Harnessing Multimodal Sensing for Multi-user Beamforming in mmWave Systems

Kartik Patel, Robert W. Heath Jr.

***Abstract*—Sensor-aided beamforming reduces the overheads associated with beam training in millimeter-wave (mmWave) multi-input-multi-output (MIMO) communication systems. Most prior work, though, neglects the challenges associated with establishing multi-user (MU) communication links in mmWave MIMO systems. In this paper, we propose a new framework for sensor-aided beam training in MU mmWave MIMO system. We leverage the beamspace representation of the channel that contains only the angles-of-departure (AoDs) of the channel's significant multipath components. We show that a deep neural network (DNN)-based multimodal sensor fusion framework can estimate the beamspace representation of the channel using sensor data. To aid the DNN training, we introduce a novel supervised soft-contrastive loss (SSCL) function that leverages the inherent similarity between channels to extract similar features from the sensor data for similar channels. Finally, we design an MU beamforming strategy that uses the estimated beamspaces of the channels to select analog precoders for all users in a way that prevents transmission to multiple users over the same directions. Compared to the baseline, our approach achieves more than 4 times improvement in the median sum-spectral efficiency (SE) at 42 dBm equivalent isotropic radiated power (EIRP) with 4 active users. This demonstrates that sensor data can provide more channel information than previously explored, with significant implications for machine learning-based communication and sensing systems.**

I. Introduction

Hybrid MIMO architectures enable large antenna arrays to support high-bandwidth applications like mmWave communication. Configuring communication links with hybrid MIMO architectures in commercial systems like 5G and IEEE 802.11ay make use of beam management protocols [1]–[4]. These protocols follow a two-stage procedure: in the first stage, known as beam training, the transmitter and receiver measure candidate beams from codebooks to configure analog precoders and combiners. In the second stage, with the analog precoders and combiners fixed, they perform channel estimation and design the digital precoder and combiner for multi-stream MIMO transmission. Despite the hierarchical nature of this process, the beam training stage introduces significant overhead in configuring MIMO communication links.

Kartik Patel is with the Chandra Department of Electrical and Computer Engineering, the University of Texas at Austin, Austin, TX, 78712 USA. Email: kartikpatel@utexas.edu.

Robert W. Heath Jr. is with the Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA. Email: rwheathjr@ucsd.edu.

This material is based upon work supported by the National Science Foundation under grant nos. NSF-ECCS-2435261, NSF-CCF-2435254, NSF-CNS-2433782 and is supported in part by funds from federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program.

Prior work has proposed using out-of-band sensor data to reduce the beam training overhead by narrowing the set of candidate beam pairs to try. These sensor data include position information [5]–[9], radar signals [10]–[15], camera images [16]–[20] and LiDAR point clouds [21]–[26]. While some early work used model-based signal processing [10], [11], most recent approaches leverage machine learning (ML) to learn the relationship between sensor information and the optimal beam configuration. Further fueled by the publicly available datasets like Raymobtime [27], e-FLASH [28] and DeepSense 6G [29], prior work has also explored using a combination of position information, camera images, and LiDAR point clouds to improve the accuracy of sensing-aided beam prediction strategies [25], [26], [30]–[35].

While existing ML-based approaches [5]–[9], [12], [13], [15]–[26], [30]–[36] show promise in the single user (SU) setting, their performance in the MU setting remains unclear (as summarized in Fig. 1a). Specifically, predicting the optimal beams for each user equipment (UE) separately might overlook inter-user interference; a crucial aspect of MU communication. Therefore, a framework is needed that is specifically designed for MU setting that leverages sensor data for low-overhead beam training while considering inter-user interference.

Some prior work has explored ML-based solutions for MU beam training using the history of channel state information (CSI) and signal strength measurements [37]–[41]. There is, however, limited work that leverages sensor data for MU beam training. One approach uses camera images for MU beam training and proposes a two-stage process with user selection and beam training stages, each using separate DNN models [42]. That approach, however, is limited by an assumption of the existence of a single dominant line-of-sight (LOS) path to each user, which is not realistic in practice [43].

In this paper, we propose using multimodal sensor data – camera images, LiDAR point clouds, and position information – for beam training in MU mmWave systems (Fig. 1b). Specifically, we show that these sensors can collectively provide more information about the AoDs of the significant multipath components in the mmWave channel than was previously assumed in prior works [5]–[9], [12]–[19], [22]–[24], [30]–[35], [44], [45].

To estimate the AoDs of the channel paths from sensor data, we employ a DNN-based multimodal fusion network similar to [31]. Instead of identifying a single optimal beam, however, we aim to identify AoDs of the strong channel paths. This requires transforming the problem from a classification to a multi-label identification problem. Thus, each quantized AoD pair (azimuth and elevation) is treated as a label. This approach

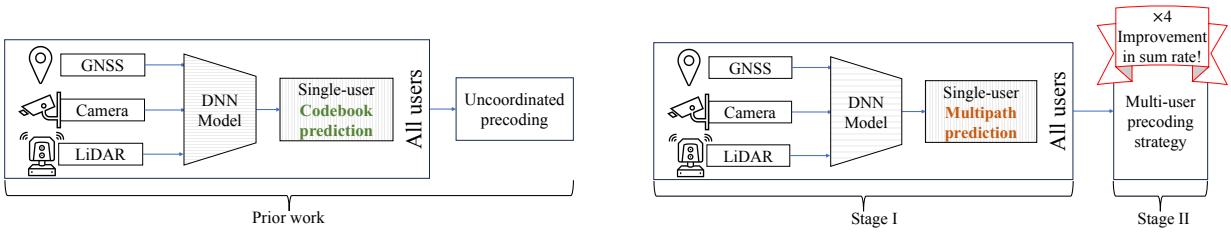


Fig. 1: Improving MU communication using side-information: (a) Prior work uses single-user codebook prediction that could suffer from inter-user interference in MU setting. (b) We show that predicting multipath components and using a dedicated beamforming strategy designed for MU scenario can quadruple the achievable rates for MU systems.

enables a fixed size DNN to predict the AoDs of a variable number of multipath components.

While the multi-label identification problem offers a basic methodology for DNN training, it has a limitation. Conventional multi-label identification tasks assume categorical labels with no notion of *proximity* between them. The labels can only be the same or different. Therefore, the samples are similar only if they share a few identical labels in the ground truth. In contrast, AoDs of channel paths have an inherent notion of proximity – they can be closer or farther apart, not just the same or different. Consequently, there is also a notion of similarity between the channels, even when their paths do not share identical AoDs. Furthermore, this notion of similarity between the channels can also enhance DNN training by encouraging it to extract similar features from sensor data when the channels have similar AoDs.

In this paper, we formalize this notion of similarity by introducing a novel *soft encoding* for channels. We then define a metric to measure the similarity between two channels based on their soft-encodings. We further introduce a novel loss function called SSCL to aid DNN training by incentivizing the intermediate features extracted by the DNN network to align or dis-align in proportion to the similarity of the channels. SSCL is inspired from supervised contrastive loss (SCL) [46] which incentivizes the DNN in a similar manner but is only designed for categorical classes without capturing the proximity between them. We believe SSCL loss function can be valuable for applying supervised ML algorithms to practical systems like wireless communication, where labels deal with non-categorical data and there exists a notion of similarity across labels. To the best of our knowledge, this is the first time such an encoding and loss function have been used for training supervised ML models.

Finally, we propose a novel MU beamforming algorithm that uses predicted AoDs of channel paths. Specifically, the algorithm uses a beamspace representation of the channel containing only the predicted AoDs and designs analog precoders that avoid simultaneous transmission to multiple users over the same multipath clusters. As a result, the base station (BS) can estimate the *effective channels* for all users simultaneously with minimal inter-user interference, thus, enabling the subsequent design of an optimized digital precoder. We show that this MU strategy achieves 4 times improvement in the median sum-SE with 4 active users compared to MU extensions of SU beam prediction strategies, and significantly reduces the link establishment overhead.

The contributions of this paper are summarized as follows:

1) We propose a framework for sensor-aided MU beam training at the BS. We use the beamspace representation of a channel, which focuses only on the AoDs of channel paths. We design an ML-based multimodal fusion network trained using the camera, LiDAR and position information, and the beamspaces of the associated channels. We formulate the problem as a multi-label identification task with each label representing a quantized AoD pair (azimuth and elevation).
2) We propose a novel soft-encoding technique and a corresponding similarity metric that captures the similarity between channels based on the proximity of the AoDs of their channel paths. Furthermore, we introduce the SSCL, a novel loss function that incentivizes the network to extract similar features from the sensors for similar beamspaces. We believe SSCL has broader applicability to various ML tasks in wireless systems that often involve non-categorical labels.
3) We present a novel MU beamforming algorithm that uses the estimated beamspace representations of the channels to identify the analog precoders at the BS. Subsequently, conventional preamble-based channel estimation techniques and regularized zero-forcing (RZF) are used to design the digital precoders.
4) We conduct a comprehensive analysis on the Raymobtime dataset [27], assessing the end-to-end performance of the proposed sensor-aided MU beamforming against two baselines, SU approaches extended to MU setting and Full CSI setting. This evaluation showcases the ability of sensor data to extract significantly more information about channels and the benefit of using a joint MU beamforming strategy.

Organization: The rest of the paper is structured as follows: Section II presents the system model and defines the beamspace of a channel. Section III details the sensor-aided beamspace prediction, including sensor data preprocessing, beamspace encoding, the DNN architecture, and metrics used for training and testing. Section IV describes the MU beamforming algorithm that leverages the estimated beamspaces of the channels for selecting analog and digital precoders to minimize inter-user interference across users. Section V presents the performance analysis of the sensor-aided MU beamforming

strategy. Finally, Section VI provides the concluding remarks and directions for future research.

Notations: Small bold letters ($\mathbf{a}$) and capital bold letters ($\mathbf{A}$) denote a vector and matrix, respectively. $\bar{\mathbf{a}}, \mathbf{a}^{\mathrm{T}}, \mathbf{a}^{*}$ denotes conjugate, transpose, and Hermitian of the vector/matrix $\mathbf{a}$. $[a]$ denotes the set $\{0, \ldots, a-1\}$. $\mathrm{vec}(\mathbf{A})$ denotes the vectorization operator on the matrix $\mathbf{A}_{M\times N}$ such that the i,j-th element of the matrix $\mathbf{A}$ maps to $(iN+j)$-th element of the column vector $\mathrm{vec}(\mathbf{A})$, $\forall i \in [M], j \in [N]$. The hat $\hat{\mathbf{a}}$ denotes estimated value of $\mathbf{a}$. $\mathbf{1}_N$ denotes the vector of length N with all elements equal to one and $\mathbf{I}_N$ denotes the identity matrix of dimensions $N \times N$. Finally, $\mathrm{ADiff}(\theta_1, \theta_2)$ denotes the smallest absolute difference between two angles θ_1, θ_2.

II. System Model

In this section, we first describe the system model and the performance metrics used in this paper. We then define the beamspace representation of a channel.

A. Antenna model

We consider MU-MISO OFDM system. The BS is equipped with N_{S} symbol streams and N_{RF} radio frequency (RF) chains. It serves U UEs simultaneously using MU beamforming.

In this paper, we only focus on the beam training at the BS. We assume the UEs are equipped with omni-directional antennas with the same polarization as the BS antenna array. Consequently, each UE receives a single symbol stream from one RF chain of the BS antenna array, resulting in a configuration where $N_{\mathrm{S}} = N_{\mathrm{RF}} = U$. Extending this paper to consider UEs with multiple antennas, hence, a MIMO system, is a potential area for future research.

