FedAttention: Federated Attention-Based Fusion Learning for Multi-Modal Beamforming in IoV

Jinxuan Chen*, Eric Samikwa*, Torsten Braun*, Kaushik Chowdhury†
*Institute of Computer Science, University of Bern, Switzerland
†Department of Electrical and Computer Engineering, University of Texas at Austin, USA
Email: *{jinxuan.chen, eric.samikwa, torsten.braun}@unibe.ch, † {kaushik}@utexas.edu

ICC 2025 - IEEE International Conference on Communications | 979-8-3315-0521-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICC52391.2025.11161204

***Abstract*—Advanced beamforming techniques enable stable vehicular communication and address mmWave limitations by accurately directing the signal. However, traditional beamforming techniques struggle in high-speed vehicles due to time-intensive codebook processing and image-based feedback adjustments. Multi-modal beamforming using real-time data like GPS, cameras, and LiDAR to train the Deep Learning (DL) models can provide adaptive beam steering, improving reliability in dynamic conditions. Despite this, centralized systems involving large raw data transmission are vulnerable to saturation and malicious interference, and they neglect privacy concerns, necessitating a new framework. This paper proposes a novel federated attention-based fusion learning framework named FedAttention for multi-modal beamforming in the Internet-of-Vehicle (IoV). FedAttention further improves the model generalization ability by utilizing the CNN-Transformer architecture and making full use of the Multi-access Edge Computing (MEC) servers for the potential federated split learning to enhance efficiency. Based on the real-world datasets, FedAttention achieves 98.16% in Top-5 accuracy and 82.09% in Top-1 accuracy, a 26.86% improvement compared to the current FLASH framework with less wall clock time, showing its training efficiency and robustness.**

***Index Terms*—federated learning, multi-modal beamforming, fusion learning, transformer, attention**

I. Introduction

In 6G Vehicle-to-everything (V2X) communication, Integrated Sensing and Communication (ISAC) is a pillar technology for improving environmental awareness by collecting data from multiple sensors such as radar, LiDAR, cameras, and GPS, leading to safer and more efficient autonomous driving [1]. Multi-access Edge Computing (MEC) further enhances V2X communication by putting the computational resources and data processing closer to vehicles and infrastructure, thereby substantially pruning the E2E communication latency and supporting real-time decision-making [2]. Meanwhile, mmWave frequencies offer extensive bandwidth for rapid data transmission and high-resolution sensing capabilities in the Internet-of-Vehicle (IoV), which helps vehicles to achieve faster and more efficient communication for applications such as cooperative perception [3].

mmWave signals are vulnerable to blockages from obstacles like buildings and trees, limiting their effectiveness in complex environments [4]. Advanced beamforming techniques can improve signal strength, coverage, and reliability by precisely adjusting beams in specific areas. However, traditional beamforming struggles in high-speed vehicle environments, as it relies on time-intensive codebook processing and image-based feedback adjustments, which struggle to maintain stable transmission amid rapid movement and uneven roads [5]. Moreover, these methods demand increasingly sophisticated mathematical solutions, reducing their efficiency in dynamic V2X scenarios.

In contrast, sensor-based beamforming utilizes real-time environmental data for Deep Learning (DL) models to enable more precise and adaptive beam steering, significantly improving efficiency and reliability in dynamic conditions. For instance, Zheng *et al.* [6] proposed a 3D Convolutional Neural Network (CNN) for optimizing power dissipation and beam selection in V2X scenarios, leveraging 5G mmWave communication and LiDAR sensor data for enhanced performance. Reus *et al.* [7] proposed a DL-based data fusion approach using visual edge devices and localization sensors to reduce beam selection overhead and detect blockages. Xu *et al.* [8] proposed 3D object detection to extract vehicle data and design a DNN for optimal beam pair inference without pilot signals, along with a vision-based BCT prediction method to improve transmission rates. Yang *et al.* [9] developed a mmWave network architecture that uses street camera images and user identification to predict the optimal beam index and blockage state without pilot training, achieving ultra-reliable low-latency communication (URLLC).

