Multimodal Beam Prediction With Missing Modality Adaptability: A Contrastive Learning Approach

Lebin Yao, *Graduate Student Member, IEEE*, Ke Ma, Yiliang Sang, *Graduate Student Member, IEEE*, Pengyu Wang, Mingkun Li, *Graduate Student Member, IEEE*, Zihan Liu, Zhaocheng Wang, *Fellow, IEEE*, Zhu Han, *Fellow, IEEE*, and Sheng Chen, *Life Fellow, IEEE*

***Abstract*—Multimodal beam prediction based on deep learning can significantly reduce the beam training overhead in millimeter-wave multiple-input multiple-output (MIMO) systems. However, the performance of existing multimodal approaches is often degraded by missing modalities, and their intricate modality fusion mechanisms typically incur high inference complexity. To address these challenges, a novel contrastive learning approach that inherently accommodates arbitrary patterns of missing modalities and enables low-complexity inference is proposed in this paper. First, a unified latent space is constructed through contrastive learning, which aligns the unimodal embeddings with a set of learnable candidate beam embeddings. This alignment facilitates simple yet effective modality fusion by averaging the embeddings from available modalities, as well as low-complexity beam selection via similarity search. Then, to establish a semantically meaningful initial structure within the latent space, we leverage the inter-beam correlations to initialize the candidate beam embeddings, thereby guiding the pretraining of unimodal encoders. Finally, the standard information noise contrastive estimation loss is generalized by incorporating the relevance among training samples and candidate beams to provide richer supervisory signals for model training. Simulation results demonstrate that the proposed approach can establish a semantically meaningful latent space, and achieve superior performance and lower complexity compared to existing approaches in multimodal beam prediction with missing modalities.**

***Index Terms*—Millimeter-wave communications, beam prediction, multimodal learning, missing modality, contrastive learning.**

Received 5 January 2026; revised 28 May 2026 and 10 August 2026; accepted 14 September 2026. Date of publication 22 September 2026; date of current version 7 October 2026. This work was supported by the National Natural Science Foundation of China under Grant U22B2057. The associate editor coordinating the review of this article and approving it for publication was J. Park. *(Corresponding authors: Zhaocheng Wang; Ke Ma.)*

Lebin Yao, Ke Ma, Yiliang Sang, Pengyu Wang, Zihan Liu, and Zhaocheng Wang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: yaolb24@mails.tsinghua.edu.cn; make15@tsinghua.org.cn; sangyl23@mails.tsinghua.edu.cn; wangpengyu@mail.tsinghua.edu.cn; zh-l23@mails.tsinghua.edu.cn; zcwang@tsinghua.edu.cn).

Mingkun Li is with the Institute of Data and Information, Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China (e-mail: lmk23@mails.tsinghua.edu.cn).

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea (e-mail: hanzhu22@gmail.com).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K., and also with the Faculty of Information Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: sqc@ecs.soton.ac.uk).

Digital Object Identifier 10.1109/TCOMM.2026.3735831

I. Introduction

GIVEN the vast bandwidth resources in the millimeter-wave (mmWave) frequency band, mmWave communication is widely regarded as a key technology for the sixth-generation (6G) wireless systems [1], [2]. However, mmWave signals are subject to severe propagation pathloss, which poses a significant challenge to establishing reliable communication links [3]. To overcome this attenuation, large-scale antenna arrays are deployed at both the base station (BS) and user equipment (UE) to form high-gain directional beams, thereby compensating for the signal degradation [4], [5], [6].

To achieve high mmWave received power, the beam training technology is widely employed to identify the transmitting and receiving beam pair with the highest beamforming gain [7], [8], [9], [10], [11]. Specifically, the most fundamental approach is an exhaustive search that evaluates every candidate beam pair from a predefined codebook to identify the optimal one [7]. Although this approach guarantees the optimality, it generally incurs prohibitive overhead and complexity. To mitigate this issue, several heuristic approaches have been developed. For instance, the hierarchical search approach identifies a promising wide beam first, and then refines the search to a narrow beam within its coverage for reducing the search space [8], [9]. Another common approach is the interactive search, which sweeps the transmitting and receiving beams independently rather than jointly [10], [11]. Nevertheless, these heuristic approaches still impose a substantial beam training burden, and this motivates the exploration of more efficient approaches.

Benefiting from the strong nonlinear fitting capability of deep learning, many works have proposed to predict the optimal beam with deep learning for reducing the beam training overhead [12], [13], [14], [15], [16], [17], [18]. By fitting the complex mapping from various auxiliary measurements to the optimal beam, this deep learning based approach enables direct beam prediction without cumbersome search. One line of researches utilizes the wireless signal measurements. For instance, the study [12] used the received power from a subset of beams to infer the optimal one, while the work [13] predicted the high-resolution narrow beams based on low-resolution wide beam measurements. Furthermore, the sub-6 GHz channel information was utilized in [14] for mmWave beam prediction. Another research stream focuses

0090-6778 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

on leveraging sensory data, such as positional data [15], light detection and ranging (LiDAR) point clouds [16], radar signals [17], and camera images [18]. These sensing modalities include the complementary information for beam prediction. The positional data directly conveys the UE-to-BS geometric relationship, while the LiDAR point clouds and the radar measurements describe the range and angular distribution of surrounding objects through different physical principles [16], [17]. However, all these approaches [12], [13], [14], [15], [16], [17], [18] rely on a single input modality and fail to exploit the complementary information from different modalities, which limits their performance and robustness.

To overcome this limitation, fusing multiple modalities for beam prediction has received much attention [19], [20], [21], [22], [23], [24], [25]. Early works in this area typically employed a straightforward fusion strategy, which concatenates the features from different modalities to form a fused feature for subsequent processing. This concatenation based approach has been harnessed to fuse positional data with camera images, LiDAR measurements, or both of them [19], [20], [21], [22]. The study [23] has further incorporated a wider range of sensory inputs, including position, LiDAR, radar, and camera. However, such simple concatenation often fails to capture the sophisticated, high-order interactions between modalities. To more effectively model these complex inter-modal relationships, recent works have focused on designing advanced fusion mechanisms. For instance, the study [24] introduced a cross-modal feature enhancement (CMFE) mechanism that performs fusion through cross attention among unimodal features to adaptively interchange information between different modalities, while the work [25] adopted multimodal compact bilinear pooling (MCBP), which fuses modalities by approximating the outer product of unimodal features to compactly capture bilinear interactions between modality features. In addition, contrastive learning has recently been incorporated into multimodal beam prediction to align and enhance features from different modalities prior to modality fusion [26], [27]. By leveraging the complementary information from multiple sources, these multimodal approaches [19], [20], [21], [22], [23], [24], [25], [26], [27] achieve significant performance gains compared to their unimodal counterparts.

Despite the enhanced performance, existing multimodal approaches suffer from two serious shortcomings that hinder their practical deployment. On the one hand, most of the existing multimodal approaches are designed under the assumption that all modalities are constantly accessible. This assumption, however, usually fails in real-world deployments where data streams may be interrupted by various factors such as sensor malfunctions, hardware limitations, or privacy constraints [28], [29]. In the scenarios with missing modalities, the unavailable data are generally imputed with zero or random values, which can destabilize the input distribution of deep learning models and degrade prediction performance [30]. To alleviate this problem, the channel attention mechanism was adopted in [31] to adaptively fuse the available modalities, but it still relies on zero-value imputation and suffers from the performance loss caused by the unstable input. On the other hand, the existing multimodal beam prediction approaches typically include intricate modality fusion and post-processing mechanisms, which can lead to high inference complexity.

To address the dual challenges of missing modalities and high inference complexity, a novel multimodal beam prediction framework based on contrastive learning is developed in this paper. This framework projects all modalities onto the embeddings of the same dimension, and constructs a unified latent space with semantically meaningful structure to facilitate efficient modality fusion and beam selection. Concretely, motivated by the success of contrastive learning in multi-view representation alignment [26], [27], [32], [33], [34], the modality embeddings from unimodal encoders are aligned with a set of learnable candidate beam embeddings via contrastive training. Through this alignment, the modality embeddings corresponding to the same optimal beam are clustered together around the corresponding candidate beam embedding, forming semantically meaningful groups within the latent space. The resulting unified latent space offers two *critical advantages*. Firstly, the modality fusion is simplified to averaging the available modality embeddings, which inherently accommodates arbitrary patterns of missing modalities without additional complexity. Secondly, the prediction task is reformulated as low-complexity beam selection via similarity search, where the candidate beam whose embedding is most similar to the fused modality embedding is selected as the prediction result. Through these designs, the proposed framework is fundamentally distinguished from existing multimodal approaches in two aspects. On the one hand, unlike existing contrastive learning based multimodal approaches that use contrastive learning only as a feature enhancement method before modality fusion [26], [27], our framework leverages contrastive learning to construct a unified latent space with semantically meaningful structure that inherently accommodates potential missing modalities during both training and inference. On the other hand, instead of relying on intricate fusion networks such as cross attention [24] or bilinear pooling [25], our framework performs inference through simple embedding averaging and similarity search, achieving significantly reduced complexity.