We assume the BS uses fully connected hybrid antenna architecture with N_{RF} RF chains connected to a uniform rectangular antenna array of $N^{\mathrm{x}}_{\mathrm{BS}} \times N^{\mathrm{y}}_{\mathrm{BS}}$ elements spaced at half-wavelength distance apart. The BS antenna array is positioned on the XY-plane with its broadside direction along the Z-axis (see Fig. 2a) [9], [47]. We denote the azimuthal and elevation angles by $\theta \in [-\pi, \pi)$ and $\phi \in [0, \pi]$, respectively, and define $\Omega_x = \cos\theta \sin\phi, \Omega_y = \sin\theta\sin\phi$. Then, the Vandermonde vectors along the X and Y directions can be defined as

$$\mathbf{a}_{\mathrm{x}}(\theta,\phi) = \frac{1}{\sqrt{N^{\mathrm{x}}_{\mathrm{BS}}}}[1, e^{-\mathrm{j}\pi\Omega_x}, \ldots, e^{-\mathrm{j}(N^{\mathrm{x}}_{\mathrm{BS}}-1)\pi\Omega_x}]^{\mathrm{T}}, \quad (1)$$

$$\mathbf{a}_{\mathrm{y}}(\theta,\phi) = \frac{1}{\sqrt{N^{\mathrm{y}}_{\mathrm{BS}}}}[1, e^{-\mathrm{j}\pi\Omega_y}, \ldots, e^{-\mathrm{j}(N^{\mathrm{y}}_{\mathrm{BS}}-1)\pi\Omega_y}]^{\mathrm{T}}. \quad (2)$$

The antenna array response matrix, denoted by $\mathbf{A}(\theta,\phi) \in \mathbb{C}^{N^{\mathrm{y}}_{\mathrm{BS}} \times N^{\mathrm{x}}_{\mathrm{BS}}}$, can be defined as

$$\mathbf{A}(\theta,\phi) = \mathbf{a}_{\mathrm{y}}(\theta,\phi)\mathbf{a}_{\mathrm{x}}(\theta,\phi)^{\mathrm{T}}. \quad (3)$$

Using this antenna model, we now define the channel and signal models.

B. Channel model

We consider the wideband channel model with D-taps as defined in [48]. We assume the channel is static over one time slot. At a time-slot n, let $\left(\alpha^{(n)}_{u,c,\ell}, \tau^{(n)}_{u,c,\ell}\right), \forall \ell \in [L^{(n)}_{u,c}], c \in [C^{(n)}_u]$ be the gain and delay of the ℓ-th path in the c-th ray cluster to UE u. Let $\left(\theta^{(n)}_{u,c,\ell}, \phi^{(n)}_{u,c,\ell}\right), \forall \ell \in [L^{(n)}_{u,c}], c \in [C^{(n)}_u]$ be the azimuthal and elevation angles of the ℓ-th path in the c-th ray cluster to UE u. Further, let $p(t)$ denote the pulse shaping function evaluated at time t. Consequently, the frequency response of the channel $\mathbf{H}^{(n)}_u[k] \in \mathbb{C}^{N^{\mathrm{y}}_{\mathrm{BS}} \times N^{\mathrm{x}}_{\mathrm{BS}}}$ over the sub-carriers $k \in \{-\lfloor \frac{K-1}{2} \rfloor, \ldots, \lceil \frac{K-1}{2} \rceil\}$ can be defined as

$$\mathbf{H}^{(n)}_u[k] = \sum_{d=1}^{D}\sum_{c=1}^{C^{(n)}_u}\sum_{\ell=1}^{L^{(n)}_{u,c}} \Big[\alpha^{(n)}_{u,c,\ell} p\left(dT_{\mathrm{S}} - \tau^{(n)}_{u,c,\ell}\right) \times \mathbf{A}\left(\theta^{(n)}_{u,c,\ell}, \phi^{(n)}_{u,c,\ell}\right) e^{-\mathrm{j}\frac{2\pi dk}{K}}\Big]. \quad (4)$$

We also denote the vectorized channel $\mathrm{vec}(\mathbf{H}^{(n)}_u[k])$ by $\mathbf{h}^{(n)}_u[k]$ in the subsequent signal modeling.

C. Signal model

We consider a downlink MU-MISO system with linear beamforming where each user receives a single data stream (see Fig. 2b). Let $\mathbf{s}[k] \in \mathbb{C}^{N_{\mathrm{S}}\times 1}$ be a vector of symbols such that each element $\mathbf{s}_i[k]$ denotes the symbol to a UE over the k-th sub-carrier. We assume $\mathbb{E}[\mathbf{s}[k]\mathbf{s}^{*}[k]] = \frac{P}{N_{\mathrm{S}}}\mathbf{I}_{N_{\mathrm{S}}}$, where P is the average transmitted power over k-th sub-carrier.

The BS uses hybrid precoding by first digitally precoding each sub-carrier using the digital precoder $\mathbf{F}_{\mathrm{BB}}[k] = [\mathbf{f}^{\mathrm{BB}}_1[k], \cdots, \mathbf{f}^{\mathrm{BB}}_{N_{\mathrm{S}}}[k]] \in \mathbb{C}^{N_{\mathrm{RF}}\times N_{\mathrm{S}}}$, followed by the RF precoding (or analog precoding) using the RF precoder $\mathbf{F}_{\mathrm{RF}} = [\mathbf{f}^{\mathrm{RF}}_1, \cdots, \mathbf{f}^{\mathrm{RF}}_{N_{\mathrm{RF}}}] \in \mathbb{C}^{N^{\mathrm{x}}_{\mathrm{BS}}N^{\mathrm{y}}_{\mathrm{BS}}\times N_{\mathrm{RF}}}$. By denoting the additive Gaussian noise $\mathsf{n}_u[k] \sim \mathcal{N}(0,\sigma^2)$, the received signal at the intended UE u can be written as

$$\mathsf{y}_u[k] = \mathbf{h}^{\mathrm{T}}_u[k]\mathbf{F}_{\mathrm{RF}}\mathbf{f}^{\mathrm{BB}}_u[k]\mathbf{s}_u[k] + \mathbf{h}^{\mathrm{T}}_u[k]\mathbf{F}_{\mathrm{RF}}\sum_{u'=1,u'\neq u}^{U}\mathbf{f}^{\mathrm{BB}}_{u'}[k]\mathbf{s}_{u'}[k] + \mathsf{n}_u[k]. \quad (5)$$

Note that the second term in the signal model characterizes the inter-user interference caused by other active users. We set the elements of $\mathbf{F}_{\mathrm{RF}}$ to have a unit magnitude for modeling the RF precoding by a passive analog phased array. Accordingly, the choice of $\mathbf{F}_{\mathrm{RF}}$, specifically, the phase of each element of $\mathbf{F}_{\mathrm{RF}}$, defines the beam pattern of the BS antenna.

D. Metric: Achievable spectral efficiency (SE)

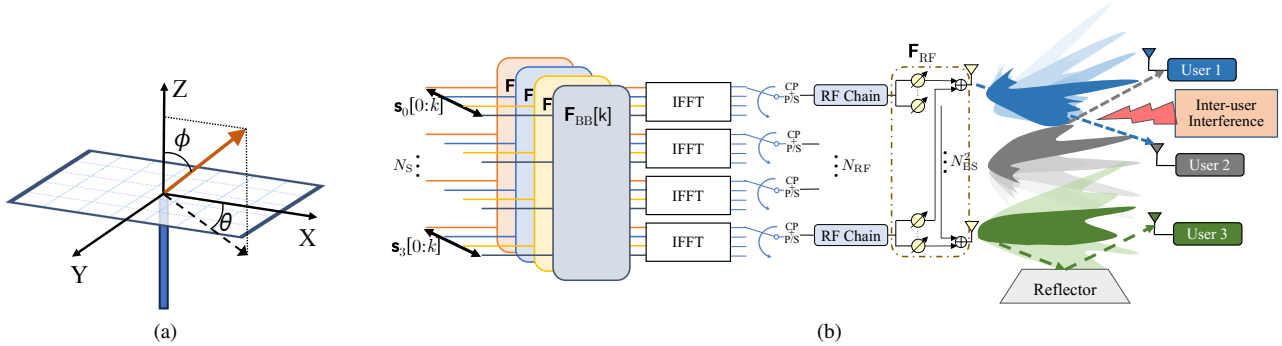
With the proposed signal model, we now define the achievable sum SE of the system to evaluate the performance of the MU communication system. Assuming Gaussian signaling and treating inter-user interference as Gaussian noise, the achievable SE of the UE u at a time defined as

$$R_u = \frac{1}{K}\sum_{k=0}^{K-1}\log_2\left(1 + \frac{\frac{P}{N_{\mathrm{S}}}\left|\mathbf{h}^{\mathrm{T}}_u[k]\mathbf{F}_{\mathrm{RF}}\mathbf{f}^{\mathrm{BB}}_u[k]\right|^2}{\sigma^2 + \frac{P}{N_{\mathrm{S}}}\sum_{u'\neq u}\left|\mathbf{h}^{\mathrm{T}}_u[k]\mathbf{F}_{\mathrm{RF}}\mathbf{f}^{\mathrm{BB}}_{u'}[k]\right|^2}\right). \quad (6)$$

Further, we define the sum-SE of the system as $R_{\mathrm{S}} = \sum_u R_u$.

E. The beamspace of a channel

We call the list of AoD pairs of significant paths, $\left(\theta^{(n)}_{u,c,\ell}, \phi^{(n)}_{u,c,\ell}\right), \forall \ell \in \left[L^{(n)}_{u,c}\right], c \in C^{(n)}_u$, in the channel the



(a) (b)

Fig. 2: (a) Geometry of the BS antenna array. (b) Schematic diagram: We consider a MU-multi-input-single-output (MISO) orthogonal frequency division multiplexing (OFDM) system with a frequency-selective channel model. The BS equipped with a planar antenna array with hybrid antenna architecture transmits to $U = 4$ UEs equipped with an omnidirectional antenna array. Greedily choosing the optimal beam for each UE can introduce significant inter-user interference.

AoD-list of the channel. Further, we let the 2-D angular space of azimuthal and elevation angles is quantized on the grid of size $G_\theta \times G_\phi$, such that the (i,j)-th index of the grid is associated to the direction (θ_i, ϕ_j), where $\theta_i = \frac{2\pi}{G_\theta} i - \pi, \phi_j = \frac{\pi}{G_\phi} j$. Then, $\forall i \in [G_\theta], j \in [G_\phi]$, we define the beamspace representation of the channel as

$$\tilde{\mathbf{G}}^{(n)}_{u,(i,j)} = \sum_{c=1}^{C_u^{(n)}} \left| \frac{1}{L^{(n)}_{u,c}} \sum_{\ell=1}^{L^{(n)}_{u,c}} \mathbf{a}^*_{\mathrm{y}}(\theta_i, \phi_j) \times \mathbf{A}\left(\theta^{(n)}_{u,c,\ell}, \phi^{(n)}_{u,c,\ell}\right) \times \bar{\mathbf{a}}_{\mathrm{x}}(\theta_i, \phi_j) \right|^2 . \quad (7)$$

The proposed beamspace representation of the channel represents the expected strength of the channel along the directions indexed by (i,j) assuming the total gains of all ray clusters are identical. This beamspace representation of the channel is a normalized projection of the *physical* channel on the oversampled antenna array manifold [49]. This projection provides a geometric perspective on the channel as observed by the BS equipped with planar arrays as defined in Section II-A. Consequently, the beamspace representation of a channel also captures the impact of the grating lobes produced by the planar array. For example, if a beam along the direction (θ_i, ϕ_j) produces a grating lobe, and if there exists a path along the direction of the grating lobe, then the beamspace representation captures its impact. Furthermore, the beamspace representation does not contain rapidly varying channel features, hence, we characterize it with a frequency flat response.

