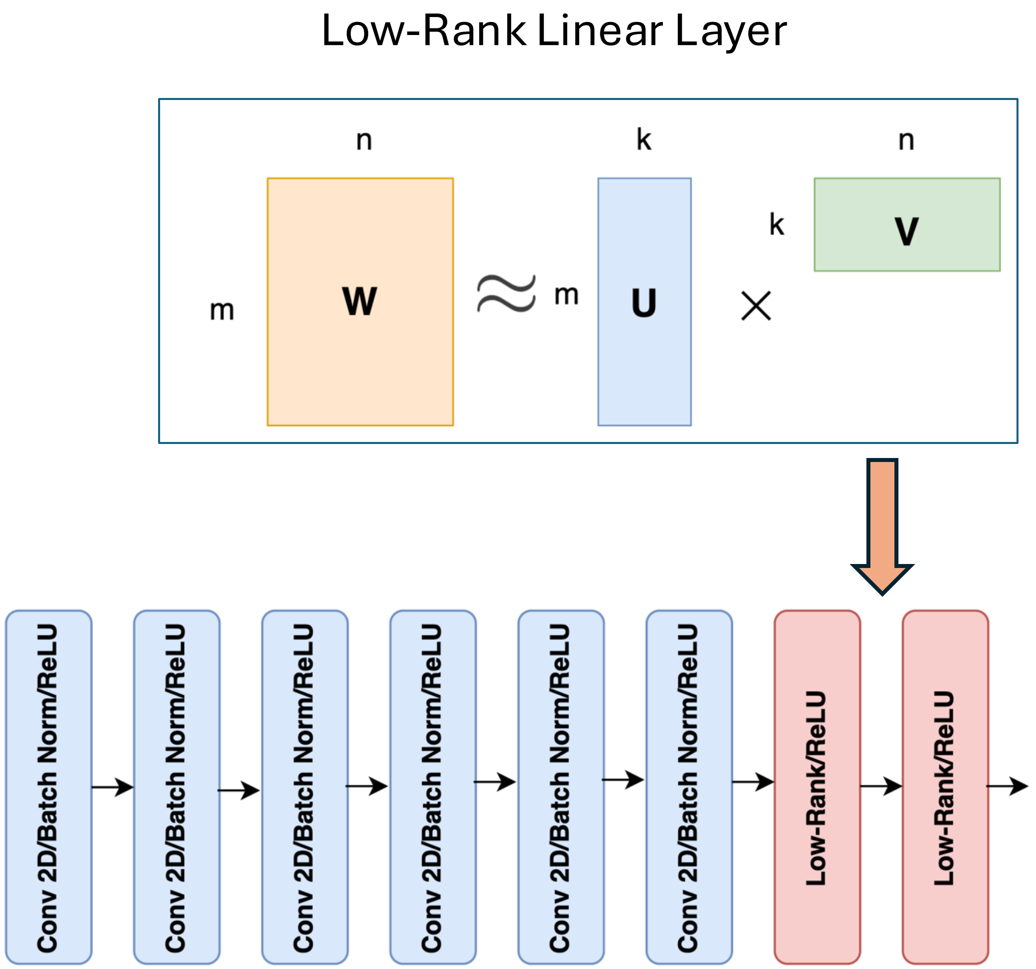


Model Selection: Low-Rank Linear Layers

- Low-Rank Linear Layer:
  - Use the product of two lower-rank matrices U and V to approximate matrix W
  - Reduce the number of trainable parameters