To establish a more semantically meaningful structure within the latent space, two methodological refinements to the standard contrastive learning framework are incorporated to exploit the rich structural information within the beamforming gain distributions. On the one hand, for each candidate beam, its beamforming gains across the training set are collected into a vector and compressed into the latent space, which serves as the initialization of the corresponding learnable embedding. Two implementations of this compression are presented to allow flexible balance between computational complexity and representation quality, namely a training-free random projection technique [35] with ultra-low complexity, and a learned low-rank factorization method that further adapts to the inter-beam statistical structure at the cost of additional offline training. These initialized embeddings are then used to guide the pretraining of unimodal encoders. This strategy ensures that the embeddings for candidate beams with similar spatial coverage properties are co-located in the latent space from the outset, forming a semantically meaningful initial struc-

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

ture before joint contrastive training. On the other hand, the standard information noise contrastive estimation (InfoNCE) loss [36] is generalized to provide richer supervisory signals for model training. This generalization enhances the standard InfoNCE loss by introducing additional weighted loss terms for semantically high-relevance samples and candidate beams, which provide moderated attraction instead of repulsion in contrastive learning. These additional terms guide the model to learn more semantically meaningful embeddings that capture the underlying correlations among different samples and candidate beams, thereby enhancing its performance.

To validate the effectiveness of our proposed approach, numerical simulations are conducted on the real-world dataset DeepSense6G [37]. Simulation results demonstrate that our contrastive learning approach successfully establishes a semantically meaningful structure within the latent space, and significantly outperforms its traditional counterparts in the beam prediction task, especially for the scenarios with missing modalities. Furthermore, ablation studies confirm that both of our methodological refinements contribute to improving the beam prediction performance.

The main contributions of this paper are summarized as follows:

- A novel contrastive learning based multimodal beam prediction framework is developed. By constructing a unified latent space, this framework facilitates the modality fusion based on embedding averaging and the beam selection via similarity search, which inherently accommodates arbitrary missing modality patterns without additional complexity.
- An initialization strategy for the candidate beam embeddings, coupled with the pretraining of unimodal encoders, is proposed. To establish a semantically meaningful initial structure within the latent space, the dataset-wide beamforming gain patterns are compressed into the latent space as the initialization of candidate beam embeddings, and the unimodal encoders are then pretrained based on these initialized embeddings.
- The standard InfoNCE loss for contrastive learning is generalized to provide richer supervisory signals. The proposed loss incorporates additional terms to soften the repulsion for semantically high-relevance samples and candidate beams, guiding the model to learn more semantically meaningful embeddings that capture their underlying correlations.

This paper is organized as follows. Section II introduces the system model and formulates the multimodal beam prediction problem. Section III elaborates on the proposed contrastive learning framework and the two methodological refinements. Section IV presents the simulation results, and finally Section V concludes the paper.

Notations: $\mathbb{R}^{M\times N}$ and $\mathbb{C}^{M\times N}$ denote the $M\times N$ real and complex spaces, respectively. The vectors and matrices are denoted by boldface lower case and upper case letters, e.g., $\boldsymbol{a}$ and $\boldsymbol{A}$. $[\cdot]_i$ is the i-th element of a vector, while $[\cdot]_{i,j}$ is the (i,j)-th element of a matrix. $[\cdot]_{i,:}$ and $[\cdot]_{:,j}$ represent the i-th row and the j-th column of a matrix, respectively. $(\cdot)^{\mathrm{T}}$ is the transpose operator. $|\cdot|$ represents the absolute value for a scalar or the number of elements for a set. $\|\cdot\|_2$ is the 2-norm operator. The sets are denoted by calligraphic letters, e.g., $\mathcal{A}$.

II. System Model and Problem Formulation

A. System Model

Consider the uplink transmission in a mmWave vehicle to infrastructure (V2I) wireless system, where the BS and the UE are equipped with an N_{BS}-element antenna array and a single antenna, respectively [24], [37]. It is noted that the proposed approach can be directly applied to the systems with multi-antenna UE. With $\boldsymbol{h}\in\mathbb{C}^{N_{\mathrm{BS}}}$ denoting the channel vector, the received signal y is given by

$$y=\sqrt{P}\boldsymbol{f}^{\mathrm{T}}\boldsymbol{h}x+n, \tag{1}$$

where P is the transmitting power, and $\boldsymbol{f}\in\mathbb{C}^{N_{\mathrm{BS}}}$ is the adopted receiving beam vector from a predefined K-beam codebook $\mathcal{F}=\{\boldsymbol{f}_k\}_{k=1}^{K}$, while x represents the transmitted data symbol satisfying the power constraint $|x|=1$, and $n\sim\mathcal{CN}(0,\sigma^2)$ is the complex-valued additive white Gaussian noise (AWGN) with power σ^2.

B. Problem Formulation

Based on the above system model, the receiving signal-to-noise ratio (SNR) γ can be expressed as

$$\gamma=\frac{P|\boldsymbol{f}^{\mathrm{T}}\boldsymbol{h}|^2}{\sigma^2}. \tag{2}$$

To maximize the receiving SNR, beam training aims to find the optimal candidate beam index k^* satisfying

$$k^*=\arg\max_{k\in\{1,2,\cdots,K\}}|\boldsymbol{f}_k^{\mathrm{T}}\boldsymbol{h}|. \tag{3}$$

To reduce the beam training overhead, the multimodal beam prediction technology utilizes deep learning to predict the optimal receiving beam based on multiple auxiliary measurements. Concretely, the received power of equally spaced beams with spacing K_0 is measured to provide direct information about the wireless channel [12]. In addition to these wireless signal measurements, multiple sensory data are utilized to provide rich complementary information about the environment, including the positional data from the global positioning system (GPS), the LiDAR point clouds obtained by LiDAR sensors, and the radar measurements from a frequency modulated continuous wave (FMCW) radar [37]. Let $\mathcal{M}=\{\mathrm{beam},\mathrm{position},\mathrm{LiDAR},\mathrm{radar}\}$ denote the set of modalities, and $\boldsymbol{s}_i\in\mathbb{R}^{N_i}, i\in\mathcal{M}$, denote the measurements of modality i with dimension N_i. Since the modality data are not always available in practical systems, we further introduce binary masks $m_i\in\{0,1\}, i\in\mathcal{M}$, to indicate the availability of each modality, where $m_i=1$ if the modality i is available, otherwise $m_i=0$. For the missing modalities, their measurements $\boldsymbol{s}_i$ are filled with zero values. The deep learning based multimodal beam prediction problem can be then formulated as

$$\hat{k}^*=g(\{\boldsymbol{s}_i\}_{i\in\mathcal{M}},\{m_i\}_{i\in\mathcal{M}};\Theta), \tag{4}$$

where $g(\cdot)$ is the prediction model with learnable parameters Θ, and $\hat{k}^*$ is the predicted optimal beam index.

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

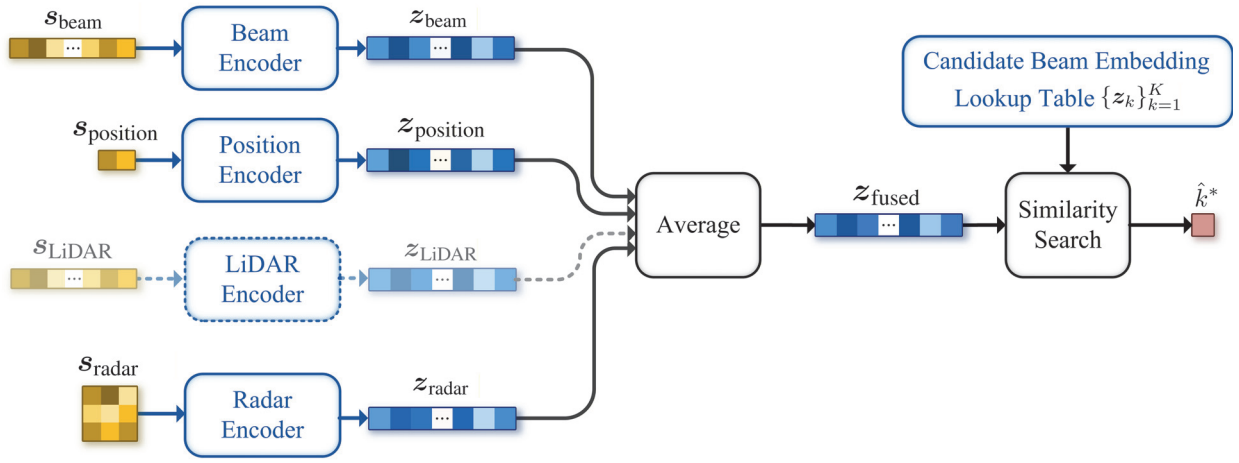


Fig. 1. Proposed contrastive learning based multimodal beam prediction framework, where missing modalities are shown in faded style.

The metric for beam prediction performance is the normalized beamforming gain η, which is defined as

$$\eta = \frac{|\boldsymbol{f}_{\hat{k}^*}^{\mathrm{T}} \boldsymbol{h}|^2}{|\boldsymbol{f}_{k^*}^{\mathrm{T}} \boldsymbol{h}|^2}. \tag{5}$$

III. Contrastive Learning Based Multimodal Beam Prediction

A. Data Preprocessing

To sufficiently exploit the environment information reflected by modality measurements, the positional data and radar measurements are preprocessed before being fed into deep learning models to extract meaningful features, as introduced below [17], [21].