We acknowledge that there can be a large number of multipath components in mmWave channels [50]. Hence, a complete beamspace of a mmWave channel can include a lot of multipath components which are infeasible to predict purely from the sensor data. Therefore, we only consider the $\mathcal{L}(< L^{(n)}_{u,c})$ strongest paths from each ray cluster to define the *truncated* beamspace representation of the channel. In this paper, we show that the multimodal sensors in the system can estimate the truncated beamspace representation of the channel (in Section III) and the estimated truncated beamspace representation can be used for MU beamforming in mmWave communication (in Section IV). Since all discussion for the rest of the paper applies to both – truncated and non-truncated – beamspace representations, we avoid the term truncated for consistency unless explicitly required.

While the concept of beamspace exists in prior work [49], our definition diverges slightly. We assume a path gain of unity for each channel path, and a normalized gain of each multipath cluster because of the difficulty in predicting path gains solely from sensor data. Exploring alternative methods for path gain prediction can be a promising avenue for future research.

Our proposed beamspace representation, focusing solely on AoDs, aligns with concepts used in prior mmWave channel estimation techniques. For instance, compressive sensing-based approaches include an intermediate step for identifying the *support set* of the sparse beamspace representations [51]–[54]. This support set, defined as the indices of the non-sparse elements in the beamspace representation of the channel, denotes the directions of channel paths without their path gains. Similarly, side-information-aided beamforming solutions use out-of-band channel characteristics to derive a prior over the AoDs, essentially creating a prior over the beamspace without path gains [55]–[57]. These intermediary steps in prior work can be viewed as a form of beamspace representations without using the path gains. Like these techniques, our algorithm also leverages the proposed AoD-based beamspace as an intermediate step to map sensor data to precoder selection.

III. Stage I: Sensor-aided Beamspace Estimation

In this section, we describe the process of the sensor-aided prediction of the channel's beamspace representation. We begin by describing the sensor data preprocessing and the encoding techniques for AoDs of the channel paths. We then present the DNN-based multimodal sensor fusion network and loss functions employed for network training.

A. Preprocessing sensor data

The method of preprocessing the input data (i.e., sensor data) has a great impact on the choice of the DNN architecture. Thus, in this subsection, we explain the preprocessing steps used in this paper and accordingly, present the DNN architecture in Section III-C.

We consider UEs equipped with a LiDAR device and a positioning service, while the BS is equipped with a camera.

Given our focus on capturing only the large-scale channel effects for beamforming and assuming a sufficiently low beam retraining period (approximately 500 ms) [58], we assume the sensors capture samples synchronously with a frequency of at least 10 Hz. We now describe the preprocessing steps for each sensing modality.

1) LiDAR samples: Each LiDAR sample provides a snapshot of the environment surrounding the UE in the form of a point cloud, where each point represents a reflection of emitted light pulses. Objects introduce a higher density of points due to numerous reflections. Therefore, the number of points in a LiDAR point cloud varies depending on the environment. Such variable-length inputs raise a challenge for designing DNN models requiring fixed-size inputs. Thus, the key objective of LiDAR data processing is to encode the variable-length sensor data samples into fixed-size samples.

We adopt the standard LiDAR preprocessing method discussed in [23], [31], [59], which encodes the variable-length sensor data into a fixed-size 3D matrix. Specifically, we first define a space corresponding to the BS coverage using the coordinates $(X_{\min}, X_{\max})$, $(Y_{\min}, Y_{\max}), (Z_{\min}, Z_{\max})$. We then uniformly partition the 3D space by quantizing each dimension to $b_{\mathrm{x}}, b_{\mathrm{y}}, b_{\mathrm{z}}$ levels, respectively. Next, we assign the value -2 to the bin containing the BS and the value -1 to the bin containing the UE. Finally, for the remaining bins indexed by $(i, j, k) \in [b_{\mathrm{x}}^{\mathcal{L}}] \times [b_{\mathrm{y}}^{\mathcal{L}}] \times [b_{\mathrm{z}}^{\mathcal{L}}]$, we assign the value 1 if there exists at least one LiDAR point that falls within the specific bin.

2) Camera images: The camera captures a visual snapshot of the environment around the BS in the form of an image. We assume the BS captures a 180-degree field-of-view (FoV) using either a single camera or multiple cameras with images stitched together. In this paper, we resize each 180-degree FoV image to a grayscale image of size $b_{\mathrm{h}}^{\mathcal{C}} \times b_{\mathrm{w}}^{\mathcal{C}}$ and normalize the pixel values between 0 and 1.

3) Coordinates: We assume that the UEs have positioning capability from RF-based localization, Global Navigation Satellite Systems (GNSS), sensor-based Simultaneous Localization and Mapping (SLAM) [60] or other such methods. In this paper, we convert the coordinates of UEs with respect to the BS location and define a one-dimensional input of length 2.

B. Encoding the AoD-list of the channel for training

In this subsection, we first discuss the motivation for using a novel approach for encoding the AoD-list of the channel and then propose two encoding techniques. The encoded AoD-list is used as a "ground truth" for training the multimodal fusion network described in Section III-C. Accordingly, we assume that for each sample of the sensor data, the AoD-list of the associated channel is available during the training process.

1) Motivation: In this paper, we model the task of predicting AoD-pairs as a multi-label identification problem [61] due to a fundamental difference between our work and the prior work like [22], [23], [31], [62]. The prior work have focused on identifying a single beam (or top-K beams) from a codebook based on processed sensor data. Therefore, they framed the problem as a classification task with each beam being a class encoded using one-hot encoding. We, however, focus on identifying the AoD-list of channels with more than one channel path. Thus, the AoDs of the channel paths can not be encoded using one-hot encoding. Therefore, we consider a quantized AoD-pair as a label and treat the AoD-list prediction problem as the multi-label identification problem with the multi-label encoding for the AoD-list of the channel.

In addition to the standard multi-label encoding, we also require an encoding for the AoD-list that can capture a notion of similarity between channels. For example, a channel path with AoD $(\theta, \phi) = (0^\circ, 0^\circ)$ is closer to the path with AoD $(2^\circ, 2^\circ)$ than $(10^\circ, 10^\circ)$. As a result, the channel having only one AoD $(0^\circ, 0^\circ)$ is closer to the channel having only AoD $(2^\circ, 2^\circ)$ than the channel having only one AoD $(10^\circ, 10^\circ)$. The conventional solutions to multi-label identification problem and multi-label encoding techniques do not model this similarity between the ground truths, because of the implicit assumption of "incomparability" across categorical labels.

A common approach to leveraging the similarity of the ground truth labels is to model the problem as a regression task. In that case, the DNN model can take the processed sensor data as input and predict the directions of all channel paths. A typical mmWave channel, however, can have a variable number of channel paths, while a fixed-size DNN can not handle a variable number of outputs.

This motivates us to design an encoding technique for the AoD-list of a channel that *(i)* is an extension of multi-label encoding to allow a variable number of AoDs to be predicted, and *(ii)* can capture the notion of similarity across channels.

2) Encoding techniques: We define two types of encoding for the list of AoDs of channel paths: *(a)* hard encoding, *(b)* soft encoding. See Fig. 3 for a reference.

a) Hard encoding: We define the AoD-pair as a label for the multi-label encoding. Let $(\theta_{i,\ell}, \phi_{i,\ell})$ denote the azimuthal and elevation angles (in degrees) of the ℓ-th path associated with the channel of sample i. Let Q be a quantization map such that $Q(\theta, \phi) = (\theta^{\mathrm{Q}}, \phi^{\mathrm{Q}})$ is an index of (θ, ϕ) quantized on a uniform grid of size (Θ, Φ). We denote the hard encoding of the AoD-list associated to the sample i as $\mathbf{y}_i \in \{0,1\}^{\Theta \times \Phi}$ where $\mathbf{y}_i[\theta_{i,\ell}^{\mathrm{Q}}, \phi_{i,\ell}^{\mathrm{Q}}] = 1, \forall \ell$ such that $Q(\theta_{i,\ell}, \phi_{i,\ell}) = (\theta_{i,\ell}^{\mathrm{Q}}, \phi_{i,\ell}^{\mathrm{Q}})$.

b) Soft encoding: We define soft encoding as a perturbation on hard encoding (see Fig. 3). Formally, considering a perturbation range of $\pm\frac{\Delta}{2}^\circ$ on hard encoding, we define the soft encoding of the AoD-list associated to the sample i as

$$\tilde{\mathbf{y}}_i[\theta^{\mathrm{Q}}, \phi^{\mathrm{Q}}] = 1 - \min_{(\ell,\theta,\phi):Q(\theta,\phi)=(\theta^{\mathrm{Q}},\phi^{\mathrm{Q}})} \max\left\{\frac{\mathrm{ADiff}(\theta, \theta_{i,\ell})}{\Delta/2}, \frac{\mathrm{ADiff}(\phi, \phi_{i,\ell})}{\Delta/2}\right\}. \tag{8}$$

This perturbation adds non-zero, linearly decaying values to angles close to the directions of the channel paths.

Based on the soft encoding of AoD-list, we measure the similarity between the channels associated to samples i, j as

$$\rho(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j) = \frac{\tilde{\mathbf{y}}_i^{\mathrm{T}} \tilde{\mathbf{y}}_j}{||\tilde{\mathbf{y}}_i||_2 ||\tilde{\mathbf{y}}_j||_2}. \tag{9}$$

The proposed metric measures the overlap of channel path directions within the defined perturbation range. Higher overlap

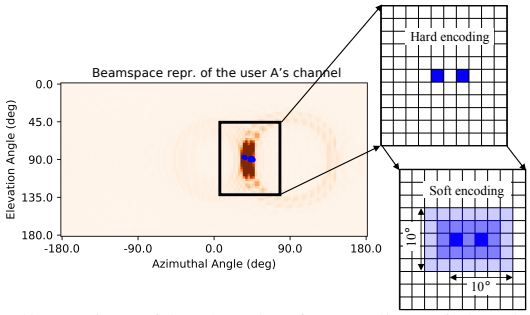


Fig. 3: Illustration of hard and soft encoding of the channel's AoD-list: The blue squares represent a value of 1, indicating the presence of a channel path along those specific directions. The soft encoding incorporates a perturbation range of $\Delta = 10^\circ$ in both elevation and azimuthal directions. The value within each square decays linearly from the nearest square containing the actual channel path direction. While the grid is shown for a limited area for illustrative purposes, it would encompass all azimuthal and elevation angle pairs.

in the soft encoding of the AoD-list of two samples within the perturbation range translates to a higher similarity score. Additionally, the normalization in the metric ensures that both channels are weighted equally regardless of the number of paths they contain. This also guarantees that if one channel encoding $\tilde{\mathbf{y}}_i$ has more paths, the other channel $\tilde{\mathbf{y}}_j$ would require a higher number of paths in similar directions to get considered as highly similar.

The hard and soft encodings capture different aspects of the channel's AoD-list. Hard encoding defines a clear boundary between valid and invalid channel directions for a specific sample, while the soft encoding incorporates the notion of similarity between different channels even if they are not identical. In this paper, we leverage both these aspects in DNN training, employing different loss functions for each encoding, as detailed in Section III-D.