Nearly all the studies focused on either single-sensor-based beamforming or deep centralized learning-based approaches. However, relying on a single sensor-assisted model can be compromised by external factors such as adverse weather conditions like heavy rain and complex road surfaces, which can degrade the quality of sensing data, ultimately reducing beamforming performance and reliability [10]. Moreover, many of the discussed techniques rely on centralized systems, which introduce challenges such as high bandwidth demands for raw data transmission, leaving them vulnerable to saturation and malicious interference. Additionally, these approaches often neglect crucial data privacy concerns, posing significant risks in sensitive or large-scale deployments.

Federated Learning (FL) [11] is an emerging distributed method that enables multiple clients to train models collaboratively without sharing raw data, thereby preserving data privacy and preventing the overloading of control channels. FLASH [12] explores the application of FL by employing CNN for local model training on devices, followed by global

aggregation of model weights. However, FLASH [12] suffers from limitations such as very low accuracy for the best beam sector prediction (even for the vast number of FL rounds). It also exhibits reduced computational efficiency by leveraging FL and very deep learning architecture, hindering the overall effectiveness of FL in real-world V2X scenarios. Therefore, our work aims to determine how to enhance the best beam sector prediction accuracy in FL while maintaining training efficiency.

In this paper, we propose `FedAttention`, a novel federated attention-based fusion learning framework for multi-modal beamforming in IoV, as illustrated in Fig. 1. Our main contributions are as follows:

- FedAttention leverages the combined strengths of CNN and transformer architectures to effectively capture both local and global features. This hybrid design significantly improves model accuracy and training efficiency, addressing the shortcomings of current state-of-the-art FL-based multi-modal beamforming methods.
- By making full use of the joint computing resources of the MEC servers and the local vehicles, we introduce an *offloading indicator* that enables federated split learning (FSL) for vehicles with limited capacity to perform local model training, thereby enhancing training efficiency.
- We evaluate FedAttention's superior model generalization and adaptability against existing FL-based multi-modal frameworks, considering the heterogeneity of the dataset. The results demonstrate that FedAttention significantly improves accuracy, exceeding the upper limit of the FLASH. To achieve the same global model accuracy as FLASH, our framework requires fewer computational resources and available data samples, as well as reduced wall clock time, showing its enhanced training efficiency.

The remainder of this paper is organized as follows. Section II introduces the design of the federated attention-base fusion learning framework for multi-modal beamforming. Section III evaluates the performance of FedAttention based on real-world datasets with different variables. Finally, conclusions are drawn in Section IV.

II. FEDERATED ATTENTION-BASED LEARNING FOR MULTI-MODAL BEAMFORMING

A. System Model

Consider a set of autonomous vehicles Φ in the coverage of a gNodeB that collaborates with MEC servers for training the FL global model. Each vehicle $\phi \in \Phi$ is equipped with multi-modal sensors GPS, camera, and LiDAR, which can be used to collect the multi-modal local dataset $D_\phi = \{X_{\phi,C}, X_{\phi,I}, X_{\phi,\zeta}\}$ of size $|D_\phi|$. The parameters $X_{\phi,C} \in \mathbb{R}^{D_\phi \times 2}$, $X_{\phi,I} \in \mathbb{R}^{D_\phi \times d_0^I \times d_1^I}$, and $X_{\phi,\zeta} \in \mathbb{R}^{D_\phi \times d_0^\zeta \times d_1^\zeta \times d_2^\zeta}$ represents the GPS, camera, and LiDAR data, respectively. GPS data can be formulated as a simple vector with two elements representing latitude and longitude, the image data forms a matrix with dimensions $(d_0^I \times d_1^I \times 3)$ representing height, width, and the RGB values, while the LiDAR system