$$\boxed{\eta = \frac{(m+n)k}{mn}}$$

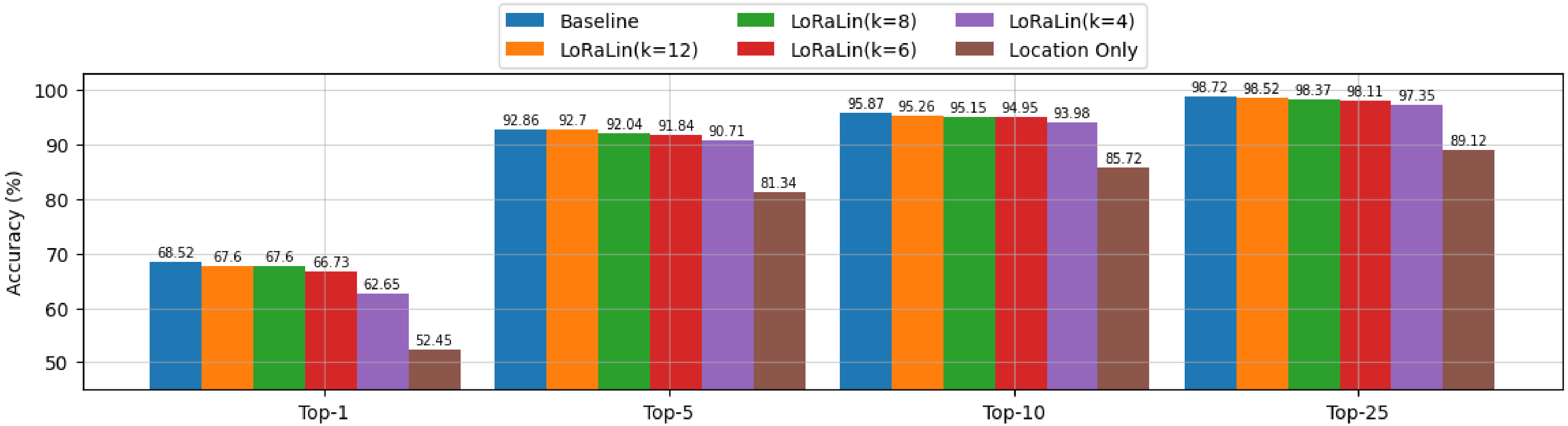




Centralized Training Results (1)

- Centralized data collection and centralized training
- Evaluations: top-K accuracy
  - The optimal beam pair is in the top-K most probable beam pairs

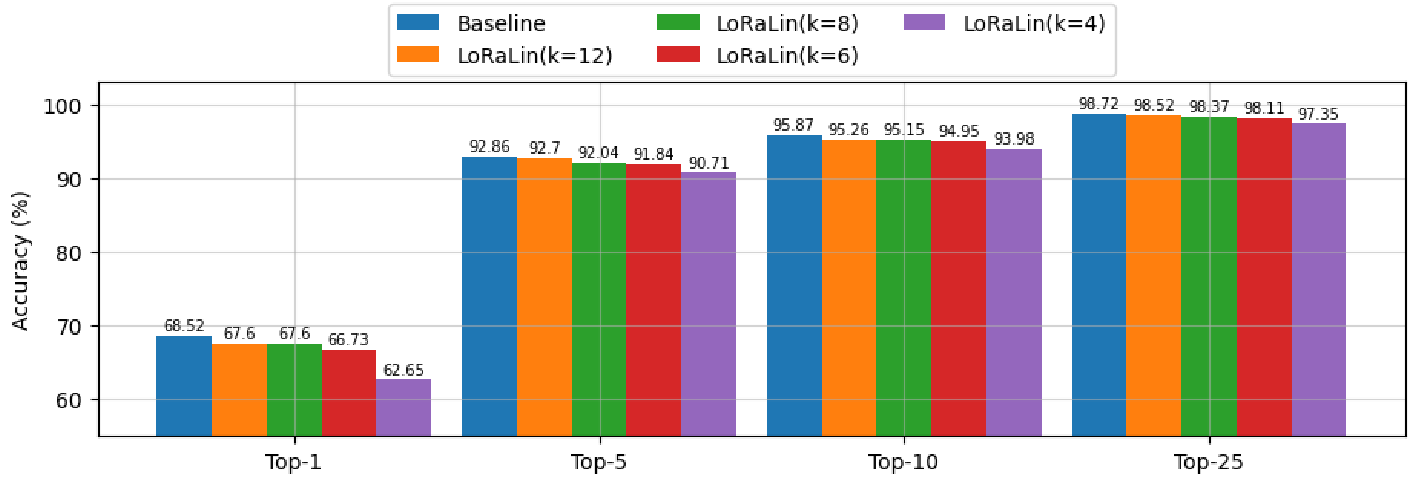
Models	Top-1 Acc.	Top-5 Acc.	Top-10 Acc.	Top-25 Acc.
Baseline	68.52%	92.86%	95.87%	98.72%
LoRaLin (k = 12)	67.6%	92.7%	95.26%	98.52%
LoRaLin (k = 8)	67.6%	92.04%	95.15%	98.37%
LoRaLin (k = 6)	66.73%	91.84%	94.95%	98.11%
LoRalin (k = 4)	62.65%	90.71%	93.98%	97.35%
Location Only	52.45%	81.34%	85.72%	89.12%