We can further extend the proposed encoding methods beyond the current definition of the index for hard and soft encodings, which represent angle pairs (θ, ϕ) (referred to as angle-based encodings). As an alternative, we can define hard- and soft-encoded vectors indexed by quantized triplets of cosines $(\sin(\theta), \cos(\theta), \sin(\phi))$ (cosine-based encodings). This approach introduces a non-uniform quantization of angles compared to the uniform quantization used in angle-based encodings. Intuitively, cosine-based encoding should be a better choice for training since precoder design depends on the cosines of angles and not the exact angles themselves. Accordingly, we evaluate the performance of both angle-based and cosine-based encodings during the training and testing phases of the DNN model.

Note that the soft encoding in itself is not enough for the DNN model to learn the similarity/dissimilarity between channels. We also require an appropriate loss function for training DNN, which we propose in Section III-D.

C. Multimodal sensor fusion: Design of DNN

In this subsection, we describe the DNN-based multimodal sensor fusion model, which leverages the pre-processed sensor

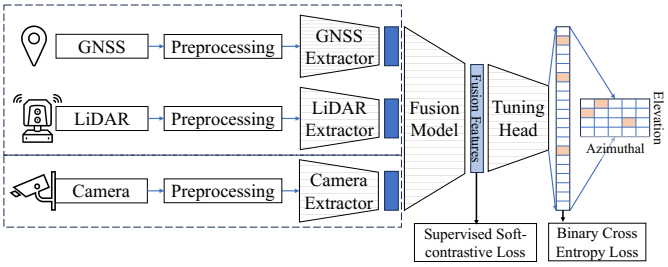


Fig. 4: Schematic diagram of the proposed DNN-based multimodal fusion network

data and hard- and soft-encoded AoD-lists of the associated channels. We first outline the schematic of our model, containing unimodal feature extractors and fusion network. We then describe the two-part structure of the fusion network – the key novelty of our model.

Our model draws inspiration from [31]. We, however, modify the hyper-parameters near the network's head to accommodate the larger output dimensions required for our task. Similar to [31], our model operates in two stages as illustrated in Fig. 4:

1) The pre-processed data from each modality is passed through its corresponding *unimodal feature extractor* (detailed in Fig. 5). These extractors generate modality-specific features.
2) The features extracted from each modality are then concatenated and fed as the input to the *fusion network* (detailed in Fig. 5).

The key novelty in our model is the two-part structure of the fusion network. Specifically, the first part, called *fusion feature extractor*, takes the concatenated features as an input and generates *fusion features* of dimension 512. We denote the fusion features associated with the i-th sensor data sample by $\hat{\mathbf{z}}_i$. These fusion features are used to incentivize the unimodal feature extractors and the fusion network to output similar fusion features for similar channels. We describe the incentive in Section III-D.

The second part of the fusion network, called *tuning head*, uses the fusion features $\hat{\mathbf{z}}_i$ to predict the hard-encoded AoD-list $\mathbf{y}_i$ in the form of probabilities denoted by $\hat{\mathbf{y}}_i$. We consider a value greater than 0.5 in $\hat{\mathbf{y}}_i$ as a positive prediction. Finally, for each sample i associated to UE u and time-slot n, the estimated $\hat{\mathbf{y}}_i$ is used to obtain the beamspace representation of the channel $\hat{\tilde{\mathbf{G}}}_u^{(n)}$.

D. Loss functions and evaluation metrics

In this subsection, we describe the loss functions used to train the entire multimodal fusion model, and the metrics used to evaluate its performance.

1) Binary cross entropy (BCE) loss: BCE loss is used during training to minimize the disparity between the predicted AoD-list $\hat{\mathbf{y}}_i$ and the ground truth represented by the hard-encoded AoD-list $\mathbf{y}_i$. Mathematically, for a batch of sensor observations denoted by $\mathcal{B}$, the BCE loss is defined as

$$\text{BCE}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(\mathbf{y}_i^{\mathrm{T}} \log(\hat{\mathbf{y}}_i) + (\mathbf{1} - \mathbf{y}_i^{\mathrm{T}}) \log(1 - \hat{\mathbf{y}}_i) \right). \tag{10}$$

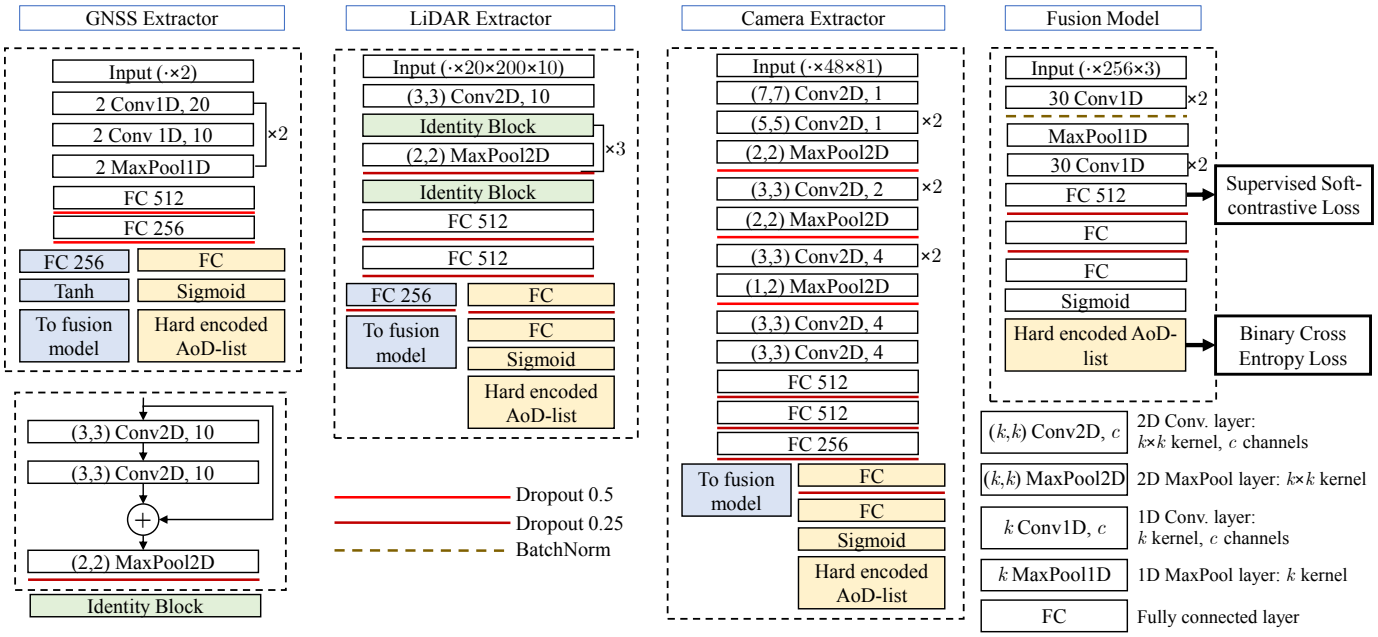


Fig. 5: Detailed block diagrams of unimodal feature extractors and the fusion network: The results from the *yellow*-shaded branches of the unimodal extractors are used to predict the AoD-list of a channel from the respective sensor modality. The output of the *blue*-shaded branches of the unimodal extractors are used to feed the fusion model.

2) Supervised soft-contrastive loss (SSCL): SSCL is used during training to incentivize the DNN to learn the notion of similarity between channels by encouraging to extract similar features from the fusion of sensor data for similar channels. For that, we introduce two objectives on the fusion features $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{z}}_j$ associated to sensor observations i and j: *(i)* Aligning the fusion features $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{z}}_j$ when the corresponding sensor observations i and j have similar channels, and *(ii)* dis-aligning the fusion features $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{z}}_j$ when the observations exhibit considerably different channels. This allows robust training of DNN against the noise in the sensor observations, and further reducing the amount of data required for the training [46].

To formally define SSCL, recall the similarity metric $\rho_{i,j} := \rho(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j)$ defined in (9). Then, for a batch of sensor observations $\mathcal{B}$, we define SSCL as

$$\text{SSCL}(\hat{\mathbf{z}}, \tilde{\mathbf{y}}) = \frac{1}{|\mathcal{B}|^2} \sum_{i,j} \Bigg[-\rho_{i,j} \exp\left(\frac{\hat{\mathbf{z}}_i^{\mathrm{T}} \hat{\mathbf{z}}_j}{||\hat{\mathbf{z}}_i||_2 ||\hat{\mathbf{z}}_j||_2} \right) + (1-\rho_{i,j}) \exp\left(\frac{\hat{\mathbf{z}}_i^{\mathrm{T}} \hat{\mathbf{z}}_j}{||\hat{\mathbf{z}}_i||_2 ||\hat{\mathbf{z}}_j||_2} \right) \Bigg], \tag{11}$$

$$= \frac{1}{|\mathcal{B}|^2} \sum_{i,j} (1 - 2\rho_{i,j}) \exp\left(\frac{\hat{\mathbf{z}}_i^{\mathrm{T}} \hat{\mathbf{z}}_j}{||\hat{\mathbf{z}}_i||_2 ||\hat{\mathbf{z}}_j||_2} \right). \tag{12}$$

The first and second parts of (11) incentivize the fusion model to align and dis-align the fusion features depending on the similarity and dissimilarity of beamspaces in the batch. Not only that, the incentives for alignment and dis-alignments are proportional to the similarity and dissimilarity metrics $\rho_{i,j}$, $(1-\rho_{i,j})$, respectively. Thus, SSCL proposed in (12) balances the trade-offs between alignment and dis-alignment of fusion features.

The proposed SSCL differs from SCL presented in [46] in how similarity is weighted. Our approach leverages a continuous metric $\rho(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j)$ to weigh the alignment/disalignment of fusion features, whereas the original SCL relies on a binary similarity based on class labels. This fundamental difference necessitates modifications to the original SCL formulation. Due to the continuous similarity metric as a weight, the multiplicative dis-alignment factor in the denominator of SCL in [46] can not be ensured to be normalized. Thus, we use an additive dis-alignment factor as in (11) instead of a multiplicative factor as in SCL [46].

Finally, based on the ground truth encodings $(\mathbf{y}, \tilde{\mathbf{y}})$, we use the sum of $\text{BCE}(\hat{\mathbf{y}}, \mathbf{y})$ and $\text{SSCL}(\hat{\mathbf{z}}, \tilde{\mathbf{y}})$, as the loss function to train the DNN.

3) Mean angular distance (MAD) metric: MAD is used during evaluation to quantify the average discrepancy between the predicted and actual AoDs. To calculate the MAD, we follow four steps: First, for each predicted AoD pair, we identify the nearest AoD pair in the ground truth AoD-list. Second, we compute the angular distance between the true and predicted AoD pairs. The angular distance is defined as the inverse cosine of cosine distance between the unit vectors along predicted and true AoD pairs. Third, we average the measured distances across all predicted AoDs within a sample. Finally, we calculate the average of the sample-wise MAD across all samples in a batch.

4) Mean absolute error (MAE) in cosines metric: Since the antenna array response vectors depend on the cosine-triplet of AoD-pairs, defined as $(\sin(\theta), \cos(\theta), \sin(\phi))$, MAE in cosines is used during evaluation to quantify the discrepancy between the predicted and actual cosine-triplets. The calculation of MAE in cosines is similar to that of MAD; however, instead of calculating the angular distance between AoD pairs, we calculate the L_1-distance between the cosine-triplets of

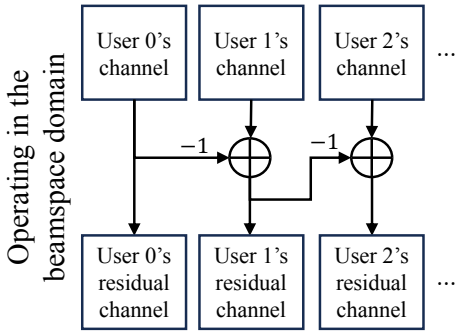


Fig. 6: Block diagram for determining residual channels of users based on beamspace estimates: The BS iteratively removes the channel paths that overlap with other UEs to get the residual channels. These residual channels of UEs are used to find the RF precoder $\mathbf{F}_{\mathrm{RF}}$.

predicted and true AoD pairs.