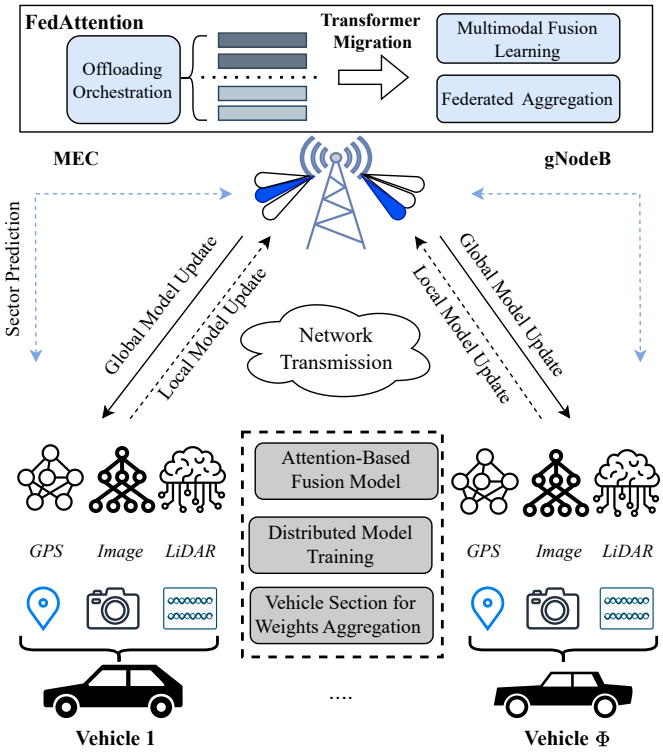


Fig. 1. Schematic of FedAttention framework.

generates a point cloud represented in a grid and can be formulated as a matrix with dimensions $(d_0^\zeta \times d_1^\zeta \times d_2^\zeta)$.

As shown in Fig.1, in FL, multiple vehicles collaboratively train the model based on their local multi-sensor datasets without sharing them. Each vehicle is equipped with fusion models for advanced training. We define the vehicle ϕ loss $F_\phi(\omega)$ given the model parameters ω as:

$$F_\phi(\omega) = \frac{1}{|D_\phi|} \sum_{(x,y)\in D_\phi} f(\omega, x, y), \tag{1}$$

where $f(\omega, x, y)$ is the local loss on the sample $(x, y) \in D_\phi$ and x represents a tuple in the dataset $\{X_{\phi,C}, X_{\phi,I}, X_{\phi,\zeta}\}$.

The model parameters across the participating vehicles will be shared with MEC for federated aggregation after the multi-modality biased selection module. Then, the aggregated model will be broadcast to the next-turn participating vehicles for continuous training. The goal is to minimize a global loss function $F(\omega)$, which is the weighted sum of local loss functions $F_\phi(\omega)$ across all devices. This is expressed as:

$$F(\omega) = \sum_{\phi=1}^{\Phi} \frac{|D_\phi|}{|D|} F_\phi(\omega), \tag{2}$$

where $|D|$ is the total size of all vehicle datasets combined.

B. CNN-Transformer-Based Fusion Model

Transformers rely on self-attention mechanisms to process sequences of data, enabling efficient parallelization and better handling of long-range dependencies [13]. In this paper, we propose a CNN-Transformer-based fusion learning framework

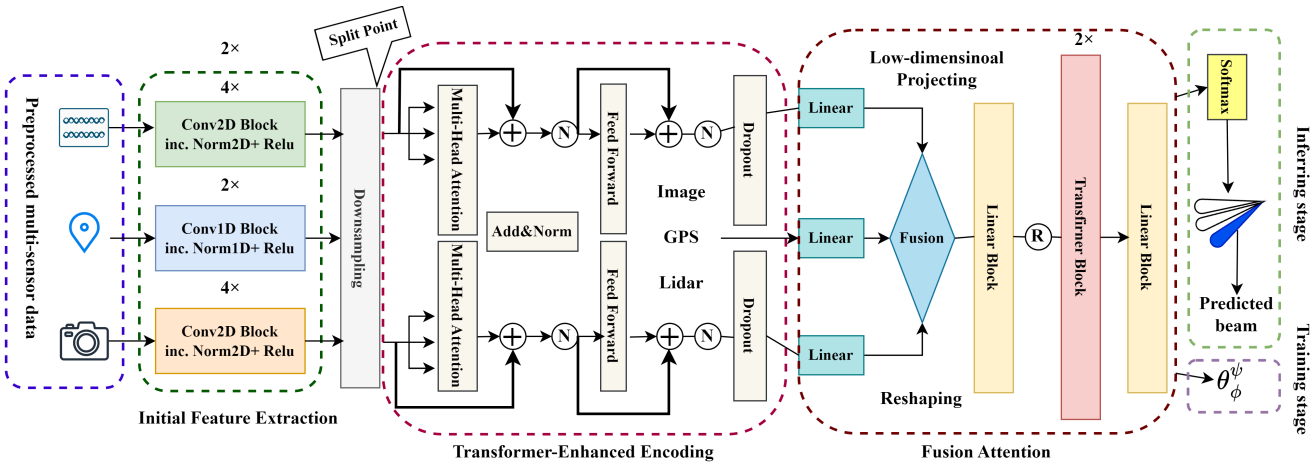


Fig. 2. CNN-Transformer-Based Fusion Architecture for Multimodal Beamforming.

for FL, namely `FedAttention`, designed for multi-modal beamforming. The detailed logical model architecture is presented in Fig. 2.