1) Position: The positional data offers information about the geometric relationship between the UE and the BS, which is crucial for determining the approximate range of the optimal beam. Each original positional data sample is a 2-element vector representing the latitude and longitude of the mobile UE. Besides, the latitude and longitude of the stationary BS are known in advance. Based on these two coordinates, the relative position of the UE with respect to the BS is calculated in a local Cartesian coordinate system, yielding east-west and north-south distances in meters as the input features of the positional data. Notably, the positional data acquired from GPS may suffer from substantial errors in urban environments, which makes additional sensing modalities crucial in practical deployments [26].

2) Radar: The radar measurements offer rich information about the environment, such as the distance and direction of surrounding objects. At each sampling instant, the FMCW radar captures the reflected chirp signals with its multiple receiving antennas. To extract the range and angle information from the original radar measurements, the 2D fast Fourier transform (FFT) is applied to the time sample axis and the antenna axis to obtain the range-angle spectrum. This spectrum is then normalized as the input features of the radar data.

B. Contrastive Learning Framework

To efficiently fuse the multimodal information for accurate beam prediction under the scenario with missing modalities, we propose a novel multimodal beam prediction framework based on contrastive learning, which is shown in Fig. 1. Specifically, this framework employs a set of unimodal encoders to extract discriminative features from different modality measurements and project them onto a unified latent space. To construct a semantically meaningful structure within this latent space, we introduce a set of K learnable candidate beam embeddings, each representing one candidate beam in the codebook $\mathcal{F}$. The unimodal encoders and candidate beam embeddings are then jointly trained via a contrastive learning objective, which aligns each sample's modality embeddings with its corresponding ground-truth beam embedding, while pushing them away from the other candidate beam embeddings. Through this contrastive training, the modality embeddings are clustered around their corresponding candidate beam embeddings, and the modality fusion is simplified to averaging the aligned embeddings from available modalities, which inherently accommodates arbitrary missing modality patterns without additional complexity. Moreover, the beam selection is reduced to a low-complexity similarity search that selects the candidate beam whose embedding is the closest to the fused embedding as the prediction result. The details of the proposed framework are presented as follows.

1) Design of Unimodal Encoders: To effectively extract the modality-specific information, a specifically-designed unimodal encoder $\phi_i(\cdot;\boldsymbol{\theta}_i) : \mathbb{R}^{N_i} \to \mathbb{R}^d$ parameterized by $\boldsymbol{\theta}_i$ is employed to map the input measurements $\boldsymbol{s}_i$ onto a d-dimensional embedding $\boldsymbol{z}_i \in \mathbb{R}^d$ for each modality $i \in \mathcal{M}$, which can be expressed as

$$\boldsymbol{z}_i = \phi_i(\boldsymbol{s}_i;\boldsymbol{\theta}_i). \tag{6}$$

To adapt to the characteristics of various modalities, different network architectures are adopted for different unimodal encoders. Specifically, the fully connected network (FCN) is adopted for the partial beam and position modalities due to

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

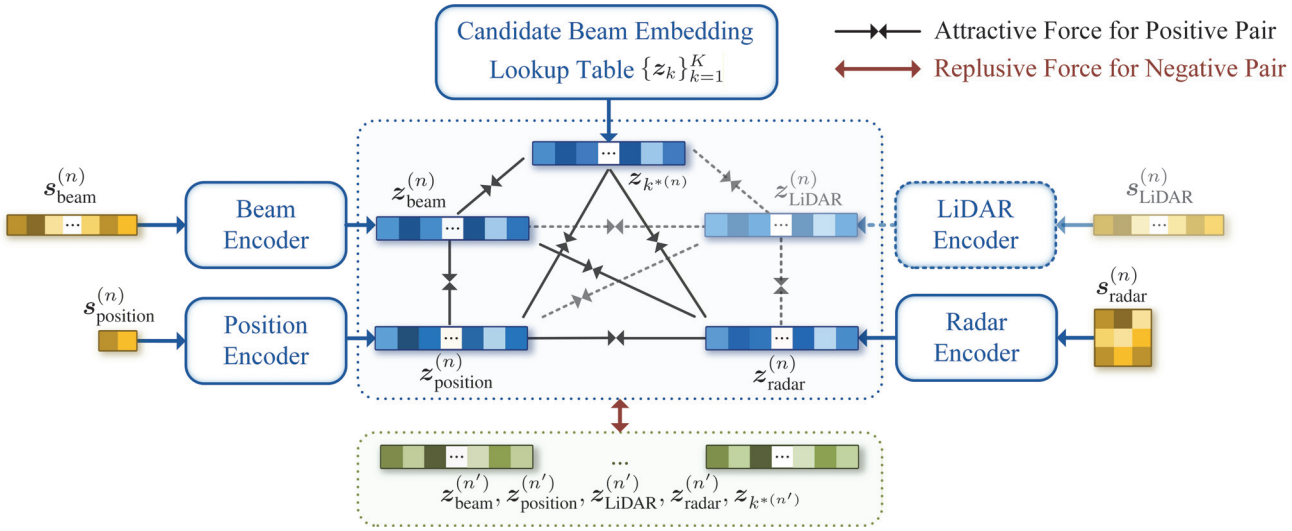


Fig. 2. Illustration of effect of contrastive training, where the missing modalities are shown in faded style.

their low-dimensional nature, while the 1D and 2D convolutional neural networks (CNNs) are respectively employed for the LiDAR and radar modalities to effectively capture their spatial structures with relatively low computational complexity. By setting the dimension of embedding d to be identical across all modalities, the embeddings from different modalities become comparable, which is crucial for subsequent embedding alignment and modality fusion.

2) Design of Candidate Beam Embeddings: To construct a semantically meaningful structure within the latent space, a set of learnable candidate beam embeddings $\{\boldsymbol{z}_k\}_{k=1}^K$ is introduced in parallel with the unimodal encoders as the anchors in this space, where $\boldsymbol{z}_k \in \mathbb{R}^d$ represents the embedding vector for the k-th candidate beam in the codebook $\mathcal{F}$. By aligning the modality embeddings with their corresponding ground-truth candidate beam embeddings, the modality embeddings corresponding to the same optimal beam are clustered together around the anchor, while the embeddings corresponding to different beams are pushed away from each other, thus forming a semantically meaningful structure in the latent space.

3) Contrastive Learning Based Embedding Alignment: The embedding alignment is achieved by jointly training the unimodal encoders and the candidate beam embeddings in a contrastive learning manner. Specifically, for the n-th training sample, we denote its modality embeddings as $\{\boldsymbol{z}_i^{(n)}\}_{i\in\mathcal{M}}$, ground-truth beam index as $k^{*(n)}$, and modality availability masks as $\{m_i^{(n)}\}_{i\in\mathcal{M}}$. To incorporate the learnable candidate beam embeddings that serve as anchors in the latent space, we introduce a virtual modality indexed by anchor with availability mask $m_{\text{anchor}}^{(n)} = 1$ for all training samples. For the n-th training sample, its virtual modality embedding is obtained by looking up the learnable candidate beam embedding table

$$\boldsymbol{z}_{\text{anchor}}^{(n)} = \boldsymbol{z}_{k^{*(n)}}. \tag{7}$$

With this formulation, the extended modality set is defined as $\bar{\mathcal{M}} = \mathcal{M} \cup \{\text{anchor}\}$. To uniquely identify each embedding in the training set, we use the index (n, i) to specify the embedding $\boldsymbol{z}_i^{(n)}$ from modality $i \in \bar{\mathcal{M}}$ of sample n.

A principled and widely adopted objective for contrastive learning is the InfoNCE loss [36], which encourages the embeddings from the same sample to be similar while pushing the embeddings from different samples apart. For a pair of embeddings $(\boldsymbol{z}_i^{(n)}, \boldsymbol{z})$, the pairwise contrastive loss is defined as

$$\ell(\boldsymbol{z}_i^{(n)}, \boldsymbol{z}) = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i^{(n)}, \boldsymbol{z})/\tau)}{\sum_{(n',j)\neq(n,i)} m_j^{(n')} \exp(\text{sim}(\boldsymbol{z}_i^{(n)}, \boldsymbol{z}_j^{(n')})/\tau)}, \tag{8}$$

where $\tau > 0$ is the temperature hyperparameter that modulates the penalty, and $\text{sim}(\cdot,\cdot)$ is the cosine similarity function defined as

$$\text{sim}(\boldsymbol{z}, \boldsymbol{z}') = \frac{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}'}{\|\boldsymbol{z}\|_2 \|\boldsymbol{z}'\|_2}. \tag{9}$$

Then, the InfoNCE loss is calculated as

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{\sum_{n=1}^N \sum_{i\in\bar{\mathcal{M}}} m_i^{(n)}} \sum_{n=1}^N \sum_{i\in\bar{\mathcal{M}}} \frac{m_i^{(n)}}{|\mathcal{P}(n,i)|} \times \sum_{(n,j)\in\mathcal{P}(n,i)} \ell(\boldsymbol{z}_i^{(n)}, \boldsymbol{z}_j^{(n)}), \tag{10}$$

where $\mathcal{P}(n,i)$ is the positive pair index set for embedding $\boldsymbol{z}_i^{(n)}$ defined as

$$\mathcal{P}(n,i) = \{(n,j) : j \in \bar{\mathcal{M}}, j \neq i, m_j^{(n)} = 1\}, \tag{11}$$

which includes the indices of other available embeddings from the same training sample.