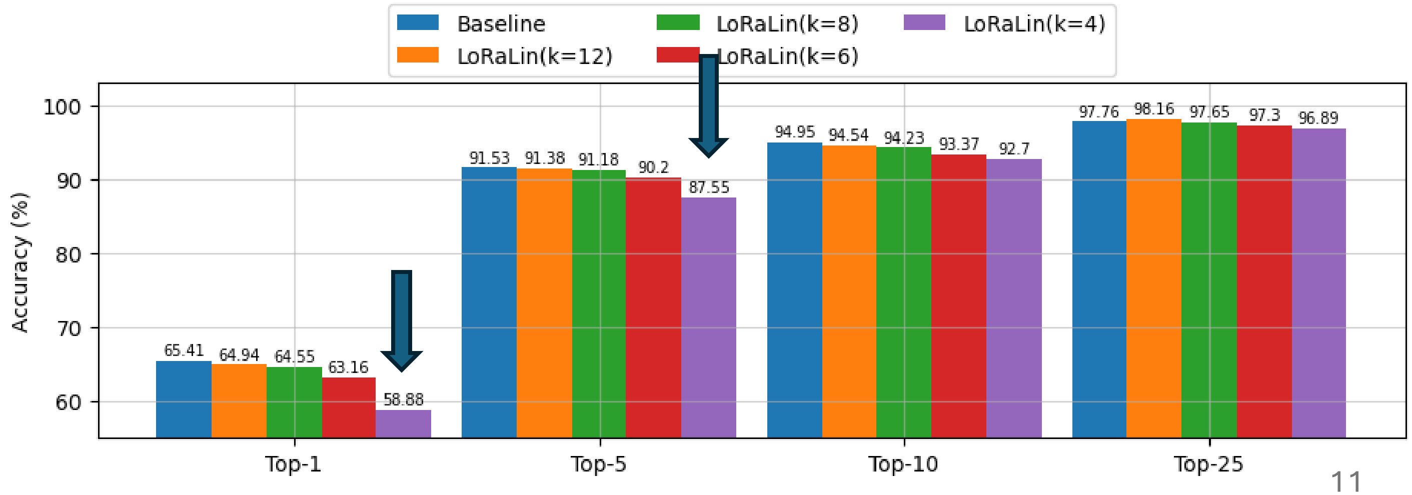
Centralized Training Results (2)

- Evaluate the robustness of models with less training data
  - Train all the model with half of the training size
  - Performance degradation for low-rank models with smaller k