Remark 1. *The number of paths can vary across different samples. Consequently, MAD and MAE-based metrics, which depend on predicted angles, are not suitable as loss functions for DNN training due to challenges with gradient-based back-propagation. Nevertheless, they remain valuable for assessing the DNN's performance in terms relevant to the practical application.*

With the specified design of the network and the loss functions, we have the first stage of the sensor-aided MU beamforming: A model to predict the beamspace representation $\hat{\tilde{\mathbf{G}}}_u$ from sensor observations. We report the performance of training and testing with various sensor modalities in Section V-D. In the following, we move on to the second stage, MU beamforming strategy using the predicted beamspaces.

IV. Stage II: Beamspace-based MU beamforming algorithm

In this section, we describe the MU beamforming algorithm that uses the beamspace representations of the users' channels. Given the estimate of the beamspace from the multimodal sensor fusion network, we first introduce the intuition behind the MU beamforming algorithm followed by the outline.

A. Key intuition behind the MU beamforming algorithm

The algorithm is split into two phases. The first phase focuses on designing RF precoder from the beamspace representations (see Fig. 6). The key idea in this phase is to assign one transmission direction to at most one UE. For instance, consider a scenario where the BS communicates with two UEs, A, B. The beamspaces of their channels are shown in Fig. 7a and 7b, respectively. Notice that both UEs have an overlapping multipath component at AoD-pair $(45°, 90°)$. The BS starts with the UE having the least number of predicted paths (in this case, UE A), selects one direction from its estimated beamspace, and removes that direction from the beamspaces of the subsequent users. In this case, for the UE A, the algorithm selects the (only) direction from its beamspace and removes that direction from the beamspace of UE B. We call the resultant beamspaces, the residual channels of UEs. The BS then designs the RF precoders, $\mathbf{f}_A^{\mathrm{RF}}, \mathbf{f}_B^{\mathrm{RF}}$, that maximize the transmission energy along the peak of the UEs' residual channels. This process allows for assigning only one UE to each cluster of the channel.

In the second phase of the algorithm, the BS optimizes the digital precoder $\mathbf{F}_{\mathrm{BB}}$ using conventional MU-MISO strategies based on the estimation of equivalent channel and channel feedback. Since the dimension of $\mathbf{F}_{\mathrm{BB}}$ is significantly smaller than $N_{\mathrm{BS}}^{\mathrm{x}} \times N_{\mathrm{BS}}^{\mathrm{y}}$, this phase contributes minimally to the overall overhead.

Algorithm 1 MU beamforming using the beamspace representations (detailed in Section IV-B)

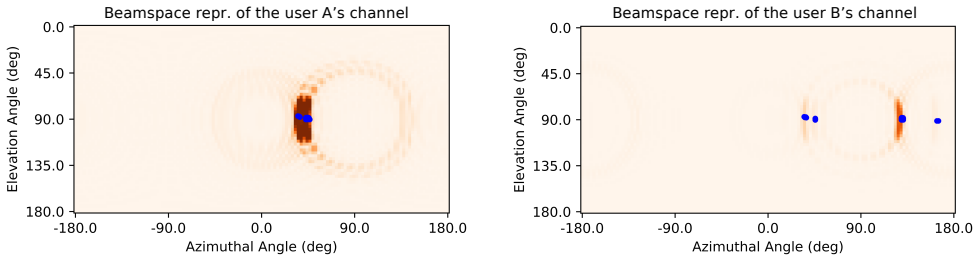
Input: An ordered list of UEs $\mathcal{U} = (u_1, \ldots, u_U)$ to be communicated, sorted in the ascending order of the number of the predicted paths.
Input: Beamspace estimate from sensors $\hat{\tilde{\mathbf{G}}}_u, \forall u \in \mathcal{U}$
First stage: Determining $\mathbf{F}_{\mathrm{RF}}$
for $i = 1, 2, \ldots, U$ **do**
  Estimate residual beamspace $\check{\mathbf{G}}_i = \hat{\tilde{\mathbf{G}}}_i - \sum_{j=0}^{i-1} \hat{\tilde{\mathbf{G}}}_j$
  Find a peak direction in $\check{\mathbf{G}}_i$, denoted by $(\hat{\theta}_i, \hat{\phi}_i)$.
  Design a directional beamformer $\mathbf{f}_i^{\mathrm{RF}} = \mathrm{vec}(\bar{\mathbf{A}}(\hat{\theta}_i, \hat{\phi}_i))$.
end for
Set $\mathbf{F}_{\mathrm{RF}} = [\mathbf{f}_1^{\mathrm{RF}}, \cdots, \mathbf{f}_{N_{\mathrm{RF}}}^{\mathrm{RF}}]$.
Second stage: Determining $\mathbf{F}_{\mathrm{BB}}$
BS sets $\mathbf{F}_{\mathrm{BB}}[k] = \mathbf{I}_{N_{\mathrm{RF}}}$ and transmit training sequence to all UEs simultaneously.
UEs estimate the effective channels $\hat{\mathbf{h}}_u^{\mathrm{T}}[k] = \mathbf{h}_u^{\mathrm{T}}[k]\mathbf{F}_{\mathrm{RF}}$ and feed them back to BS.
Using the effective channels, the BS designs RZF digital precoder $\mathbf{F}_{\mathrm{BB}}[k]$ according to (13).

B. Algorithm outline

The pseudocode of the algorithm is presented in Algorithm 1. Consider the $U = N_{\mathrm{S}} = N_{\mathrm{RF}}$ number of active UEs. In the first phase, the BS receives the predicted AoD-list and the associated beamspace estimates $\hat{\tilde{\mathbf{G}}}_u$ of all UEs from the fusion network. The BS then sorts UEs in the ascending order of the number of predicted paths. Then for each UE i, the BS creates the residual beamspace representation, denoted by $\check{\mathbf{G}}_i$, by subtracting the residual beamspaces of all UEs $j < i$. It then identifies the directions $(\hat{\theta}_i, \hat{\phi}_i)$ associated with the peak of $\check{\mathbf{G}}_i$, and selects $\mathbf{f}_i^{\mathrm{RF}} = \mathrm{vec}(\bar{\mathbf{A}}(\hat{\theta}_i, \hat{\phi}_i))$ as the RF precoder for UE i. This process is repeated until all UEs are assigned an RF precoder.

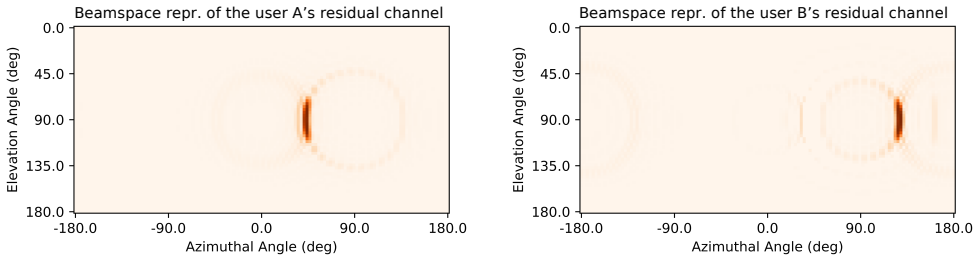
Once $\mathbf{F}_{\mathrm{RF}}$ is determined, the BS then optimizes $\mathbf{F}_{\mathrm{BB}}$ in the second phase. This process is similar to the procedure used in preamble-based MU-MIMO channel estimation [37], [63], [64]. The BS transmits a common preamble sequence to all UEs. Each UE estimates its effective channel $\hat{\mathbf{h}}_u^{\mathrm{T}}[k] = \mathbf{h}_u^{\mathrm{T}}[k]\mathbf{F}_{\mathrm{RF}}$, and feeds it back to the BS through the control channel. Finally, the BS collects the estimated channels from all UEs, and designs the RZF digital precoder $\mathbf{F}_{\mathrm{BB}}[k]$ as follows

$$\mathbf{F}_{\mathrm{BB}}[k] = \frac{\left(\sum_{u'=1}^{U} \bar{\hat{\mathbf{h}}}_{u'}[k]\hat{\mathbf{h}}_{u'}^{\mathrm{T}}[k] + \mathbf{I}_{N_{\mathrm{RF}}}\right)^{-1} \bar{\hat{\mathbf{h}}}_u[k]}{\left\| \left(\sum_{u'=1}^{U} \bar{\hat{\mathbf{h}}}_{u'}[k]\hat{\mathbf{h}}_{u'}^{\mathrm{T}}[k] + \mathbf{I}_{N_{\mathrm{RF}}}\right)^{-1} \bar{\hat{\mathbf{h}}}_u[k] \right\|_2}. \quad (13)$$



(a) Beamspace representation of user A's channel

(b) Beamspace representation of user B's channel



(c) Beamspace representation of user A's residual channel

(d) Beamspace representation of user B's residual channel

Fig. 7: Intuition behind MU beamforming algorithm: The overlapping directions in the beamspace representation of user A and user B's channels are removed from the beamspace of user B. User A gets the priority because it has fewer AoDs. The resultant beamspace representations are called residual channels and are used to determine the initial RF precoder $\mathbf{F}_{\mathrm{RF}}$.

With the chosen precoders $\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}[k], \forall k \in [K]$, the BS can transmit to all UEs simultaneously.

Remark 2. *While our approach and traditional parametric channel estimation methods [51]–[54] both aim to estimate AoDs, there is key difference. Conventional parametric channel estimation relies on training sequences and, thus, requires an initial beam selection to ensure adequate signal-to-noise ratio (SNR) for accurate parameter estimation. In contrast, our method uses out-of-band sensors to predict AoDs, effectively eliminating the need for expensive beam search and reducing overall system overhead. This distinction highlights that our sensor-aided approach offers a more efficient solution compared to traditional parametric channel estimation approaches.*

Remark 3. *Using path gains in the precoder design: We emphasize here that while our approach does not involve predicting path gains from sensor data and does not use path gains for designing analog precoders in the first stage, it uses the path gains for designing the digital precoder in the second stage. Specifically, the path gains are embedded in the estimated effective channels of the UEs.*

While using path gains in the first stage could enhance analog precoder design, it would require predicting these gains purely from sensor data. We believe that predicting path gains from sensor data presents a promising area for future research.

V. Numerical Results

In this section, we present numerical evaluation of the proposed sensor-aided MU beamforming strategy. We first describe the dataset used for evaluation, followed by a comparison of the achieved MU SE R_{S} with the baselines. We then discuss the impact of using different sensing modalities on the performance and compare the performance of MU communication with the proposed beamforming strategy to a time-division multiplexed SU strategy. Finally, we quantify the reduction in the link establishment overhead due to the sensor-aided beamforming approach.

A. Dataset

To evaluate the performance of the sensor-aided beamspace prediction model, we use the Raymobtime dataset [27]. This dataset comprises a collection of synthetic ray tracing datasets designed for simulating wireless channels. We specifically employ datasets `s008` and `s009`. These datasets offer a rich combination of multimodal sensing data, including LiDAR, camera images, and coordinates, alongside ray tracing-based channels at 60 GHz for all ten mobile receivers, in a synthetic 3D scenario of the city Rosslyn, Virginia. These datasets capture 2086 and 2000 scenes, respectively, with each scene separated by 30 seconds.