Fig. 2 illustrates the effect of contrastive training. The numerator of contrastive loss $\exp(\text{sim}(\boldsymbol{z}_i^{(n)}, \boldsymbol{z})/\tau)$ establishes attractive forces between positive pairs, pulling the modality embeddings and the ground-truth candidate beam embedding from the same training sample together to form semantically meaningful clusters in the latent space, where each cluster corresponds to a specific optimal candidate beam. Meanwhile, repulsive forces embodied in the denominator of contrastive

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

loss $\sum_{(n',j)\neq(n,i)} m_j^{(n')} \exp(\mathrm{sim}(\boldsymbol{z}_i^{(n)}, \boldsymbol{z}_j^{(n')})/\tau)$ push negative pairs (the embeddings from different training samples) apart, ensuring that the latent space maintains a discriminative structure across different samples. Notably, by including the candidate beam embeddings $\{\boldsymbol{z}_k\}_{k=1}^K$ in the contrastive training, the modality embeddings are not only aligned with each other, but also anchored to their corresponding candidate beam embeddings, which directly facilitates the subsequent similarity based beam selection. Moreover, since missing modalities are excluded from the loss calculation by the availability masks $m_i^{(n)}$, the framework naturally accommodates arbitrary missing modality patterns without requiring specialized handling mechanisms.

4) Modality Fusion and Beam Selection: With the unified latent space constructed through the contrastive training, both the modality fusion and the beam prediction can be accomplished in a remarkably simple yet effective manner, as illustrated in Fig. 1. For a new test sample with modality measurements $\{\boldsymbol{s}_i\}_{i\in\mathcal{M}}$ and availability masks $\{m_i\}_{i\in\mathcal{M}}$, we first extract the embeddings from all available modalities using the trained unimodal encoders, and then fuse them by averaging to obtain the fused embedding $\boldsymbol{z}_{\text{fused}}$, which can be expressed as

$$\boldsymbol{z}_{\text{fused}} = \frac{1}{M}\sum_{i\in\mathcal{M}} m_i \boldsymbol{z}_i = \frac{1}{M}\sum_{i\in\mathcal{M}} m_i \phi_i(\boldsymbol{s}_i;\boldsymbol{\theta}_i), \tag{12}$$

where $M = \sum_{i\in\mathcal{M}} m_i$ is the number of available modalities. This averaging based fusion is enabled by the fact that the modality embeddings from the same sample have all been aligned to its corresponding candidate beam embedding during training, ensuring that they reside in neighboring regions of the latent space and can be meaningfully combined through simple arithmetic averaging. It should be noted that the fusion in (12) treats all the available modalities equally regardless of their reliability, and an inaccurate embedding from a degraded modality may thus bias the fused embedding and impair the beam prediction performance. Since the qualities of different modalities are comparable and each of them remains relatively stable in the considered V2I scenarios, the equal-weight averaging is adopted for its simplicity. Nevertheless, the equal-weight averaging in (12) can be readily extended to a confidence-aware weighted averaging that assigns larger weights to more reliable modality embeddings, which is expected to be beneficial when the modality qualities are highly heterogeneous or vary rapidly over time. Subsequently, the beam selection is conducted by searching for the candidate beam whose embedding has the highest cosine similarity to the fused embedding, which is formulated as

$$\hat{k}^* = \arg\max_{k\in\{1,2,\cdots,K\}} \mathrm{sim}(\boldsymbol{z}_{\text{fused}}, \boldsymbol{z}_k). \tag{13}$$

To efficiently implement this similarity search, all well-trained candidate beam embeddings are first normalized after contrastive training. During each inference, the fused embedding of testing sample is normalized, after which the K cosine similarity calculations can be carried out in parallel through a single matrix-vector multiplication between the matrix of normalized candidate beam embeddings and the normalized fused embedding. Benefiting from both the similarity search based design and this efficient implementation, the proposed beam selection mechanism achieves significantly lower computational complexity compared to the intricate post-processing networks typically employed in traditional approaches.

C. Candidate Beam Embedding Initialization and Unimodal Encoder Pretraining

While the contrastive learning framework described in Subsection III-B provides a principled approach to construct the unified latent space, random initialization of candidate beam embeddings typically results in a latent space without semantically meaningful structure, and it is challenging to establish such a structure through training from this arbitrary starting point. To address this issue, we propose a synergistic strategy that combines the initialization of candidate beam embeddings with the pretraining of unimodal encoders to establish a semantically meaningful initial structure within the latent space before joint contrastive training by exploiting inter-beam correlations. Specifically, we first initialize the candidate beam embeddings by the compression of their dataset-wide beamforming gain patterns, which ensures that the beams with similar spatial coverage properties are co-located in the latent space from the outset. Subsequently, these initialized embeddings are used as the fixed anchors to guide the pretraining of unimodal encoders, enabling the encoders to learn the embeddings that respect this semantically meaningful structure before joint contrastive training. This combined initialization and pretraining strategy provides a strong structural foundation for the latent space, facilitating more effective embedding alignment and beam selection in the subsequent joint contrastive training phase.

For the initialization of candidate beam embeddings, we first analyze the beamforming gain patterns across all training samples to capture the inter-beam correlations from the training dataset. For each training sample $n \in \{1, 2, \cdots, N\}$, the normalized beamforming gain of the k-th candidate beam is defined as

$$g_{n,k} = \frac{|\boldsymbol{f}_k^{\mathrm{T}} \boldsymbol{h}^{(n)}|^2}{|\boldsymbol{f}_{k^{*(n)}}^{\mathrm{T}} \boldsymbol{h}^{(n)}|^2}, \tag{14}$$

where $\boldsymbol{h}^{(n)}$ denotes the channel vector of the n-th training sample. This normalization eliminates the effect of varying received power levels across different samples, allowing us to focus on the relative beamforming gain distribution of each beam. By collecting the normalized gains for all beams and all samples, we construct a normalized beamforming gain matrix $\boldsymbol{G} \in \mathbb{R}^{N\times K}$, where $[\boldsymbol{G}]_{n,k} = g_{n,k}$. From this matrix, we extract the beam-wise beamforming gain pattern vector for each beam k as $\boldsymbol{g}_k = [\boldsymbol{G}]_{:,k} \in \mathbb{R}^N$, which captures how beam k performs across all training samples. For the beams with similar spatial coverage properties, their beamforming gains will exhibit similar patterns across the training samples, resulting in similar normalized gain pattern vectors $\boldsymbol{g}_k$ in the N-dimensional space. Notably, to be consistent with the similarity metric used in the InfoNCE loss and beam selection, all vector similarities in this subsection are measured by cosine similarity defined in (9).

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

To initialize the candidate beam embeddings in the d-dimensional latent space while preserving the inter-beam correlations captured by these beamforming gain pattern vectors, two methods that offer different tradeoffs between computational complexity and representation quality are presented, namely the random projection and the learned low-rank factorization.

1) Random Projection: The random projection maps the beamforming gain pattern vectors via a random linear projection, which is training-free and incurs negligible computational cost. Specifically, a Gaussian random projection matrix $\boldsymbol{R} \in \mathbb{R}^{d\times N}$ with entries independently sampled from $\mathcal{N}(0, 1/d)$ is generated, and the initial embedding for the k-th candidate beam is computed as

$$\boldsymbol{z}_k^{\text{init}} = \frac{\boldsymbol{R}\boldsymbol{g}_k}{\|\boldsymbol{R}\boldsymbol{g}_k\|_2}. \tag{15}$$

According to the Johnson-Lindenstrauss lemma [38], the random projection approximately preserves the pair-wise similarities between different beamforming gain pattern vectors with high probability, and this ensures that the beams with similar spatial coverage properties remain close to each other in the reduced d-dimensional latent space.

2) Learned Low-Rank Factorization: While the random projection enjoys ultra-low complexity, its data-agnostic nature does not adapt to the actual statistical structure of the beamforming gain matrix $\boldsymbol{G}$, leaving room for more accurate compression. To pursue a tighter preservation of inter-beam correlations, we further introduce the learned low-rank factorization, which employs an autoencoder to factorize $\boldsymbol{G}$ by minimizing the reconstruction error over all training samples, and derives the embeddings from the learned factor matrix. To facilitate this sample-wise reconstruction, we extract the sample-wise beamforming gain pattern from $\boldsymbol{G}$ as

$$\boldsymbol{g}^{(n)} = ([\boldsymbol{G}]_{n,:})^{\mathrm{T}} = [g_{n,1}, g_{n,2}, \cdots, g_{n,K}]^{\mathrm{T}} \in \mathbb{R}^K, \tag{16}$$

which collects the normalized beamforming gains of all candidate beams for the n-th sample. Based on this sample-wise view, an autoencoder is built with an encoder $\psi_{\text{enc}}(\cdot;\boldsymbol{\theta}_{\text{enc}}) : \mathbb{R}^K \to \mathbb{R}^d$ that maps each sample-wise beamforming gain pattern vector $\boldsymbol{g}^{(n)}$ to a d-dimensional latent code, and a linear decoder parameterized by a weight matrix $\boldsymbol{W} \in \mathbb{R}^{K\times d}$ without bias that reconstructs the gain vector by $\hat{\boldsymbol{g}}^{(n)} = \boldsymbol{W}\psi_{\text{enc}}(\boldsymbol{g}^{(n)};\boldsymbol{\theta}_{\text{enc}})$. Here, the encoder is implemented as an FCN with a bottleneck that forces the network to capture the compact and informative latent codes. The autoencoder is then trained by minimizing the mean squared reconstruction loss as

$$\mathcal{L}_{\text{AE}} = \frac{1}{N}\sum_{n=1}^{N}\left\|\boldsymbol{g}^{(n)} - \boldsymbol{W}\psi_{\text{enc}}(\boldsymbol{g}^{(n)};\boldsymbol{\theta}_{\text{enc}})\right\|_2^2. \tag{17}$$