Full Training Dataset



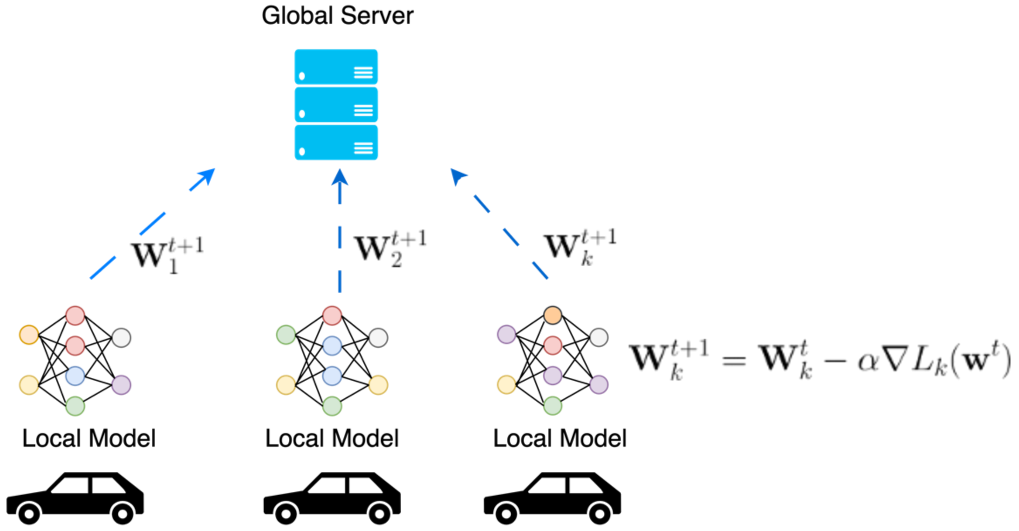
Half Training Dataset



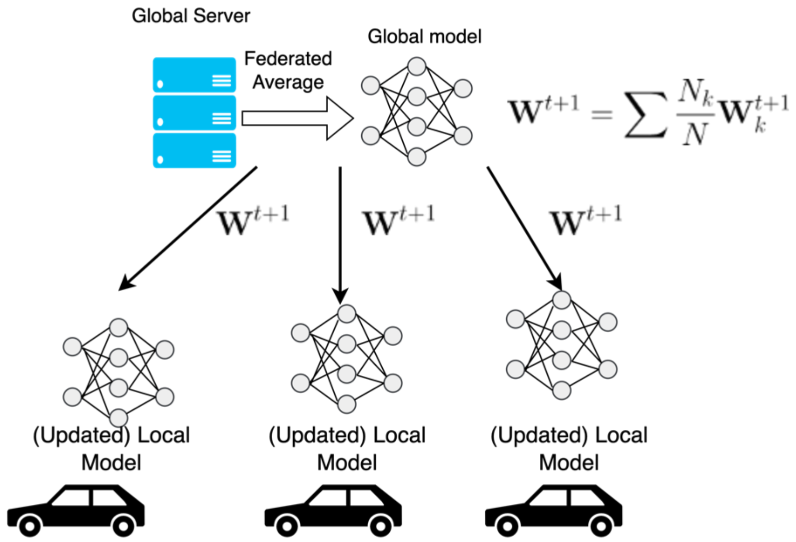
Distributed Training: Federated Learning Framework

- Federated learning framework:
  - Global model initialization
  - Local update: $\mathbf{W}_k^{t+1} = \mathbf{W}_k^t - \alpha \nabla L_k(\mathbf{w}^t)$
  - Global aggregation: $\mathbf{W}^{t+1} = \sum \frac{N_k}{N} \mathbf{W}_k^{t+1}$
  - Training stop criterion: # aggregation rounds or metrics achieved