Crucially, `s008` and `s009` datasets include the underlying mmWave channel associated with each sensing data sample. This valuable feature allows training and testing of the proposed prediction model using the sensor data as input and the hard- and soft-encoded beamspace estimates of the underlying mmWave channels as output. In contrast, other real-world multimodal sensing datasets such as DeepSense 6G [29] and e-FLASH [31], used for mmWave beam predictions, only contain the mapping of the beams in the codebook and measurements of the received powers, but not the complete CSI. This limits their use in our work.

Similar to previous work [23], [31], we use datasets `s008` for training our fusion model, while `s009` is used for validating the obtained results.

B. System setup and parameters for DNN

Our system uses a BS equipped with 2×1 RF chains with 16×8 antenna elements for $U = 2$, and $\frac{U}{2} \times 2$ RF chains with $4U \times 16$ antenna elements for $U = \{4, 6, 8\}$. The granularity for the beamspace representation is $G_\theta \times G_\phi = 64 \times 32$ for $U = 2$ and $G_\theta \times G_\phi = 16U \times 64$ for $U = \{4, 6, 8\}$. We consider a total of $\mathcal{L} = 25$ strongest paths from all ray clusters combined to define the *truncated* beamspace of the channel. Following 5G NR specifications, we use OFDM symbols with 792 subcarriers spaced at 120 kHz [65]. We consider all users transmit simultaneously on all subcarriers and do not consider resource allocation over the resource grid.

For LiDAR point clouds preprocessing, we define the coverage space of BS as $(X_{\min}, X_{\max}) = (744\ \text{m}, 767\ \text{m})$, $(Y_{\min}, Y_{\max}) = (429\ \text{m}, 679\ \text{m})$, $(Z_{\min}, Z_{\max}) = (0\ \text{m}, 10\ \text{m})$. Any points outside this range are removed from the point cloud. The remaining data is then quantized into the grid of $b^{\mathcal{L}}_{\text{x}} \times b^{\mathcal{L}}_{\text{y}} \times b^{\mathcal{L}}_{\text{z}} = 20 \times 200 \times 10$ bins in X, Y and Z directions. The camera images are scaled to the size $b^{\mathcal{C}}_{\text{h}} \times b^{\mathcal{C}}_{\text{w}} = 48 \times 81$.

Finally, for the angle-based encoding of beamspaces, we uniformly quantize the azimuthal and elevation angle pairs into $\Theta \times \Phi = 90 \times 45$ bins. For cosine-based encoding, we uniformly quantize the values of $(\sin\theta, \cos\theta, \sin\phi)$ in $40 \times 40 \times 40$ bins. Unless otherwise mentioned, we use angle-based encoding for the subsequent analysis. For soft-encoding, we consider the perturbation range of $\Delta = 10$.

For the training of the DNN model, we use both BCE and SSCL losses with equal weights. We use the batch size of 32. We consider an initial learning rate of 0.0001 and reduce it by a factor of 0.99 when the total training loss has stopped improving for 10 consecutive epochs.

C. MU spectral efficiency

We evaluate the performance of our complete framework on `s009` dataset. We consider simultaneous transmission to 2, 4, 6, and 8 mobile UEs out of the 10 available in each episode. For each episode, we consider all possible combinations of UEs called *user-clusters*. For each cluster, we use all available sensing modalities – camera images, LiDAR point clouds, and coordinates of the UEs – to predict the beamspace of the channel. We then use the MU beamforming algorithm to determine the RF precoder $\mathbf{F}_{\text{RF}}$ and digital precoder $\mathbf{F}_{\text{BB}}$ and calculate SE R_u achieved by each user u in the user-cluster. We then select the user-clusters that provide the maximum SE for each user. Finally, we plot the median and the 25th and 75th percentile values of the sum-SE of all such user-clusters across all episodes.

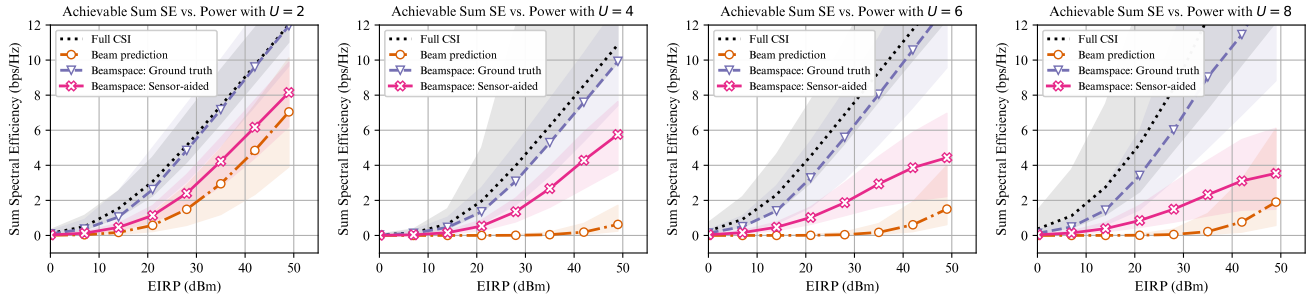
We compare our strategy against three baselines:

- Full CSI at the BS: In this scenario, the BS designs $\mathbf{F}_{\text{BB}}$ assuming noise-less channel estimate and a fully digital antenna array architecture. This represents a best case scenario for a MU-MIMO communication system.
- Beam prediction: The BS uses the optimal beam for each UE individually from a DFT codebook. This baseline represents the best case performance of the state-of-the-art multimodal sensor-aided mmWave beam training framework [31] when extended to MU setting without prediction error.
- Ground truth beamspace: The BS has access to the true truncated beamspace representation, as opposed to the sensor-aided estimate. This baseline demonstrates the effectiveness of our MU beamforming algorithm, isolating the impact of errors in the sensor-aided beamspace estimates.

It is important to note that noise affects the design of $\mathbf{F}_{\text{BB}}$ in our framework and the baselines except in the full CSI setting, as it requires estimating the channel using the preamble-based MU-MIMO channel estimation method. We also emphasize here that we do not consider the overhead of searching the optimal beam from the top-K prediction, nor do we consider the overhead of channel estimation. Our focus for the analysis is to highlight the sum-SE performance of the proposed strategy compared to the baseline. Therefore, we assume the access to an ideal beam predictor that selects the best beam and the access to noisy channel estimate without any overhead. We compare the overhead of the proposed strategy with baselines in Section V-F.

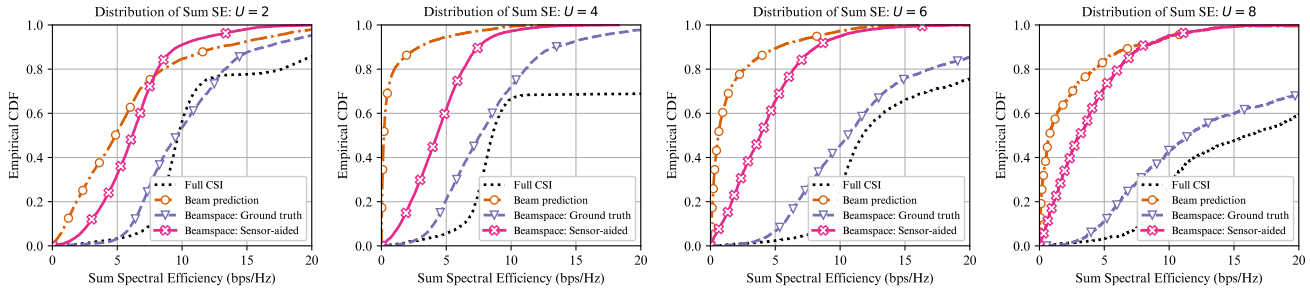
Fig. 8 shows the median sum-SE with 25% and 75%-tile range achieved in the MU setting with both $\mathbf{F}_{\text{RF}}$ and $\mathbf{F}_{\text{BB}}$ derived using Algorithm 1. This demonstrates the end-to-end performance of combining our proposed DNN-based multimodal fusion network with MU beamforming strategy. We observe consistently high performance using beamspace-aided MU beamforming compared to the MU extensions of SU baseline technique. Note that the performance of the beam prediction baseline is significantly worse with a high number of active users, indicating the impact of inter-user interference. In Fig. 9, we show the empirical CDF of achieved sum-SE at 42 dBm EIRP. We observe that the median MU sum-SE of the proposed MU beamforming strategy is significantly higher (up to $4\times$ with $U = 4$) than the MU extension of SU baseline.

Note that the prediction error in sensor-aided beamspace estimate, defined in comparison with ground truth beamspace, only depends on the sensor modalities used during training, and remains constant across different numbers of UEs. From Fig. 9, we observe that there is a larger impact on performance for the same prediction error when using more UEs. This suggests that the proposed algorithm grows more sensitive towards the errors in beamspace prediction as the number of users increases. Notably, in the ideal scenario with known ground truth beamspace, the system performance is close to that of a digital MU-MIMO system with a full CSI setting. This, however, is not the case when using sensor-aided beamspace estimate. The prediction error can introduce misalignments or identify sub-optimal directions of multipath components in the channel. Because this prediction error is independent of transmit power, the resultant loss of dominant directions has a proportionally high impact for higher transmit power. Therefore, we believe that further optimization of the multimodal fusion network has the potential to improve the



(a) $U = 2$ (b) $U = 4$ (c) $U = 6$ (d) $U = 8$

Fig. 8: MU sum-SE vs. power for $U = 2, 4, 6, 8$: The lines in the plots denote the median sum-SE while the shaded regions show the respective 25%-tile to 75%-tile range.



(a) $U = 2$ (b) $U = 4$ (c) $U = 6$ (d) $U = 8$

Fig. 9: Distribution of MU sum-SE for $U = 2, 4, 6, 8$: beamspace-aided beamforming achieves much higher sum-SE compared to the beam prediction baseline at 42 dBm EIRP.

SE performance of the proposed MU framework.

Finally, we also observe an interesting feature of the Raymobtime dataset from the performance of full CSI in Fig. 9. We see two distinct groups of user clusters. The top 20% achieve a sum SE greater than 18 bps/Hz, while the bottom 60% have a sum-SE below 9 bps/Hz. Since we only focus on user clusters offering the best individual SE to at least one user, this sharp difference suggests many users have inherently low SE regardless of other users in their cluster. We believe this low SE stems from similar AoDs in the channels of these users. This similarity results in severe inter-user interference, even with a fully digital antenna system.

D. Impact of sensing modalities

In this section, we analyze the impact of various sensing modalities on the end-to-end performance of our framework.

We consider all possible combinations of sensor information. In the case of unimodal setting, we use the individual sensor feature extractors as described in Fig. 5 to predict the beamspace of the channel. In the case of multimodal setting, we use the multimodal fusion network to predict the beamspace of the channel. For each case, we train the model using the dataset `s008` using angle-based encoding and cosine-based encoding of beamspaces. We then evaluate the performance of the trained models on the test dataset `s009`. The test results, in terms of MAD and MAE in cosines, are reported in Table I.

We observe that the cosine-based encoding performs significantly worse compared to angle-based encoding. This is due to the huge network size to handle 64000-dimensional output when using cosine-based encoding as opposed to 7200-dimensional encoding when using angle-based encoding.

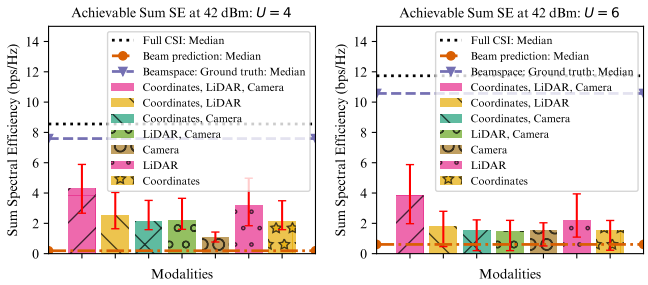
We further use these trained models to estimate the sum-SE of the system. Fig. 10 shows the MU sum-SE as a function of EIRP. Compared to the baseline using ground truth-based beamspace representation, we observe that the combination of coordinates, LiDAR, and camera performs better than other combinations. Notably only using camera images at the BS leads to lower performance compared to the ground-truth baseline which suggests the sensors at the UEs play significant role in predicting the information about the channel.