After training, the linear decoder $\boldsymbol{W}$ approximately solves a low-rank matrix factorization $\boldsymbol{G}^{\mathrm{T}} \approx \boldsymbol{W}\boldsymbol{U}$, where $\boldsymbol{U} \in \mathbb{R}^{d\times N}$ collects the latent codes of all training samples produced by the encoder. Under this factorization, the k-th row of $\boldsymbol{W}$ determines how strongly the k-th beam responds to each latent factor, and therefore two beams with similar gain patterns across the training samples naturally exhibit similar rows in $\boldsymbol{W}$. The k-th row of $\boldsymbol{W}$ is then taken as the initial embedding of the k-th candidate beam, i.e.,

$$\boldsymbol{z}_k^{\text{init}} = \frac{([\boldsymbol{W}]_{k,:})^{\mathrm{T}}}{\|[\boldsymbol{W}]_{k,:}\|_2}. \tag{18}$$

Compared with its random projection counterpart, the learned low-rank factorization adapts to the actual statistical structure of $\boldsymbol{G}$ via end-to-end training, generally yielding lower compression error and better beam prediction performance, at the cost of an additional offline training step before pretraining the unimodal encoders.

Consequently, both initialization methods endow the candidate beam embeddings with a semantically meaningful structure from the outset, where the geometric relationships in the latent space reflect the actual spatial coverage similarities among the beams. In practical deployments, these two methods can be flexibly chosen according to the complexity-performance tradeoff.

With the candidate beam embeddings initialized to reflect the inter-beam correlations, we proceed to pretrain the unimodal encoders to ensure that they learn to respect this semantically meaningful structure. During the pretraining phase, the initialized candidate beam embeddings $\{\boldsymbol{z}_k^{\text{init}}\}_{k=1}^K$ are kept fixed and serve as anchors in the latent space. For each modality $i \in \mathcal{M}$, the encoder ϕ_i is independently trained by aligning its output embeddings with the corresponding fixed candidate beam embeddings based on the InfoNCE loss. Specifically, only the training samples with available corresponding modality are utilized. For the n-th training sample with availability mask $m_i^{(n)} = 1$ and ground-truth beam index $k^{*(n)}$, there are only two embeddings to be considered, i.e., the modality embedding $\boldsymbol{z}_i^{(n)}$ and the corresponding fixed candidate beam embedding $\boldsymbol{z}_{\text{anchor}}^{(n)} = \boldsymbol{z}_{k^{*(n)}}^{\text{init}}$. These two embeddings form a positive pair, and the loss is calculated following the same principle as in (8) and (10), where the modality embedding is encouraged to align with the fixed beam embedding while being pushed away from the embeddings of other training samples. Notably, although each modality encoder is pretrained independently, they are all aligned to the same set of fixed candidate beam embeddings, and this ensures that the resulting embeddings from different modalities are compatible within the same latent space.

After pretraining, both the pretrained encoders and the initialized candidate beam embeddings are used as the starting point for the subsequent joint contrastive training, to provide a strong structural foundation that facilitates the construction of a semantically meaningful and unified latent space.

D. Generalized InfoNCE Loss

While the standard InfoNCE loss described in Subsection III-B provides an effective mechanism for the joint contrastive training of unimodal encoders and candidate beam embeddings, it treats all negative embedding pairs equally by assigning them uniform importance. However, in the context of beam prediction, the underlying relevance among training samples and candidate beams varies significantly, which is inconsistent with such uniform

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

treatment. In particular, some negative embedding pairs are indeed semantically irrelevant and should be strongly repelled, while others corresponding to highly relevant samples or beams deserve moderated treatment rather than strong repulsion. Such highly relevant cases typically arise from two types of relevance. The first type is inter-sample relevance. Two training samples may have similar beamforming gain distributions across all candidate beams, which indicates that they share similar propagation characteristics. For such highly relevant samples, their modality embeddings encode related environmental information, and should not be strongly repelled from each other. The second type is sample-beam relevance. A candidate beam may achieve high beamforming gain for a sample even though it is not the optimal beam. The embedding of such a sub-optimal beam represents a viable alternative and should not be treated the same as beams with poorer performance. Inspired by these observations, we generalize the standard InfoNCE loss to modulate the treatment of negative embedding pairs based on the relevance defined at the sample and candidate beam level, enabling the model to learn more semantically meaningful embeddings.

To quantify the semantic relevance for negative pairs, we define the two types of relevance at the sample and candidate beam level based on the sample-wise beamforming gain pattern defined in (16). The inter-sample relevance measures the similarity between two training samples, which is defined as the cosine similarity between two sample-wise beamforming gain pattern vectors

$$r_{\mathrm{SS}}(n,n') = \frac{\boldsymbol{g}^{(n)\mathrm{T}}\boldsymbol{g}^{(n')}}{\|\boldsymbol{g}^{(n)}\|_2\|\boldsymbol{g}^{(n')}\|_2}. \tag{19}$$

The sample-beam relevance quantifies how well a candidate beam performs on a given sample, which is defined as

$$r_{\mathrm{SB}}(n,k) = [\boldsymbol{g}^{(n)}]_k = g_{n,k}. \tag{20}$$

Physically, a high $r_{\mathrm{SS}}(n,n')$ indicates that samples n and n' exhibit similar spatial characteristics in their beamforming responses, while a high $r_{\mathrm{SB}}(n,k)$ indicates that beam k provides a highly competitive beamforming gain for sample n.

Based on these relevance values, we identify the high-relevance negative embedding pairs for each embedding in the training process. For a modality embedding $\boldsymbol{z}_i^{(n)}, i \in \mathcal{M}$, from sample n, the high-relevance negative index set $\mathcal{N}^+(n,i)$ consists of two parts. The first part includes the indices of modality embeddings from the samples that are highly relevant to sample n, which is given by

$$\mathcal{N}_{\mathrm{SS}}^+(n,i) = \{(n',j) : r_{\mathrm{SS}}(n,n') > \beta_{\mathrm{SS}}, j \in \mathcal{M}, m_j^{(n')} = 1, n' \neq n\}, \tag{21}$$

where β_{SS} is a threshold parameter for the inter-sample relevance. The second part includes the indices of candidate embeddings whose corresponding beams achieve high gain on sample n, which is given by

$$\mathcal{N}_{\mathrm{SB}}^+(n,i) = \{(n',\mathrm{anchor}) : r_{\mathrm{SB}}(n,k^{*(n')}) > \beta_{\mathrm{SB}}, n' \neq n\}, \tag{22}$$

where β_{SB} is a threshold parameter for the sample-beam relevance. Intuitively, $\mathcal{N}_{\mathrm{SS}}^+(n,i)$ collects modality embeddings from other samples whose beam preferences closely resemble those of sample n, whereas $\mathcal{N}_{\mathrm{SB}}^+(n,i)$ gathers the embeddings of candidate beams that also achieve high gain on sample n. The complete high-relevance negative index set for a modality embedding $\boldsymbol{z}_i^{(n)}$ is the union of these two parts, i.e.,

$$\mathcal{N}^+(n,i) = \mathcal{N}_{\mathrm{SS}}^+(n,i) \cup \mathcal{N}_{\mathrm{SB}}^+(n,i). \tag{23}$$

For a candidate beam embedding $\boldsymbol{z}_{\mathrm{anchor}}^{(n)}$ corresponding to beam $k^{*(n)}$, the high-relevance negative index set includes the indices of modality embeddings from the samples on which beam $k^{*(n)}$ achieves high gain, which is given by

$$\mathcal{N}^+(n,\mathrm{anchor}) = \{(n',j) : r_{\mathrm{SB}}(n',k^{*(n)}) > \beta_{\mathrm{SB}}, j \in \mathcal{M}, m_j^{(n')} = 1, n' \neq n\}. \tag{24}$$

With the high-relevance negative index sets defined, the generalized InfoNCE loss is formulated as

$$\mathcal{L}_{\mathrm{Gen}} = \frac{1}{\sum_{n=1}^N \sum_{i\in\bar{\mathcal{M}}} m_i^{(n)}} \sum_{n=1}^N \sum_{i\in\bar{\mathcal{M}}} \frac{m_i^{(n)}}{|\mathcal{P}(n,i)| + \alpha|\mathcal{N}^+(n,i)|} \times \left(\sum_{(n,j)\in\mathcal{P}(n,i)} \ell(\boldsymbol{z}_i^{(n)}, \boldsymbol{z}_j^{(n)}) + \alpha \sum_{(n',j)\in\mathcal{N}^+(n,i)} \ell(\boldsymbol{z}_i^{(n)}, \boldsymbol{z}_j^{(n')}) \right), \tag{25}$$

where $0 < \alpha < 1$ is a weight parameter that controls the attractive force for high-relevance negative pairs. In this formulation, positive pairs receive unit weight in the loss calculation, while high-relevance negative pairs receive reduced weight α, and other negative pairs contribute only through the denominator of the pairwise contrastive loss $\ell(\cdot,\cdot)$ in (8). Notably, when $\alpha = 0$ or $\beta_{\mathrm{SS}} = \beta_{\mathrm{SB}} = 1$, the contribution from high-relevance negative pairs vanishes, and the generalized loss reduces to the standard InfoNCE loss in (10).