Initial Feature Extraction: The preprocessed multi-sensor data are embedded as the input of the CNN stacks for the initial spatial feature extractions. Each Conv1D Block for GPS or Conv2D Block for image and LiDAR data consists of the normalization layer, the activation function using Relu, as well as the convolutional layer, which applies filters over the input data to capture essential features like edges, textures, and patterns for $X_{\phi,C}, X_{\phi,I}, X_{\phi,\zeta}$, respectively. This stage reduces the complexity of data while preserving spatial relationships, enabling the model to focus on meaningful patterns in different sensor data.

Downsampling: The downsampling technique is applied to reduce the spatial dimensions of the feature maps. The advanced AdaptiveAvgPool layer is applied, which dynamically adjusts the output size, not only reducing the computational load but also enhancing the model's ability to generalize by preserving important spatial features. Since the Transformer architecture handles the most computationally intensive tasks, and the AdaptiveAvgPool layer effectively reduces the size of the intermediate activations, it is crucial to note that we place the potential split point here for further offloading.

Transformer-Enhanced Encoding (TEE): In traditional CNNs, local features are captured well, but global relationships or dependencies can be lost. The Transformer block compensates for this by learning the global context. This block utilizes the Transformer architecture's attention mechanism and computes weighted averages of inputs after downsampling to capture long-range dependencies in the 3D image and LiDAR data.

Fusion Attention: After TEE, tensor data pass into the Linear consisting of a set of fully connected layers, which aggregate the features learned from previous layers and map the extracted features into the output space. Next, it will go into the fusion part for further training or inferring. The Linear Block will reshape the tensor data as appropriate input to the fusion Transformer Block. We use two Transformer blocks in sequence to allow for more complex feature interactions, especially enabling the model to learn interactions between the outputs of the three sub-branches. After the Transformer blocks, we use the Linear Block to map output space and the softmax for the probability of all the probe frames, which can be expressed as:

$$u_{FN}^{\phi} = \sigma(f_{\phi,\psi}^{\theta}), \tag{3}$$

where $\Psi = \{C, I, \zeta, \varpi\}$ is the index set of the model branches, ϖ is the integrated branch, and θ represents all the model parameters. We use Elastic Net regularization techniques in the Linear Block to avoid overfitting during the training [14]. Finally, for example, considering the downlink communication, we can predict the best sector r^* with the maximum received signal strength after the mapping function Γ_θ from the obtained datasets, which can be expressed as:

$$r^* = \underset{n\in\{1\cdots N\}}{\arg\max}\ \hat{y}_{r_n}\ (\Gamma_\theta : D_\phi \to \hat{y}_{r_n}), \tag{4}$$

where $\hat{y}_{r_n}$ is the predicted received signal strength at the T_x side with the R_x configured with r_n.

C. Federated Learning for Multi-modal Beamforming

During the distributed training phase, the local dataset D_ϕ belonging to each vehicle ϕ is divided into equally-sized mini-batches, denoted as ς. Let j represent the index for local iterations within the k-th round of FL. Using mini-batch gradient descent, the local model weights $\mathbf{w}_\phi$ are iteratively updated according to the following equation:

$$\mathbf{w}_\phi^{(j+1)} = \mathbf{w}_\phi^{(j)} - \eta \frac{1}{\varsigma} \sum_{i=1}^{\varsigma} \nabla_{\mathbf{w}_\phi^{(j)}} F_\phi(\omega), \tag{5}$$

where η is the learning rate, which is dynamically adjusted using the StepLR scheduler [15], $F_\phi(\omega)$ denotes the local loss function of vehicle ϕ, which quantifies the discrepancy between the true labels $y^{(i)}$ derived from Radio Frequency (RF) data and the corresponding predicted labels $\hat{y}^{(i)}$.