Local Update

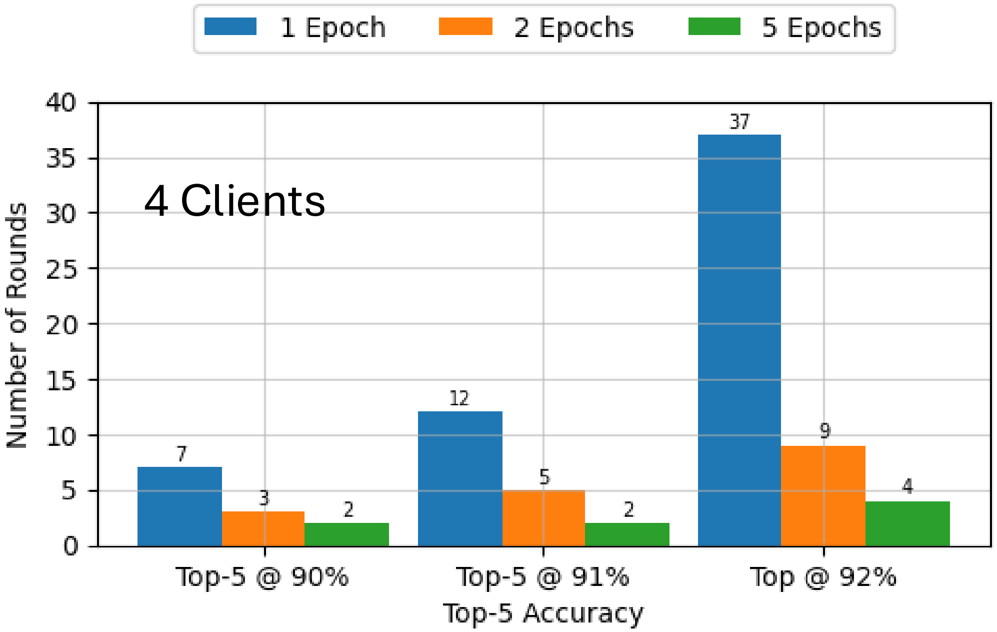


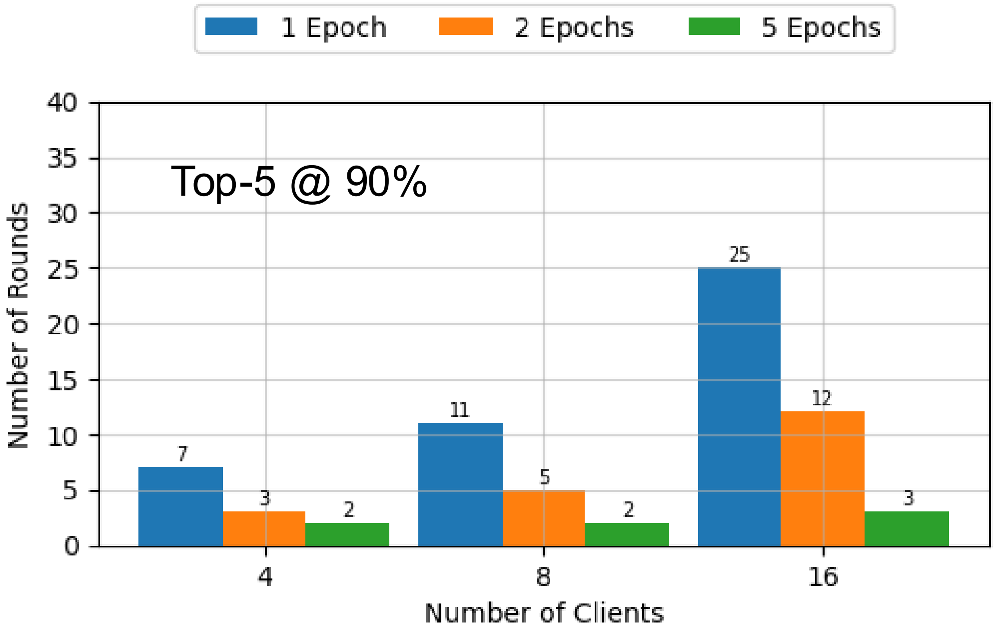
Global Aggregation



Federated Learning: Performance Evaluation (1)

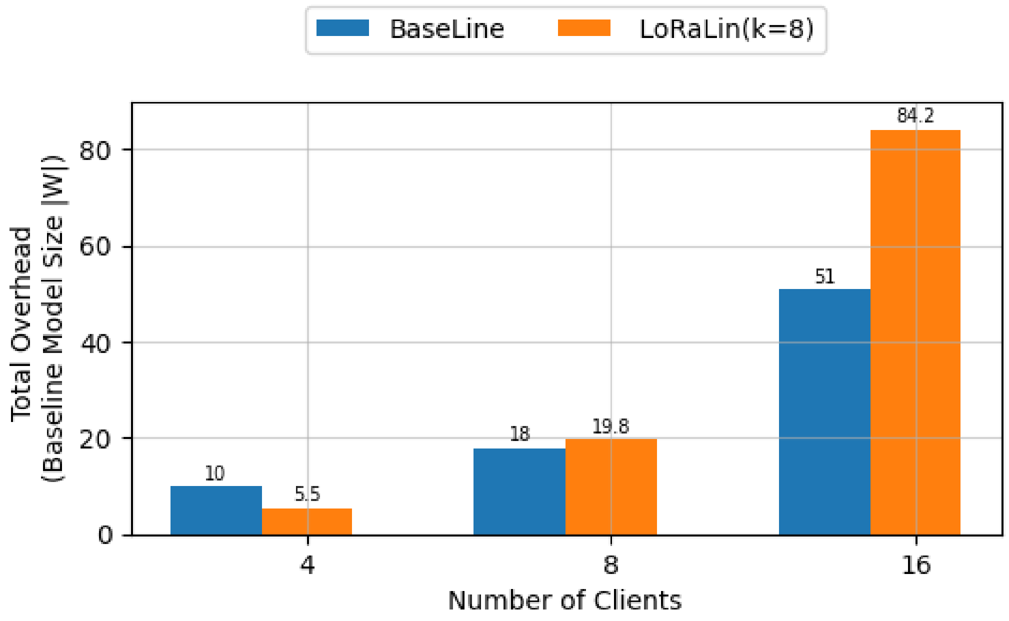
- The performance depends on many training parameters:
  - Local training epochs (between global aggregation)
  - Number of participants (clients)
  - Local models
  - Others: e.g., type of federated averaging
- Evaluation metrics: number of rounds of aggregation to reach a Top-k accuracy





Federated Learning: Performance Evaluation (2)

- The performance depends on many training parameters:
  - Number of participants (clients)
  - Local training epochs (between global aggregation)
  - Local models
  - Others: e.g., type of federated averaging
- Evaluation metrics: number of rounds of aggregation