Through this generalized formulation, the contrastive training no longer uniformly repels all negative pairs, but instead provides moderated attraction for those corresponding to highly relevant samples or candidate beams. This design preserves valuable relational information encoded in the beamforming gain patterns, enabling the model to capture the underlying correlations among samples and candidate beams.

IV. Simulation Study

A. Simulation Setup

To accurately evaluate the performance of the proposed contrastive learning approach, we conduct simulations based on the real-world multimodal mmWave communication dataset DeepSense6G [37]. Specifically, scenarios 8 and 9 are adopted, where scenario 9 serves as the primary scenario for the performance evaluation, and scenario 8 is further employed to verify the cross-scenario generalization of the proposed approach. In both scenarios, a BS serves a vehicular UE with a center frequency of 60 GHz. Besides, the BS is equipped with a uniform linear array (ULA) with $N_{\mathrm{BS}} = 16$ antennas, and employs an over-sampled codebook with $K = 64$ candidate beams. For the partial beam modality, the beam sampling

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

TABLE I
NOISE INTENSITY DISTRIBUTIONS

Noise	Intensity distributions
Gaussian noise on beam measurements	SNR $\sim \mathcal{U}(10\ \text{dB}, 15\ \text{dB})$
Gaussian noise on UE positions relative to BS	Power $\sigma^2 \sim \mathcal{U}(0\ \text{m}^2, 1\ \text{m}^2)$
Gaussian noise on LiDAR range measurements	Power $\sigma^2 \sim \mathcal{U}(0\ \text{m}^2, 3\ \text{m}^2)$
Random angular deviations on LiDAR angle measurements	$\Delta\phi \sim \mathcal{U}(-5^\circ, 5^\circ)$
Gaussian noise on radar measurements	SNR $\sim \mathcal{U}(0\ \text{dB}, 10\ \text{dB})$

TABLE II
ARCHITECTURES OF UNIMODAL ENCODERS

Modalities	Layers	Parameters
Partial beam	FC	$d_\text{i} = 8, d_\text{o} = 64$, ReLU
	FC	$d_\text{i} = d_\text{o} = 64$, ReLU
	FC	$d_\text{i} = d_\text{o} = 64$
Position	FC	$d_\text{i} = 2, d_\text{o} = 32$, ReLU
	FC	$d_\text{i} = 32, d_\text{o} = 64$, ReLU
	FC	$d_\text{i} = d_\text{o} = 64$, ReLU
	FC	$d_\text{i} = d_\text{o} = 64$
LiDAR	1D Conv	BN, $d_\text{i} = 1, d_\text{o} = 4, (3, 3, 1)$, ReLU
	1D Conv	BN, $d_\text{i} = 4, d_\text{o} = 8, (3, 3, 1)$, ReLU
	1D Conv	BN, $d_\text{i} = 8, d_\text{o} = 16, (3, 3, 1)$, ReLU
	1D Conv	BN, $d_\text{i} = d_\text{o} = 16, (3, 3, 1)$, ReLU
	AvgPool	$d_\text{i} = d_\text{o} = 16$
	FC	$d_\text{i} = 16, d_\text{o} = 64$, ReLU
	FC	$d_\text{i} = d_\text{o} = 64$
Radar	2D Conv	BN, $d_\text{i} = 1, d_\text{o} = 8, (3, 3, 1)$, ReLU
	2D Conv	BN, $d_\text{i} = d_\text{o} = 8, (3, 3, 1)$, ReLU
	2D Conv	BN, $d_\text{i} = 8, d_\text{o} = 16, (3, 3, 1)$, ReLU
	AvgPool	$d_\text{i} = d_\text{o} = 16$
	FC	$d_\text{i} = 16, d_\text{o} = 64$, ReLU
	FC	$d_\text{i} = d_\text{o} = 64$

interval is set to $K_0 = 8$. In our simulations, the full beam-gain profiles over all candidate beams required to construct the training gain matrix $\boldsymbol{G}$ are taken directly from the beam measurements provided by the DeepSense6G dataset. The acquisition of such profiles constitutes an offline training-data preparation cost and is not required during online inference. In practical deployments, the corresponding training data can be collected offline through exhaustive beam sweeping or obtained from reconstructed channel information.

In practical communication systems, different scenarios may have different modality availability and quality. To validate the robustness of the proposed framework under such realistic conditions, we randomly mask modalities and inject noise with varying levels to the available modalities for simulating diverse modality availability and quality encountered in real-world deployments. Specifically, for all the training and testing samples, each modality is independently masked with a predefined probability, with the masked modality data $\boldsymbol{s}_i^{(n)}$ replaced by zeros and the corresponding availability mask $m_i^{(n)}$ set to zero. Subsequently, we inject Gaussian noise with random intensity to the available modalities to simulate the measurement noise. In addition, for the available LiDAR data, a random angle offset is added aside from the Gaussian noise. The detailed noise intensity distributions are given in Table I.

As introduced in Subsection III-B, we employ specifically-designed unimodal encoders to extract modality-specific features for each modality. The detailed architectures of unimodal encoders are shown in Table II. In Table II, d_i and d_o denote the dimensions of input and output features, respectively. Besides, FC, Conv, BN, and AvgPool represent fully connected layer, convolutional layer, batch normalization, and average pooling, respectively. For convolutional layers, the parameters are specified in the format of (kernel size, stride, padding). Moreover, ReLU denotes the rectified linear unit activation function applied after the corresponding layer. As shown in Table II, all the unimodal encoders output embeddings with a dimension of 64, which is the dimension of the latent space d in our contrastive learning framework.

Both the pretraining and the joint contrastive training are conducted using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 64 for 300 epochs. To prevent overfitting, we apply the early stopping strategy with a patience of 20 epochs. Besides, all the trainings are repeated 5 times and the average performance is reported to stabilize the results.

To comprehensively evaluate the proposed approach, we consider four representative multimodal baselines, which are described as follows.

- *CMFE* [24]: This baseline performs modality fusion through cross attention among the unimodal features. To accommodate missing modalities, the missing modality data are zero-imputed during both training and testing.
- *MCBP* [25]: This baseline fuses modalities by approximating the outer product of unimodal features to compactly capture high-order bilinear interactions. The missing modality data are also zero-imputed during both training and testing.
- *Channel Attention* [31]: This baseline adaptively re-weights the unimodal features through a channel attention module, where the weights corresponding to the missing modalities are set to zero.
- *Decision Fusion* [39]: Inspired by the decision fusion strategy in [39], this baseline trains an independent classifier based on each unimodal encoder to predict a per-modality beam posterior distribution, and averages the posterior distributions of all the available modalities at inference, which inherently accommodates arbitrary missing modality patterns.

For fair comparisons, all the baselines employ the same unimodal encoder architectures as the proposed approach, and are trained under identical settings. The key architectural difference between the baselines and the proposed approach lies in the fusion head and the corresponding output structure, where the baselines employ specifically-designed fusion modules to directly output the probability distribution over the candidate beams, while the proposed approach outputs

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPLEXITY COMPARISON BETWEEN DIFFERENT APPROACHES IN MODALITY FUSION AND BEAM SELECTION PHASES

Approaches	Learnable parameter numbers	Inference FLOPs
Proposed	**4,096**	**8,835**
CMFE [24]	18,656	71,958
MCBP [25]	49,344	162,816
Channel Attention [31]	20,639	41,568
Decision Fusion [39]	16,640	33,280

the latent embedding of each modality and performs beam selection via similarity search. This difference is also the source of the diverse parameter counts reported in Table III.

B. Complexity Analysis

We now analyze and compare the complexity of the proposed framework with that of the four baselines introduced in Subsection IV-A. Since all the approaches employ identical unimodal encoder architectures, we focus the comparison on the modality fusion and beam selection phases. The complexity is evaluated in terms of the number of learnable parameters, which is related to the training complexity, and the number of inference floating-point operations (FLOPs), with the detailed results presented in Table III, where the lowest-complexity approach is highlighted in bold. As shown in Table III, the proposed framework achieves significantly lower complexity than all the baselines in both learnable parameters and inference FLOPs. This substantial complexity reduction stems from our novel contrastive learning based framework, which aligns the embeddings from different modalities to construct a unified latent space. With the knowledge of inter-modality relationships and beam characteristics injected into the learnable parameters of unimodal encoders during contrastive training, we eliminate the need for complex fusion and selection mechanisms, and achieve low-complexity inference.