The FL process involves local model training at the device and federated aggregation at the MEC, where autonomous vehicles periodically share their model parameters, denoted as θ_ϕ, at a global level. To further enhance the training computation efficiency, we allow each vehicle to decide dynamically whether to utilize split learning (SL) during the model training. Considering a set of MEC servers as E, we define an *offloading indicator* as Θ, which can be expressed as:

$$\Theta = \begin{cases} 1, & \dfrac{f_N\left(\frac{B_{\phi,e}}{B_r}\right)}{f_N\left(\frac{\rho_{\phi,e}}{\rho_r}\right)} > \gamma \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where $\rho_{\phi,e}$ is the capability ratio between the vehicle ϕ and the MEC server $e \in E$, $B_{\phi,e}$ is the bandwidth between them. The function $f_N(\cdot)$ performs normalization, while B_r and ρ_r serve as reference factors, and γ is a predefined threshold. For instance, if the network throughput is sufficient and the vehicle's computing power is limited, Θ will be set to 1 based on γ. In this case, the vehicle will employ SL and offload the intermediate activations, the volume of which is expressed as V_ϕ to the MEC server before the Transformer-Enhanced Encoding stage and Fusion Attention stage for collaborative model training, where $A(\cdot)$ and $A'(\cdot)$ represent the split training functions on vehicle and MEC servers, respectively.

This technique leverages the robust computing power of the B5G MEC servers to handle the more complex computations of the fusion transformer architecture, which is named Transformer Migration. As a result, it can significantly reduce training time and alleviate the bottlenecks at the FL synchronization point, thereby improving overall distributed training efficiency. The FedAttention algorithm is outlined in Algorithm 1, consisting of jointly local training (refer to lines 3 to 10), federated aggregation, and model broadcasting (refer to lines 11 to 12).

To ensure the robustness of our framework, we also calculated the model size of FedAttention and compared it with the FLASH framework [12]. Our findings show that the total number of parameters in FedAttention is only 31.7% of those in the FLASH framework. Despite utilizing a self-attention mechanism, FedAttention employs a balanced combination of convolutional and linear layers, advanced downsampling techniques, and a moderately sized transformer block. This architecture significantly reduces the parameters in the linear block compared to the very deeper architecture of FLASH. The total model size is 130 MB, making it smaller than both the very deep FLASH framework and the object detection models DETR (around 159 MB to 232 MB) [16] commonly used in modern vehicles. This reduction improves efficiency during model exchange in the FL stage.

Algorithm 1: FedAttention for Fusion Learning in Multi-Modal Beamforming for IoV

Input:
$D = \{D_\phi \mid \phi \in \Phi\} = \{X_{\phi,C}, X_{\phi,I}, X_{\phi,\zeta} \mid \phi \in \Phi\}$: Distributed multi-sensor datasets on all vehicles Φ
γ: Predefined threshold for the offloading indicator.
Output:
Γ_θ: Mapping function with the global model

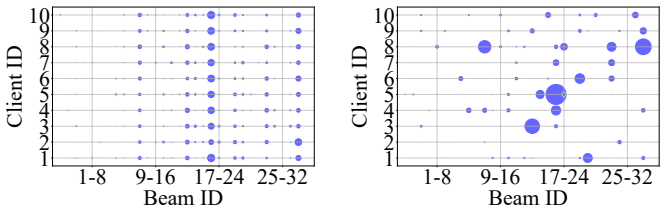
```
/* Local model nationalization                    */
θ_φ^ψ ← θ_g^ψ(0)
for k ∈ 𝕂 do
    for j ∈ 𝕁 do
        for φ ∈ Φ do
            if Θ then
                /* MEC-assisted FSL               */
                V_φ,e ← A^(k)_θ_φ^C,I,ζ (D_φ)
                ∇V_φ,e, w_φ ← A'^(k)_θ_e (V_φ,e)
            else
                w_φ ← f^(k)_θ_φ^ψ (D_φ)    // Classic FL
                /* Local loss calculation         */
                F_φ(ω) ← 1/|D_φ| Σ_(x,y)∈D_φ f(ω, x, y)
    /* Federated aggregation                      */
    θ^ψ(k) ← Σ_φ∈Φ |D_φ|/|D| θ_φ^ψ(k) for φ ∈ Φ do
        /* Global model sharing                   */
        θ_φ^ψ(k+1) ← θ^ψ(k)
```

III. EVALUATION AND SIMULATION RESULTS

A. Experimental Setup

All experiments were performed using real-world FLASH datasets [12] on our laboratory server, which is equipped with NVIDIA A40 GPUs delivering 149.6 TFLOPS for mixed-precision (FP16/FP32) tensor operations, ensuring robust parallel computing capabilities throughout the testing process.