E. Impact of SSCL loss

In Fig. 11, we show the impact of SSCL loss on the sum-SE of MU communication with $U = 4$ and 6. We observe that the SSCL loss is a key reason for the success of the proposed multimodal fusion network in predicting the beamspace representation. Since SSCL incentivizes feature similarities and dis-similarities during the training of the fusion network, we can recover key features relevant for beamspace estimation from the sensors, and therefore, achieve higher sum-SE.

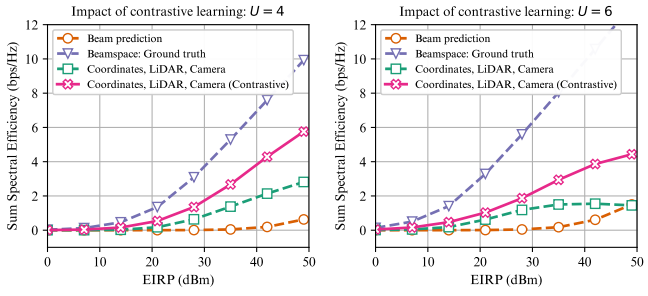
Sensor modalities	With angle-based encoding		With cosine-based encoding	
	MAD (deg)	MAE in cosine	MAD (deg)	MAE in cosine
Coordinates	7.2821	0.1258	15.2863	0.1380
Image	7.2821	0.1258	15.2863	0.1380
LiDAR	5.8630	0.1101	17.3852	0.1830
LiDAR + Image	8.0177	0.1419	17.8200	0.2041
Coordinates + LiDAR	6.2575	0.1179	15.9807	0.1867
Coordinates + Image	7.2821	0.1258	15.6233	0.1567
Coordinates + LiDAR + Image	6.1026	0.1149	17.3852	0.1830

TABLE I: Test error on `s009` dataset across modalities: MAD and MAE in cosines are defined as in Section III-D. Due to a huge network size, the cosine-based encoding performs significantly worse compared to angle-based encoding.



(a) $U = 4$ (b) $U = 6$

Fig. 10: Comparison of MU SE using different combinations of sensor modalities at 42 dBm EIRP: The error bars show 25%-tile to 75%-tile range. The combination of coordinates, LiDAR, and camera performs better than other combinations. Notably only using camera images at the BS leads to lower performance compared to the ground-truth baseline which suggests the sensors at the UEs play a significant role in predicting the information about the channel.



(a) $U = 4$ (b) $U = 6$

Fig. 11: Defining the impact of SSCL loss: By contrasting samples based on their similarity score can help recover the relevant features from the sensors for the beamspace prediction, which, in turn, helps with improving achievable SE using beamspace-aided MU beamforming strategy.

F. Overhead analysis

To further highlight the benefit of our two-stage approach (sensor-aided beamspace estimation followed by beamspace-aided beamforming) in contrast with the one-stage sensor-aided beam prediction strategy, we examine the overhead of both strategies.

1) Beamforming overhead: Our method leverages sensor data to determine the RF precoder for all users. Consequently, the first phase of Algorithm 1 does not introduce additional overhead to the communication system. In the second stage, digital precoders are identified by estimating the effective channels of all UEs through the transmission of SSBs as specified in the 5G NR standard for mmWave bands. Using the chosen RF precoders, the BS employing sensor-aided MU beamforming requires only a single synchronization signal block (SSB) for all UEs simultaneously. In contrast, a conventional BS without sensor-aided beam prediction must transmit up to 64 SSBs using different RF precoders to determine the optimal RF precoder for each UE. Therefore, sensor-aided MU beamforming can reduce beamforming overhead by up to a factor of 64.

Moreover, the state-of-the-art sensor-aided SU beam prediction strategy requires testing 10 beams (using 10 SSBs) to determine the optimal RF precoder for each UE with 90% accuracy [31]. In comparison, our two-stage approach needs only one SSB for multiple UEs, reducing the overhead by up to a factor of 10. We note that the sum-SE results presented in this paper do not account for beam selection overhead, focusing instead on the maximum achievable sum-SE. Therefore, in practice, the benefits of our two-stage approach for MU mmWave beamforming are likely to be even greater when considering beam selection overhead.

Finally, the beamforming overhead for designing the digital precoder $\mathbf{F}_{\mathrm{BB}}$ is minimal and unavoidable. It is minimal because, for N_{RF} RF chains, the BS only needs to transmit N_{RF} OFDM symbols. With a sub-carrier spacing of 120 kHz for mmWave communication, each OFDM symbol lasts 8.3 μs. For instance, transmitting to 10 UEs using 10 RF chains would require just 0.083 ms, which only decreases with increased sub-carrier spacing. This overhead is unavoidable since the BS must always adjust the digital precoder based on the phase-coherent channel estimate, irrespective of the RF precoder selection process.

2) Computational overhead: Our approach involves two main sources of computational overhead.

a) The DNN-based multimodal fusion network: We compare the computational overhead of our DNN-based fusion network with the benchmark fusion network from [31] in terms of the number of floating point operations (FLOPs). Specifically, our proposed fusion network requires 33.502M FLOPs, which is marginally lower than the 42.947M FLOPs needed by the benchmark network for per-UE beam prediction [31]. Furthermore, the modern low-power GPUs like NVIDIA Jetson Nano [66], can support hundreds of GFLOPs per second. Therefore, the inference using our proposed fusion

network only minimally contributes to the processing latency.

b) The matrix operations in Algorithm 1: The first stage of Algorithm 1, which involves U iterations, each comprising $G_\theta G_\phi$ subtractions. Thus, the computational complexity of the first stage scales as $O(UG_\theta G_\phi)$. The second stage of Algorithm 1 introduces a computational overhead of inverting a $U \times U$ matrix (recall that $N_{\mathrm{RF}} = U$) for calculating the RZF digital precoder, which scales as $O(U^3)$. Therefore, the overall computational complexity of Algorithm 1 scales as $O(UG_\theta G_\phi + U^3)$, with a fixed computation requirement for a given U.

Although our approach introduces additional computational complexity, it is feasible with modern hardware and can be executed in real-time. Importantly, it only needs to be applied at each channel coherence interval (approximately 50 ms), allowing sufficient time for execution.

VI. Conclusion and Future work

In this paper, we addressed the challenge of establishing efficient links in mmWave MU-MIMO systems by leveraging ubiquitous sensor data. Our results demonstrate that sensor data can reveal not just one dominant path, but the directions of numerous significant paths within the channel. This also highlights the potential of out-of-band sensors for extracting rich channel information to aid advance features of communication systems such as MU-MIMO.

Looking ahead, we identify three avenues for future research. Firstly, an analysis is required to validate the use of sensors for predicting the gain of the channel paths. The path gains depend on many factors, including the environment, the relative phases of different paths within a channel cluster, the amount of scattering and diffusion from the reflectors, and even the material types of the reflector. It is not clear if the sensor data, such as GPS, LiDAR, and camera may provide the precision needed to accurately infer these factors. Nevertheless, if the path gains can be estimated from the sensor data, it can allow prioritizing users based on these gains rather than on the number of channel clusters. Accurately predicting channel path gains also holds the potential to facilitate the allocation of varying numbers of data streams and adaptable power levels to each user.

Secondly, realizing the real-world implementation of our approach demands an evaluation on real-world datasets containing complete channel measurements. Existing real-world multimodal sensor datasets like DeepSense 6G and e-FLASH, unfortunately, lack complete channel measurements. Hence, experimental campaigns are required to gather multimodal sensor data alongside comprehensive channel measurements.

Lastly, for broader deployment of this framework, the adoption of generalizable DNN training alongside site-specific model tuning methodologies needs to be considered. Recent advancements in model-agnostic ML and transfer learning paradigms present promising avenues for addressing these challenges. Overcoming these obstacles, in conjunction with our proposed method, sets the stage for the development of more efficient and resilient MU mmWave communication systems.

References

[1] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. Elsevier Science, Oct. 2020.

[2] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 2019.

[3] R. M. Dreifuerst and R. W. Heath Jr., "Massive MIMO in 5G: How beamforming, codebooks, and feedback enable larger arrays," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 18–23, 2023.

[4] Y. Ghasempour, C. R. C. M. da Silva, C. Cordeiro, and E. W. Knightly, "IEEE 802.11ay: Next-generation 60 GHz communication for 100 Gb/s Wi-Fi," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 186–192, Dec. 2017.

[5] J. C. Aviles and A. Kouki, "Position-aided mm-wave beam training under NLOS conditions," *IEEE Access*, vol. 4, pp. 8703–8714, 2016.

[6] N. González-Prelcic, A. Ali, V. Va, and R. W. Heath Jr., "Millimeter-wave communication with out-of-band information," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 140–146, Dec. 2017.

[7] A. Loch, A. Asadi, G. H. Sim, J. Widmer, and M. Hollick, "mmWave on wheels: Practical 60 GHz vehicular communication without beam training," in *2017 Int. Conf. Commun. Syst. and Netw. (COMSNETS)*, 2017, pp. 1–8.

[8] J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Position aided beam prediction in the real world: How useful GPS locations actually are?" *arXiv preprint arXiv:2205.09054*, 2022.

[9] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath Jr., "Inverse multipath fingerprinting for millimeter wave V2I beam alignment," *IEEE Trans. Veh. Technol.*, Dec. 2017.

[10] A. Ali, N. González-Prelcic, and A. Ghosh, "Passive radar at the roadside unit to configure millimeter wave vehicle-to-infrastructure links," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14 903–14 917, 2020.

[11] N. González-Prelcic, R. Méndez-Rial, and R. W. Heath Jr., "Radar aided beam alignment in mmWave V2I communications supporting antenna diversity," in *2016 Inf. Theory and Appl. Workshop (ITA)*, Feb. 2016.

[12] A. Ali, N. González-Prelcic, R. W. Heath Jr., and A. Ghosh, "Leveraging sensing at the infrastructure for mmWave communication," *IEEE Commun. Mag.*, vol. 58, no. 7, pp. 84–89, 2020.

[13] U. Demirhan and A. Alkhateeb, "Radar aided 6G beam prediction: Deep learning algorithms and real-world demonstration," in *2022 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 2655–2660.

[14] A. Graff, Y. Chen, N. González-Prelcic, and T. Shimizu, "Deep learning-based link configuration for radar-aided multiuser mmWave vehicle-to-infrastructure communication," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 7454–7468, 2023.

[15] Y. Zhang, K. Patel, S. Shakkottai, and R. W. Heath Jr., "Side-information-aided noncoherent beam alignment design for millimeter wave systems," in *Proc. 20th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, ser. Mobihoc '19. Association for Computing Machinery, 2019, pp. 341–350.

[16] M. Alrabeiah, J. Booth, A. Hredzak, and A. Alkhateeb, "ViWi: Vision-aided mmWave beam tracking: Dataset, task, and baseline solutions," *arXiv preprint arXiv:2002.02445*, 2020.

[17] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *2020 IEEE 91st Veh. Technol. Conf. (VTC2020-Spring)*, 2020, pp. 1–5.

[18] B. Salehi, M. Belgiovine, S. G. Sanchez, J. Dy, S. Ioannidis, and K. Chowdhury, "Machine learning on camera images for fast mmwave beamforming," in *2020 IEEE 17th Int. Conf. Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2020, pp. 338–346.