To further verify that the proposed framework satisfies the latency requirement of practical mmWave V2I systems, we compare its end-to-end inference latency with the channel coherence time of typical mmWave V2I scenarios. Specifically, considering a vehicle moving at 72 $\mathrm{km/h}$ and a carrier frequency of 60 GHz, the channel coherence time can be estimated as $T_{\mathrm{c}} = \frac{c}{v f_{\mathrm{c}}} \approx 0.250$ ms [40], where c is the speed of light, v is the vehicle velocity, and f_{c} is the carrier frequency. We then measure the end-to-end inference latency of the proposed framework on a workstation equipped with an Intel(R) Xeon(R) Gold 6342 CPU @ 2.80 GHz and an NVIDIA GeForce RTX 4090 GPU, and the measured average inference latency is only 0.0996 ms, which is far smaller than the channel coherence time computed above. According to the 480-kHz synchronization signal block (SSB) configuration specified in the Third Generation Partnership Project (3GPP) standard [41], [42], exhaustive measurement of all 64 candidate beams requires approximately 1 ms, whereas measuring the 8 equally spaced beams used by our proposed approach requires approximately 0.125 ms. Combining the partial beam measurement latency with the model inference latency yields a total latency of approximately 0.2246 ms for the proposed pipeline, which remains below the estimated channel coherence time of 0.250 ms and represents an approximately 77.5% latency reduction relative to exhaustive beam sweeping. In addition, the model inference latency of 0.0996 ms is approximately equivalent to the SSB measurement latency of 6.4 beams under this configuration. Since the proposed approach measures only K/K_0 out of the K candidate beams, it achieves a latency reduction when $K - K/K_0 > 6.4$. Moreover, in practical deployments, the inference latency can be further reduced through dedicated edge intelligence accelerators [43]. This confirms that the proposed framework is well-suited for real-time beam prediction in high-mobility V2I scenarios.

TABLE IV
PERFORMANCE COMPARISON UNDER DIFFERENT TEMPERATURES τ

Temperatures τ	Normalized beamforming gains
0.1	0.8642
0.2	0.8689
0.3	**0.8703**
0.4	0.8635
0.5	0.8613

TABLE V
PERFORMANCE COMPARISON UNDER DIFFERENT WEIGHT PARAMETERS α

Weight parameters α	Normalized beamforming gains
0.75	0.8677
0.8	0.8685
0.85	**0.8703**
0.9	0.8685
0.95	0.8684

C. Investigation of Hyperparameters

To investigate the impact of the four hyperparameters for our proposed approach and justify the selection of their values, we evaluate the beam prediction performance under different hyperparameter values on scenario 9, with all modalities independently missing with probability 0.5. Based on extensive experiments, we set the default values of the four hyperparameters as $\tau = 0.3, \alpha = 0.85, \beta_{\mathrm{SB}} = 0.9$ and $\beta_{\mathrm{SS}} = 0.95$, and adopt the control variable method to investigate the impact of each hyperparameter, i.e., when varying one specific hyperparameter, the others are fixed at their aforementioned default values. Tables IV to VII present the normalized beamforming gain performances under different hyperparameter values, with the best hyperparameters and performances highlighted in bold.

As shown in Table IV, the temperature parameter τ in the contrastive loss (8) exhibits an optimal value at $\tau = 0.3$, and the performance degrades when τ is either too small or too large. When τ is too small, the loss concentrates gradients excessively on the hard samples, making its optimization unstable and overly sensitive to outliers or noisy samples. Conversely, when τ is too large, the softmax distribution becomes overly uniform, providing weak learning signals and

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
PERFORMANCE COMPARISON UNDER DIFFERENT THRESHOLD PARAMETERS β_{SB}

Threshold parameters β_{SB}	Normalized beamforming gains
0.8	0.8653
0.85	0.8679
0.9	**0.8703**
0.95	0.8686
1.0	0.8653

TABLE VII
PERFORMANCE COMPARISON UNDER DIFFERENT THRESHOLD PARAMETERS β_{SS}

Threshold parameters β_{SS}	Normalized beamforming gains
0.8	0.8656
0.85	0.8679
0.9	0.8694
0.95	**0.8703**
1.0	0.8635

failing to effectively separate positive pairs from negative pairs.

Table V demonstrates that the weight parameter α in our generalized InfoNCE loss (25) achieves the optimal performance at $\alpha = 0.85$. The performance remains relatively stable across the range $[0.75, 0.95]$, indicating that the framework is robust to the choice of α as long as it provides sufficient attractive force for high-relevance negative pairs. An α value around 0.85 strikes the best balance between incorporating richer supervisory signals from semantically relevant negative pairs and preserving clear boundaries from positive pairs.

Tables VI and VII reveal that the optimal thresholds are $\beta_{\mathrm{SB}} = 0.9$ for sample-beam relevance and $\beta_{\mathrm{SS}} = 0.95$ for inter-sample relevance. For both thresholds, setting the value too low includes many marginally relevant negative pairs that may introduce noise, while setting it too high excludes potentially informative high-relevance pairs, both leading to performance degradation.

Based on these investigations, we adopt $\tau = 0.3$, $\alpha = 0.85$, $\beta_{\mathrm{SB}} = 0.9$ and $\beta_{\mathrm{SS}} = 0.95$ for all subsequent simulations.

D. Simulation Results

1) Performance Comparison With Baselines: Figs. 3 and 4 compare the normalized beamforming gain and the top-5 beam prediction accuracy of the proposed framework with those of the four baselines introduced in Subsection IV-A under varying missing probability conditions on scenario 9, where the top-5 beam prediction accuracy is defined as the proportion of testing samples whose optimal beam falls within the 5 candidate beams with the highest predicted probabilities [15], [16], [17]. As expected, the results demonstrate that the proposed framework consistently outperforms the baselines across all missing probability scenarios, validating the effectiveness of our contrastive learning based design. Especially, as the missing probability increases, the performance gaps between the proposed framework and the baselines widen, which indicates that the proposed framework is more robust to missing modalities.

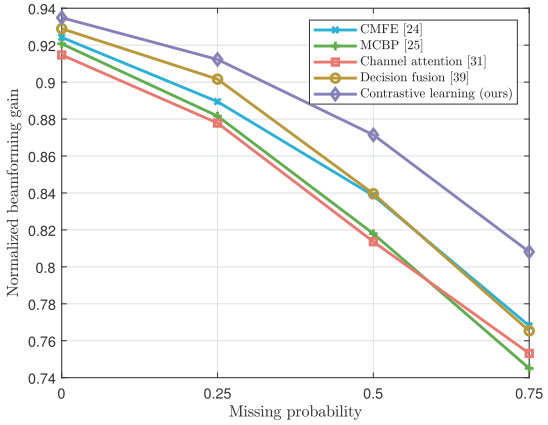


Fig. 3. Normalized beamforming gain comparison between four baselines and our proposed approach on scenario 9.

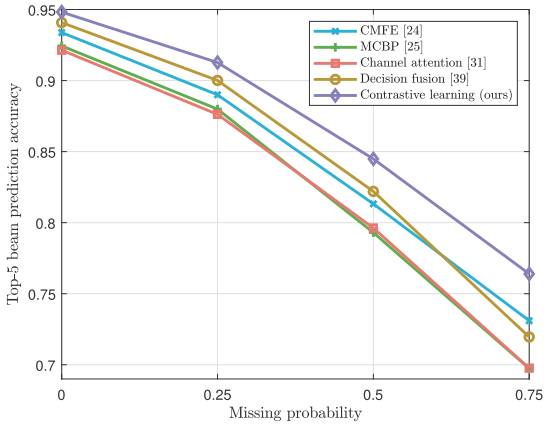


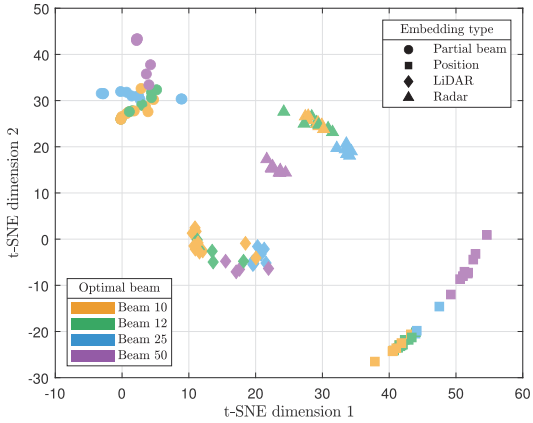
Fig. 4. Top-5 beam prediction accuracy comparison between four baselines and our proposed approach on scenario 9.

2) Latent Space Visualization: To intuitively show the effect of contrastive learning on the latent space, thereby demonstrating the superiority of our proposed approach over the existing multimodal approach, we perform the t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction on parts of the embeddings from the channel attention baseline [31] and the proposed contrastive learning approach on scenario 9, with the results presented in Fig. 5.

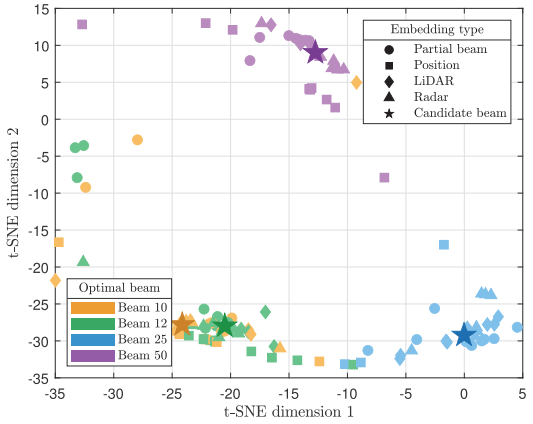
It can be clearly seen from Fig. 5(a) that the embeddings of the channel attention baseline [31] are clustered into four distinct groups, where each group corresponds to a specific modality. Within each modality cluster, the embeddings corresponding to different optimal beams are mixed together without clear separation, making it difficult for subsequent networks to leverage the discriminative information, and necessitating complex modality fusion and beam selection mechanisms.