B. Dataset Heterogenity



(a) iid settings (b) non-iid settings

Fig. 3. Label distribution based on $Dir(\alpha)$ sampling with 32 label beams.

Each client typically has its own local dataset, which may not be representative of the entire global dataset. On the other hand, the vehicles may have access to non-iid datasets, for example, vehicles collecting data from different environments, such as highly-density building scenarios and suburban scenarios. Since the global model needs to generalize across all clients, the accuracy of the vehicular FL framework is susceptible to non-iid datasets across different vehicles.

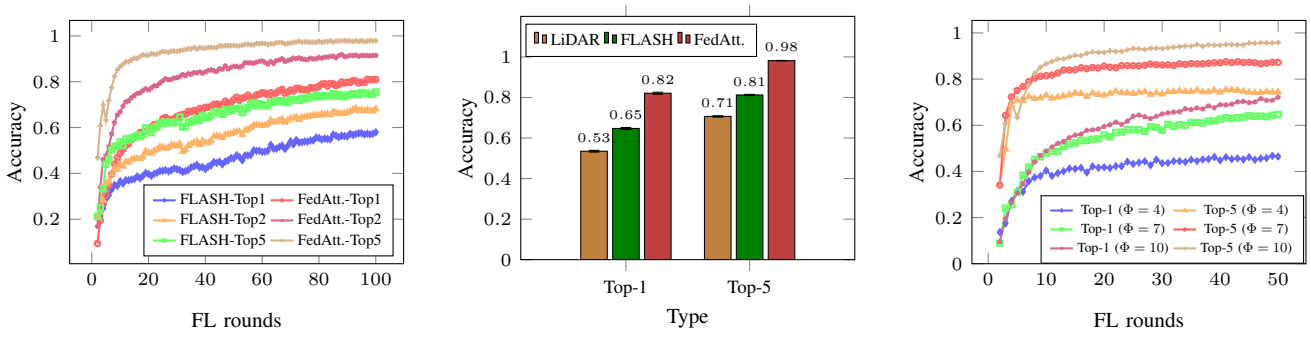


Fig. 4. (a) Comparison of accuracy with the FLASH across FL rounds, (b) Achieved final accuracy among the FL methods, (c) Effect of the number of vehicles (Available datasets) on FedAttention.

Therefore, we delve into the influence of differing non-iidness and the volume of training samples available on each device on model convergence and accuracy to prove the robustness of our framework. We redistributed the datasets from the Dirichlet distribution $Dir(\alpha)$ [17], which can be expressed as:

$$f(C_{Tx}, C_{Rx}; \alpha_1, ..., \alpha_{M+N}) = \frac{\Gamma(\sum_{i=1}^{M+N} \alpha_i)}{\prod_{i=1}^{M+N} \Gamma(\alpha_i)} \prod_{i=1}^{M+N} x_i^{\alpha-1}, \quad (7)$$

where $C_{Tx} = \{t_1, ..., t_M\}$, and $C_{Rx} = \{r_1, ..., r_N\}$ are the predefined sets of beamforming vectors that are used to specifically direct the signals for the transmitter and receiver, $\Gamma(\cdot)$ is the Gamma function, $M + N$ are the number of the probe frames required by the traditionally mandatory Sector Level Sweep (SLS) in IEEE 802.11ad standard sector initialization steps, and α_i is the concentration parameter. The α_i controls how much of the distribution's probability mass is concentrated around particular proportions of the probe frames. We set the $\{\alpha_i\}_{i=1}^{M+N} = \{0.1\}_1^{M+N} or \{10^b\}_1^{M+N}$ to represent non-iid and iid settings, where $b > 2$, the label distributions of which are shown in Fig. 3.