[19] Y. Tian, G. Pan, and M.-S. Alouini, "Applying deep-learning-based computer vision to wireless communications: Methodologies, opportunities, and challenges," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 132–143, 2020.

[20] F. Wen, W. Xu, F. Gao, C. Pan, and G. Liu, "Vision aided environment semantics extraction and its application in mmwave beam selection," *IEEE Commun. Lett.*, vol. 27, no. 7, pp. 1894–1898, 2023.

[21] D. Marasinghe, N. Jayaweera, N. Rajatheva, S. Hakola, T. Koskela *et al.*, "LiDAR aided wireless networks - beam prediction for 5G," in *IEEE 96th Veh. Technol. Conf. (VTC2022-Fall)*, 2022, pp. 1–7.

[22] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Commun. Lett.*, 2022.

[23] A. Klautau, N. González-Prelcic, and R. W. Heath Jr., "LiDAR data for deep learning-based mmWave beam selection," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 909–912, 2019.

[24] T. Woodford, X. Zhang, E. Chai, K. Sundaresan, and A. Khojastepour, "Spacebeam: Lidar-driven one-shot mmwave beam management," in *Proc. 19th Annu. Int. Conf. Mobile Syst., Appl., and Services*, ser. MobiSys '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 389–401.

[25] M. Zecchin, M. B. Mashhadi, M. Jankowski, D. Gündüz, M. Kountouris, and D. Gesbert, "LIDAR and position-aided mmwave beam selection with non-local cnns and curriculum training," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2979–2990, 2022.

[26] M. B. Mashhadi, M. Jankowski, T.-Y. Tung, S. Kobus, and D. Gündüz, "Federated mmwave beam selection utilizing LIDAR data," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2269–2273, 2021.

[27] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath Jr., "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *2018 Inf. Theory and Appl. Workshop (ITA)*, Feb. 2018, pp. 1–9.

[28] J. Gu, B. Salehi, D. Roy, and K. R. Chowdhury, "Multimodality in mmWave MIMO beam selection using deep learning: Datasets and challenges," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 36–41, 2022.

[29] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais *et al.*, "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Commun. Mag.*, 2023.

[30] M. Dias, A. Klautau, N. González-Prelcic, and R. W. Heath Jr., "Position and LIDAR-aided mmWave beam selection using deep learning," in *IEEE 20th Int. Workshop Signal Process. Advances in Wireless Commun. (SPAWC)*, 2019, pp. 1–5.

[31] B. Salehi, G. Reus-Muns, D. Roy, Z. Wang, T. Jian *et al.*, "Deep learning on multimodal sensor data at the wireless edge for vehicular network," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7639–7655, 2022.

[32] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, "Vision-position multi-modal beam prediction using real millimeter wave datasets," in *2022 IEEE Wireless Commun. Netw. Conf. (WCNC)*. IEEE, 2022, pp. 2727–2731.

[33] D. Roy, B. Salehi, S. Banou, S. Mohanti, G. Reus-Muns *et al.*, "Going beyond RF: A survey on how AI-enabled multimodal beamforming will shape the nextG standard," *Comput. Netw.*, p. 109729, Mar. 2023.

[34] G. Charan, A. Hredzak, C. Stoddard, B. Berrey, M. Seth *et al.*, "Towards real-world 6G drone communication: Position and camera aided beam prediction," in *2022 IEEE Global Commun. Conf. (GLOBECOM)*. IEEE, 2022, pp. 2951–2956.

[35] G. Reus-Muns, B. Salehi, D. Roy, T. Jian, Z. Wang *et al.*, "Deep learning on visual and location data for V2I mmWave beamforming," in *2021 17th Int. Conf. Mobility, Sens. and Netw. (MSN)*, 2021, pp. 559–566.

[36] J. Tan, T. H. Luan, W. Guan, Y. Wang, H. Peng *et al.*, "Beam alignment in mmwave V2X communications: A survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 1676–1709, 2024.

[37] R. M. Dreifuerst and R. W. Heath Jr., "Machine learning codebook design for initial access and CSI Type-II feedback in sub-6GHz 5G NR," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[38] N. J. Myers, Y. Wang, N. González-Prelcic, and R. W. Heath Jr., "Deep learning-based beam alignment in Mmwave vehicular networks," in *2020 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 8569–8573.

[39] W. Ma, C. Qi, and G. Y. Li, "Machine learning for beam alignment in millimeter wave massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 875–878, 2020.

[40] A. Molazadeh and M. Ardebilipour, "Deep learning assisted time-varying channel estimation in multi-user mmWave hybrid MIMO systems," *Physical Communication*, vol. 55, p. 101933, 2022.

[41] S. Huang, Y. Ye, and M. Xiao, "Hybrid beamforming for millimeter wave multi-user MIMO systems using learning machine," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1914–1918, 2020.

[42] H. Ahn, I. Orikumhi, J. Kang, H. Park, H. Jwa *et al.*, "Machine learning-based vision-aided beam selection for mmWave multiuser MISO system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1263–1267, 2022.

[43] I. K. Jain, R. Kumar, and S. S. Panwar, "The impact of mobile blockers on millimeter wave cellular systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 854–868, 2019.

[44] U. Demirhan and A. Alkhateeb, "Radar aided proactive blockage prediction in real-world millimeter wave systems," in *IEEE Int. Conf. Commun. (ICC)*, 2022, pp. 4547–4552.

[45] S. Wu, C. Chakrabarti, and A. Alkhateeb, "Proactively predicting dynamic 6g link blockages using LiDAR and in-band signatures," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 392–412, 2023.

[46] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian *et al.*, "Supervised contrastive learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 18 661–18 673, 2020.

[47] C. Balanis, *Antenna Theory: Analysis and Design*. Wiley, 2012.

[48] J. P. González-Coma, J. Rodríguez-Fernández, N. González-Prelcic, L. Castedo, and R. W. Heath Jr., "Channel estimation and hybrid precoding for frequency selective multiuser mmWave MIMO systems," *IEEE J. Sel. Topics in Signal Process.*, vol. 12, no. 2, pp. 353–367, 2018.

[49] R. W. Heath Jr., N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics in Signal Process.*, vol. 10, no. 3, pp. 436–453, 2016.

[50] M. K. Samimi and T. S. Rappaport, "3-D millimeter-wave statistical channel model for 5G wireless system design," *IEEE Trans. Microw. Theory Technol.*, vol. 64, no. 7, pp. 2207–2225, 2016.

[51] J. Rodríguez-Fernández, N. González-Prelcic, K. Venugopal, and R. W. Heath Jr., "Frequency-domain compressive channel estimation for frequency-selective hybrid millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2946–2960, 2018.

[52] X. Gao, L. Dai, S. Han, C.-L. I, and X. Wang, "Reliable beamspace channel estimation for millimeter-wave massive mimo systems with lens antenna array," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6010–6021, 2017.

[53] Z. Gao, L. Dai, and Z. Wang, "Channel estimation for mmwave massive MIMO based access and backhaul in ultra-dense network," in *2016 IEEE Int. Conf. Commun. (ICC)*, 2016, pp. 1–6.

[54] K. Venugopal, A. Alkhateeb, N. González-Prelcic, and R. W. Heath Jr., "Channel estimation for hybrid architecture-based wideband millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1996–2009, Sept 2017.

[55] A. Ali, N. González-Prelcic, and R. W. Heath Jr., "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, 2018.

[56] A. Ali and R. W. Heath Jr., "Compressed beam-selection in millimeterwave systems with out-of-band partial support information," in *2017 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 3499–3503.

[57] A. Ali, N. González-Prelcic, and R. W. Heath Jr., "Spatial covariance estimation for millimeter wave hybrid systems using out-of-band information," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5471–5485, 2019.

[58] V. Va, J. Choi, and R. W. Heath Jr., "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Trans. Veh. Technol.*, Nov. 2015.

[59] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *2018 IEEE/CVF Conf. Comput. Vision and Pattern Recognit.* Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 4490–4499.

[60] U. Frese, "A discussion of simultaneous localization and mapping," *Autonomous Robots*, vol. 20, pp. 25–42, 2006.

[61] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, 2012.

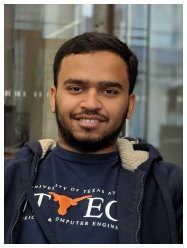
[62] Y. Wang, A. Klautau, M. Ribero, A. C. K. Soong, and R. W. Heath Jr., "Mmwave vehicular beam selection with situational awareness using machine learning," *IEEE Access*, vol. 7, pp. 87 479–87 493, 2019.

[63] A. Alkhateeb, G. Leus, and R. W. Heath Jr., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[64] R. W. Heath Jr. and A. Lozano, *Foundations of MIMO Communication*. Cambridge University Press, 2018.

[65] 3GPP, "User equipment (UE) radio transmission and reception; part 2: Range 2 standalone," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38 101-2, 06 2022, version 17.6.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3284

[66] "NVIDIA Jetson Nano — nvidia.com," https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-nano/product-development/, [Accessed 12-09-2024].



Kartik Patel (S'15 - M'24) received the B.Tech. degree in electronics and communication engineering from the Indian Institute of Technology Roorkee in 2017, the M.S. degree in electrical and computer engineering from the University of Texas at Austin in 2020, and the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2024.

Kartik has held internship positions with the Indian Institute of Science Bengaluru; Cisco Innovation Labs, CA, USA; Qualcomm Inc., CA, USA; and Nokia Bell Laboratories, NJ, USA. His research interests lie at the intersection of wireless networks, sensing, and machine learning with an active focus on system-level validation.

Robert W. Heath Jr. (S'96 - M'01 - SM'06 - F'11) is the Charles Lee Powell Chair in Wireless Communications in the Department of Electrical and Computer Engineering at the University of California, San Diego. He is also President and CEO of MIMO Wireless Inc. From 2020-2023 he was the Lampe Distinguished Professor at North Carolina State University and co-founder of 6GNC. From 2002-2020 he was with The University of Texas at Austin, most recently as Cockrell Family Regents Chair in Engineering and Director of UT SAVES. He authored "Introduction to Wireless Digital Communication" (Prentice Hall, 2017) and "Digital Wireless Communication: Physical Layer Exploration Lab Using the NI USRP" (National Technology and Science Press, 2012), and co-authored "Millimeter Wave Wireless Communications" (Prentice Hall, 2014) and "Foundations of MIMO Communication" (Cambridge University Press, 2018).

Dr. Heath has been a co-author of a number award winning conference and journal papers including recently the 2017 Marconi Prize Paper Award, the 2019 IEEE Communications Society Stephen O. Rice Prize, the 2020 IEEE Signal Processing Society Donald G. Fink Overview Paper Award, the 2021 IEEE Vehicular Technology Society Neal Shepherd Memorial Best Propagation Paper Award, and the 2022 IEEE Vehicular Technology Society Best Vehicular Electronics Paper Award. Other notable awards include the 2017 EURASIP Technical Achievement Award, the 2019 IEEE Kiyo Tomiyasu Award, and the 2021 IEEE Vehicular Technology Society James Evans Avant Garde Award. In 2017, he was selected as a Fellow of the National Academy of Inventors. In 2024, he was selected as a Fellow of the American Association for the Advancement of Science. He was a member-at-large on the IEEE Communications Society Board-of-Governors (2020-2022) and the IEEE Signal Processing Society Board-of-Governors (2016-2018). He was Editor-in-Chief of IEEE Signal Processing Magazine from 2018-2020. He is also a licensed Amateur Radio Operator, a Private Pilot, and a registered Professional Engineer in Texas.