Baseline Model, epoch = 5

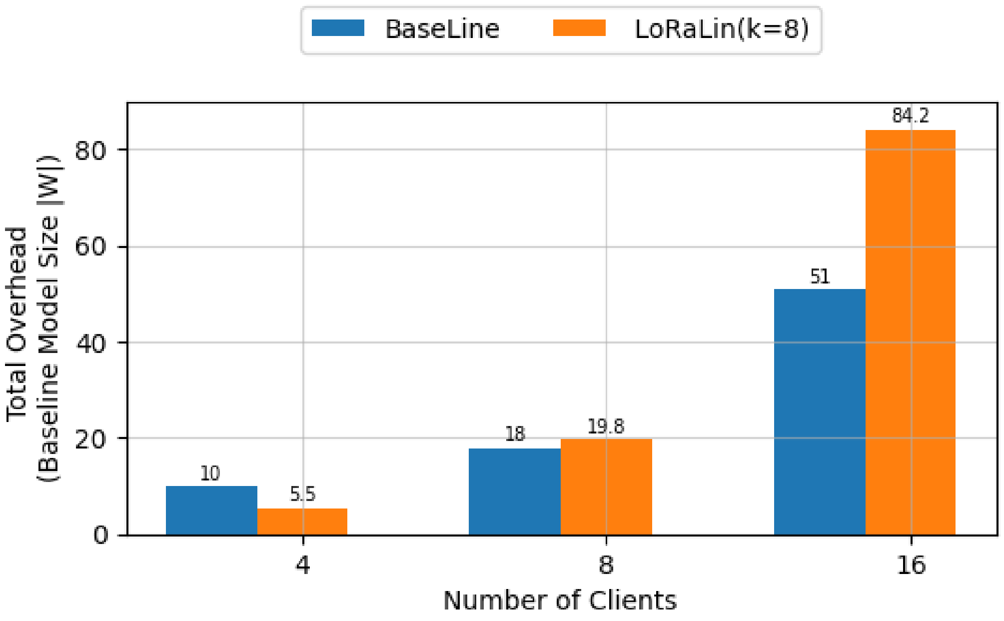
No. of Clients	No. Rounds For Top-5 %90	Uplink Overhead	Downlink Overhead	Total Overhead
4	2	8|W|	2|W|	10|W|
8	2	16|W|	2|W|	18|W|
6	3	48|W|	3|W|	51|W|

Low Rank (k=8) Model, epoch = 5

No. of Clients	No. Rounds For Top-5 %90	Uplink Overhead	Downlink Overhead	Total Overhead
4	2	4.4|W|	1.1|W|	5.5|W|
8	4	17.6|W|	2.2|W|	19.8|W|
6	9	79.2|W|	5.0|W|	84.2|W|

- |W|: baseline model size

- The performance depends on many training parameters:
    - Number of participants (clients)
    - Local training epochs (between global aggregation)
    - Local models
    - Others: e.g., type of federated averaging
- Evaluation metrics: number of rounds of aggregation



Baseline Model, epoch = 5

No. of Clients	No. Rounds For Top-5 %90	Uplink Overhead	Downlink Overhead	Total Overhead
4	2	8|W|	2|W|	10|W|
8	2	16|W|	2|W|	18|W|
6	3	48|W|	3|W|	51|W|

Low Rank (k=8) Model, epoch = 5

No. of Clients	No. Rounds For Top-5 %90	Uplink Overhead	Downlink Overhead	Total Overhead
4	2	4.4|W|	1.1|W|	5.5|W|
8	4	17.6|W|	2.2|W|	19.8|W|
6	9	79.2|W|	5.0|W|	84.2|W|

- |W|: baseline model size

Backups

TABLE I: Comparison between the proposed NN architecture and the baseline in [12], [13], both trained in a centralized manner.

Model	Top-10 accuracy	Top-10 throughput ratio	FLOPs	# of NN parameters, $\lvert\boldsymbol{\theta}\rvert$
Proposed centralized	$91.17 \pm 0.28\%$	$94.78 \pm 0.61\%$	1.72×10^6	7462
Baseline [12], [13]	$83.92 \pm 0.93\%$	$86.15 \pm 0.82\%$	179.01×10^6	403677

TABLE II: Performance of federated beam selection when initialized from an untrained global model.