C. Results Analysis

Model generalization ability: During the training phase, we configure the computing power between MEC servers and vehicles, varying from 5 to 10 times. The network throughput fluctuates dynamically between 20 and 80 Mbit/s during the parameter exchange in each FL round, with federated aggregation occurring only upon reaching the synchronization point. FedAttention combines the strengths of CNN and transformer architectures to efficiently capture both local and global features and effectively manage complex interactions within the fusion model. As shown in Fig. 4(a), it achieves 81.08% Top-1 accuracy in 100 communication rounds, representing a 39.91% improvement over the FL-based FLASH framework. Top-5 accuracy refers to the probability that the optimal beam is included in the top five candidates. Compared to FLASH, FedAttention achieves 97.90% Top-5 accuracy, reflecting a 29.67% improvement and emphasizing its superior generalization performance. Additionally, as depicted in Fig. 4(b), the fusion model outperforms the single-modal LiDAR model, with FedAttention achieving a final best beam prediction accuracy of 82.09%, surpassing all other methods. After 200 FL rounds, the FLASH model exhibits minimal improvement in Top-1 accuracy and fails to reach the 70% threshold, further highlighting the superior effectiveness of FedAttention.

Effect of the number of clients (Available datasets): The impact of varying numbers of participating clients is evaluated in FedAttention, which directly affects the amount of available data for training the global model. As shown in Fig. 4(b), when the dataset size is significantly reduced, or the number of participating vehicles is very low ($\Phi = 4$), the Top-1 accuracy declines sharply, indicating that the global model struggles to capture the diverse interests of all vehicles. However, a moderate reduction in the size of the training datasets (18045 samples) and the number of participating vehicles (Seven clients) results in only a slight decrease in accuracy, which remains within an acceptable range with 87.25% Top-5 accuracy. Notably, even with only seven vehicles involved in FL training, the accuracy still surpasses that of FLASH, while the training time is significantly reduced. This demonstrates the strong generalization capabilities of the FedAttention model.

Performance in heterogeneous environments: In the FL process, dataset heterogeneity significantly impacts the accuracy and convergence of the global model. In this phase, we set the α_i from the Dirichlet distribution to control the characteristics of the datasets. As shown in Fig. 5., when the α_i is set as 0.1 to represent the non-iid settings, the

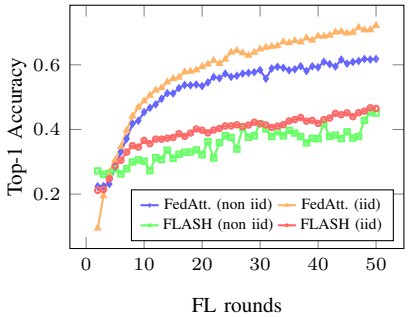


Fig. 5. Accuracy with different concentration factors across FL rounds

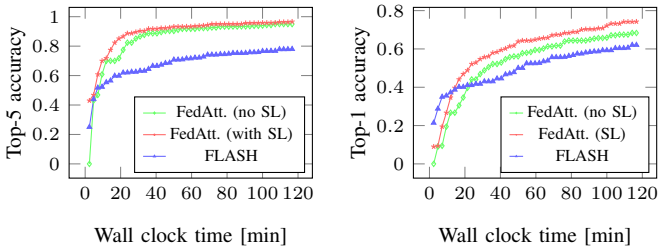


Fig. 6. Achieved maximum accuracy across the wall clock time

datasets on client 2 have only 397 data samples, only 0.063 times the data samples of client 6. Under this condition, FedAttention's Top-1 accuracy drops from 72.16% to 61.82% in 50 FL rounds, highlighting the impact of data heterogeneity. For FLASH, the accuracy also declines, and more notably, the convergence process is even less smooth compared to IID settings. Despite the non-IID environment, FedAttention significantly outperforms FLASH in terms of accuracy and convergence, demonstrating its robustness.

Computing efficiency: Since the transformer leverages the self-attention mechanism, increasing computation time per FL round, we assess the accuracy in relation to wall clock time to account for both local training and global communication time. To further enhance training efficiency, FedAttention incorporates SL to solve the potential struggler arising from the heterogeneity of the different vehicle systems. As illustrated in Fig. 6, while FedAttention may have a longer training time per FL round compared to FLASH, it requires fewer rounds to converge and achieve higher prediction accuracy. Within a fixed timeframe of 109.53 minutes, FedAttention with SL achieves a 74.17% Top-1 accuracy, outperforming both FedAttention without SL (67.47%) and FLASH (60.38%). Moreover, in terms of Top-5 accuracy, FedAttention significantly exceeds FLASH (77.85%), achieving 96.35% and 94.98% Top-5 accuracy with or without SL. This demonstrates the training efficiency of the FedAttention framework, particularly when using SL, despite the need to exchange a small amount of smashed data—only 1.01 MB per communication round.