V	N_v	$(N_a)^{0.88}$	$(O_{DL})^{0.88}$	$(O_{UL})^{0.88}$	Top-10 Acc.
5	1	19	$19\lvert\theta\rvert$	$95\lvert\theta\rvert$	90.12%
	2	13	$13\lvert\theta\rvert$	$65\lvert\theta\rvert$	90.34%
	5	10	$10\lvert\theta\rvert$	$50\lvert\theta\rvert$	89.92%
10	1	31	$31\lvert\theta\rvert$	$310\lvert\theta\rvert$	89.77%
	2	22	$22\lvert\theta\rvert$	$220\lvert\theta\rvert$	89.16%
	5	15	$15\lvert\theta\rvert$	$150\lvert\theta\rvert$	88.64%
20	1	81	$81\lvert\theta\rvert$	$1620\lvert\theta\rvert$	88.81%
	2	48	$48\lvert\theta\rvert$	$960\lvert\theta\rvert$	88.53%
	5	NA	NA	NA	87.33%

TABLE III: Performance of federated beam selection when initialized from an offline-trained model.

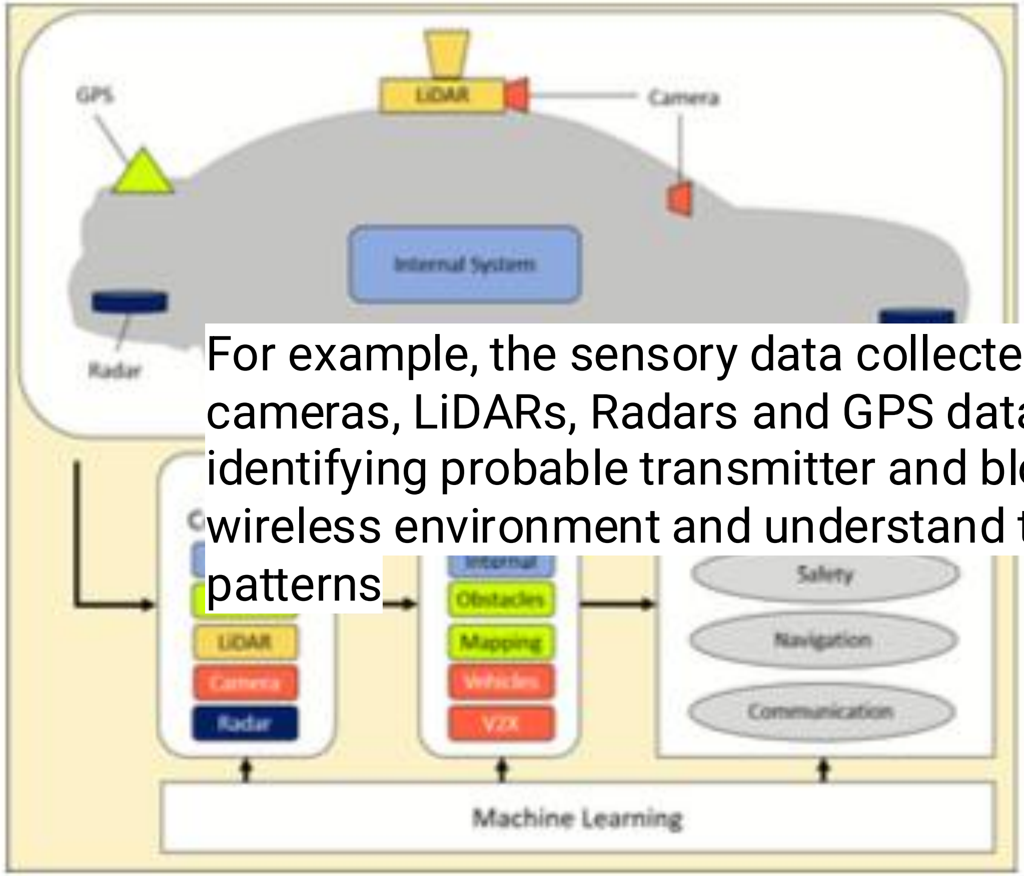
V	N_v	2K from s008 (Rosslyn)		2K from s007 (Beijing)	
		$(N_a)^{0.88}$	Top-10 Acc.	$(N_a)^{0.88}$	Top-10 Acc.
10	1	17	90.29%	23	89.35%
	2	10	89.86%	14	88.73%
	5	6	89.31%	9	88.19%