IV. CONCLUSIONS

In this paper, we propose FedAttention, a federated attention-based fusion learning framework designed for multi-modal beamforming. By fully leveraging the collaborative learning between MEC servers and local vehicle resources, FedAttention significantly enhances computational efficiency. We also demonstrate the robustness of our model across heterogeneous datasets. In terms of beam sector prediction accuracy, FedAttention achieves a Top-5 accuracy of 98.16% and a Top-1 accuracy of 82.09%, representing improvements of 26.86% and 20.84%, respectively, compared to FLASH, all while maintaining a smaller model size. By integrating CNN and transformer architectures across multiple branches and the fusion block, results based on real-world datasets validate the superiority of our model generalization ability, showing it is both capable of benefiting from FL and maintaining high accuracy.

V. ACKNOWLEDGMENT

This research is supported by the Swiss National Science Foundation (SNSF) [Grant No. 219330] and partially supported by the U.S. National Science Foundation under grant ECCS 2516080.

REFERENCES

[1] Q. Liu, R. Luo, H. Liang, and Q. Liu, "Energy-efficient joint computation offloading and resource allocation strategy for ISAC-aided 6G V2X networks," *IEEE Transactions on Green Communications and Networking*, vol. 7, no. 1, pp. 413–423, 2023.

[2] M. Emara, M. C. Filippou, and D. Sabella, "MEC-assisted end-to-end latency evaluations for C-V2X communications," in *2018 European conference on networks and communications (EuCNC)*. IEEE, 2018, pp. 1–9.

[3] K. Sakaguchi, R. Fukatsu, T. Yu, E. Fukuda, K. Mahler, R. Heath, T. Fujii, K. Takahashi, A. Khoryaev, S. Nagata *et al.*, "Towards mmWave V2X in 5G and beyond to support automated driving," *IEICE Transactions on Communications*, vol. 104, no. 6, pp. 587–603, 2021.

[4] S. S. Sarma and R. Hazra, "Pathloss attenuation analysis for D2D communication in 5G mmWave networks," in *2020 Advanced Communication Technologies and Signal Processing (ACTS)*. IEEE, 2020, pp. 1–6.

[5] Y. Sun and C. Qi, "Analog beamforming and combining based on codebook in millimeter wave massive MIMO communications," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[6] Y. Zheng, S. Chen, and R. Zhao, "A deep learning-based mmWave beam selection framework by using lidar data," in *2021 33rd Chinese Control and Decision Conference (CCDC)*. IEEE, 2021, pp. 915–920.

[7] G. Reus-Muns, B. Salehi, D. Roy, T. Jian, Z. Wang, J. Dy, S. Ioannidis, and K. Chowdhury, "Deep learning on visual and location data for V2I mmWave beamforming," in *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*. IEEE, 2021, pp. 559–566.

[8] W. Xu, F. Gao, X. Tao, J. Zhang, and A. Alkhateeb, "Computer vision aided mmWave beam alignment in V2X communications," *IEEE Transactions on Wireless Communications*, vol. 22, no. 4, pp. 2699–2714, 2022.

[9] Y. Yang, F. Gao, X. Tao, G. Liu, and C. Pan, "Environment semantics aided wireless communications: A case study of mmWave beam prediction and blockage prediction," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 7, pp. 2025–2040, 2023.

[10] D. Roy, B. Salehi, S. Banou, S. Mohanti, G. Reus-Muns, M. Belgiovine, P. Ganesh, C. Dick, and K. Chowdhury, "Going beyond RF: A survey on how AI-enabled multimodal beamforming will shape the NextG standard," *Computer Networks*, vol. 228, p. 109729, 2023.

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[12] B. Salehi, J. Gu, D. Roy, and K. Chowdhury, "Flash: Federated learning for automated selection of high-band mmwave sectors," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 1719–1728.

[13] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[15] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[17] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.