Summary and Related Research

- Background on 5G NR beam management

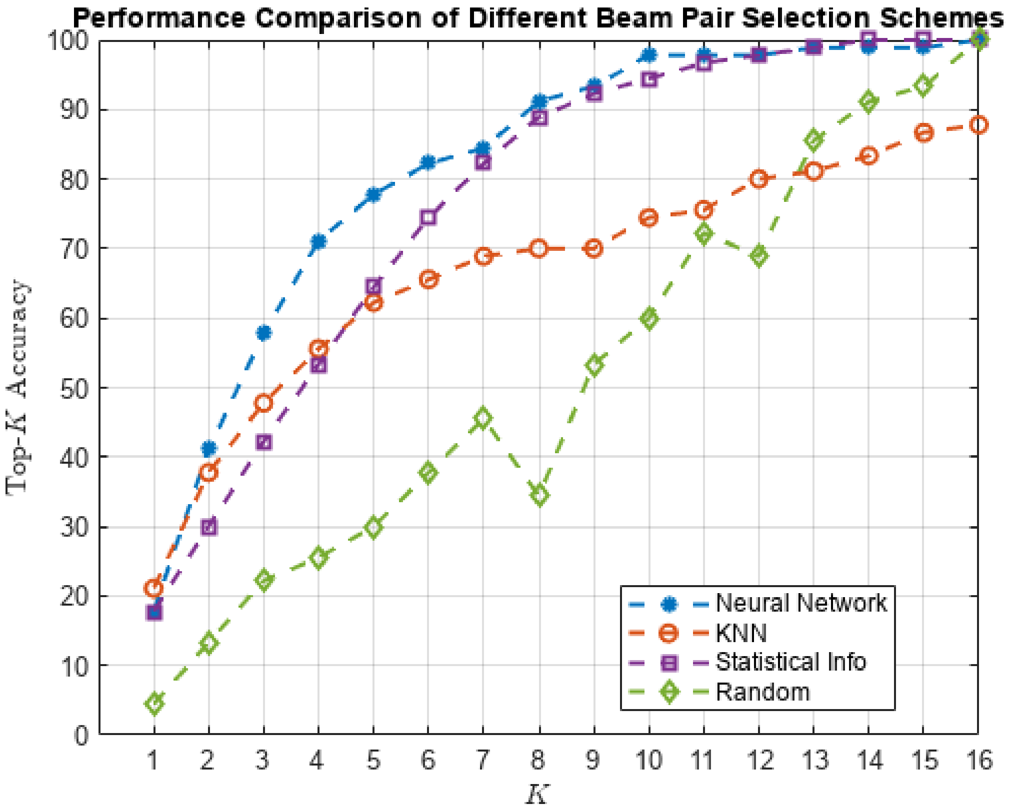
Federated Learning: Performance and Overhead Trade-off

Num of Participants	Local Epochs	Rounds for Top-5 Acc. at 90%	Rounds for Top-5 Acc. at 91%	Rounds for Top-5 Acc. at 92%	Downlink Bandwidth	Uplink Bandwidth	Total Bandwidth
4	1	7	14	37			
	2	3	5	9			
	5	2	2	4			
8	1	11	18	40			
	2	5	11	20			
	5	2	3	6			
16	1	25	50	50			
	2	12	40	40			
	5	3	6	30			



For example, the sensory data collected from RGB cameras, LiDARs, Radars and GPS data can help in identifying probable transmitter and blockages in the wireless environment and understand their mobility patterns

Fig. 6. Diagram depicting all aspects of the data fusion process in AVs, including sensors employed by the vehicle. Note that machine learning is used at each step of the data fusion process.



- The LIDAR scans were performed with the following configurations: • Model: Velodyne HDL-64E2 • Scan distance(m): 120 • Scan angle resolution: 0.17 • Viewing angle: 360 Quantization is done to convert the results into a format

https://www.mathworks.com/help/5g/ug/neural-network-for-beam-selection.html


Analog beamforming
Baseband
DAC
RF chain
Wireless channel
Analog combining
RF chain
ADC
Baseband

\mathbf{W}_k^{t+1} = \mathbf{W}_k^{t} - \alpha\nabla L_k(\mathbf{w}^t)

\mathbf{W}^{t+1} = \sum \frac{N_k}{N}\mathbf{W}_k^{t+1}

$$\mathbf{W}_k^{t+1}$$

$$\mathbf{W}_1^{t+1} \qquad \mathbf{W}_2^{t+1}$$

$$\mathbf{W}^{t+1}$$

Num of Participants	Local Epochs	Rounds for Top-5 Acc. at 90%	Rounds for Top-5 Acc. at 91%	Rounds for Top-5 Acc. at 92%	Downlink Bandwidth	Uplink Bandwidth	Total Bandwidth
4	1	7	14	37			
	2	3	5	9			
	5	2	2	4			
8	1	11	18	40			
	2	5	11	20			
	5	2	3	6			
16	1	25	50	50			
	2	12	40	40			
	5	3	6	30			