By contrast, Fig. 5(b) shows that the proposed contrastive learning approach produces a fundamentally different latent space structure, where the embeddings from different modalities corresponding to the same optimal beam are clustered

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.



(a) t-SNE visualization of embeddings of channel attention baseline [31]



(b) t-SNE visualization of embeddings of proposed contrastive learning approach

Fig. 5. Comparison of t-SNE visualizations between embeddings of channel attention baseline [31] and proposed contrastive learning approach.

TABLE VIII

NORMALIZED BEAMFORMING GAIN COMPARISON BETWEEN DIFFERENT CANDIDATE BEAM EMBEDDING INITIALIZATION METHODS

Missing probability	0	0.25	0.5	0.75
Random projection	0.9332	0.9115	0.8703	0.8074
Low-rank factorization	0.9349	0.9123	0.8714	0.8081

together around their candidate beam embedding, forming distinct clusters organized by beam identity rather than modality type. Moreover, the spatial relationships among different beam clusters reveal a well-structured latent space. As shown in Fig. 5(b), the embeddings corresponding to beam 12 are positioned closest to those of beam 10, followed by beam 25, while the embeddings corresponding to beam 50 are the farthest, reflecting the underlying similarities in their spatial coverage properties. This visualization vividly demonstrates that contrastive training can create a unified latent space where embeddings are organized by beam identity instead of modality-specific characteristics. Such a semantically meaningful structure enables the simple averaging based fusion to effectively combine the information from different modalities.

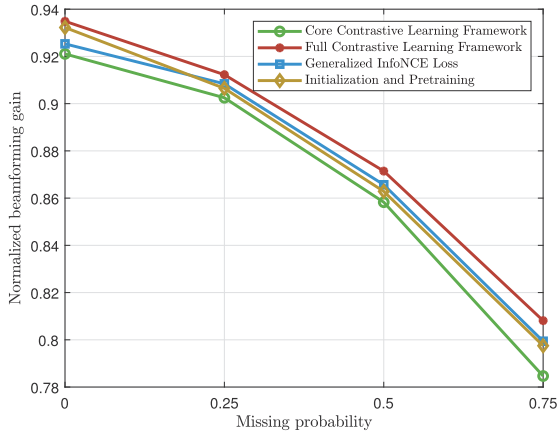


Fig. 6. Ablation study on the two proposed methodological refinements on scenario 9.

3) Performance Comparison of Candidate Beam Embedding Initialization Methods: To investigate the impact of different candidate beam embedding initialization methods on the beam prediction performance, we compare the two initialization methods presented in Subsection III-C on scenario 9 under varying missing probability conditions. The detailed results are presented in Table VIII.

As shown in Table VIII, these two different initialization methods yield similar normalized beamforming gains across all the missing probability conditions. More specifically, the learned low-rank factorization achieves a marginal performance gain at the cost of the additional training overhead, which reveals a complexity-performance tradeoff between these two different initialization methods, allowing them to be flexibly selected in practice. Based on the slight advantage observed in the above initialization comparison, the learned low-rank factorization based initialization is adopted as the default in the following ablation studies.

4) Ablation Study: To validate the effectiveness of the two proposed methodological refinements, we conduct an ablation study over four configurations, namely the core contrastive learning framework without either refinement, the core framework enhanced by only one of the two refinements, and the full framework with both of them. Fig. 6 compares the normalized beamforming gains of these four configurations as a function of missing probability.

It can be seen from Fig. 6 that the core framework attains the lowest normalized beamforming gain among the four configurations, while introducing either of the two refinements into the core framework consistently improves the beam prediction performance. In addition, the full framework equipped with both refinements achieves the best performance under all the missing probability conditions. Specifically, the improvement brought by the candidate beam embedding initialization and unimodal encoder pretraining strategy is attributed to the semantically meaningful initial structure within the latent space established by it before the joint contrastive training. Besides, the improvement achieved by the generalized InfoNCE loss validates that the extra supervisory signals in the generalized InfoNCE loss help the model learn more

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

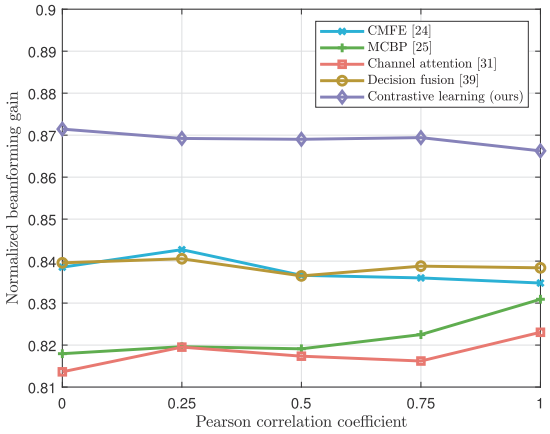


Fig. 7. Normalized beamforming gain comparison between four baselines and our proposed approach under varying correlation between partial beam and LiDAR missing patterns.

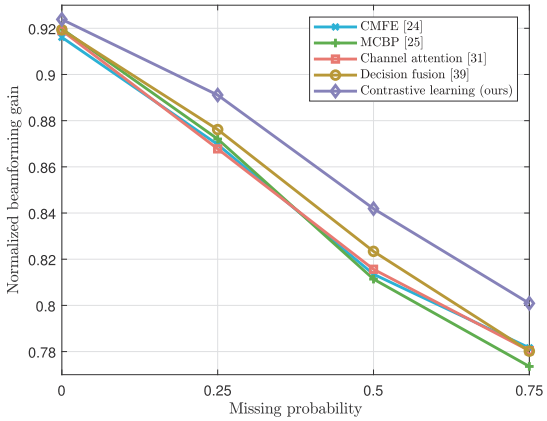


Fig. 8. Normalized beamforming gain comparison between four baselines and our proposed approach on scenario 8.

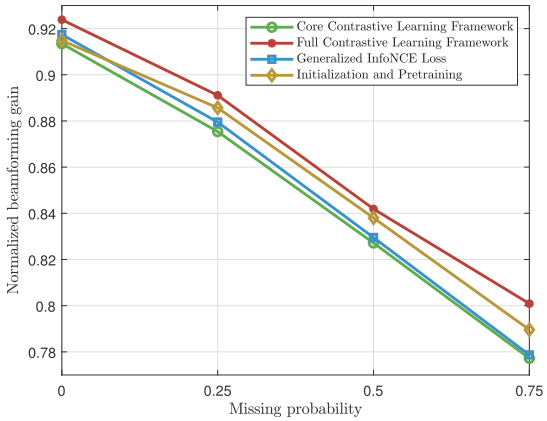


Fig. 9. Ablation study on the two proposed methodological refinements on scenario 8.

semantically meaningful embeddings. Moreover, when the modality missing probabilities are high, the performance gain brought by the generalized InfoNCE loss is more significant, as the positive embedding pairs are more sparse in these scenarios, and the introduction of extra supervisory signals enables more sufficient contrastive training.

5) Investigation of Robustness to Correlated Missing Modalities: In real-world V2I deployments, the missing patterns of different modalities are usually not strictly independent. To evaluate the robustness of the proposed framework under correlated missing patterns, we consider a representative case where the missing patterns of the partial beam and LiDAR modalities are correlated with a varying Pearson correlation coefficient [44], while the other modalities remain independently missing. The marginal missing probability of each modality is fixed at 0.5. The simulation results on scenario 9 are presented in Fig. 7.

It can be observed from Fig. 7 that the performance improvement of the proposed framework over the baselines remains essentially invariant as the Pearson correlation coefficient varies, which demonstrates the robustness of the proposed framework to correlated missing patterns. This robustness arises because the proposed framework processes each available modality individually through its unimodal encoder and then performs embedding averaging over the available modalities, rather than relying on any specific joint distribution of the missing patterns.

6) Investigation of Cross-Scenario Generalization Capability: To investigate the cross-scenario generalization capability of the proposed framework, we replicate the core experiments on scenario 8 of the DeepSense6G dataset, including the performance comparison against the four baselines and the ablation study on the proposed methodological refinements. The simulation results can be found in Figs. 8 and 9.

As shown in Fig. 8, the proposed framework still consistently achieves the highest normalized beamforming gain across all the missing probability conditions on scenario 8, and the performance gain it brings becomes more significant as the missing probability increases. Furthermore, the ablation results in Fig. 9 confirm that both the candidate beam embedding initialization with unimodal encoder pretraining and the generalized InfoNCE loss continue to bring noticeable performance improvements on scenario 8. These results together demonstrate that the proposed framework generalizes well across different V2I scenarios.

V. Conclusion

In this paper, a novel contrastive learning based multimodal beam prediction approach has been proposed to address the dual challenges of missing modality accommodation and high inference complexity. The proposed approach constructs a unified latent space by aligning modality embeddings from unimodal encoders with a set of learnable candidate beam embeddings through contrastive learning, which enables low-complexity averaging based modality fusion with adaptability to arbitrary missing patterns and similarity search based beam selection. To establish a semantically meaningful initial structure within the latent space before joint contrastive training, the candidate beam embeddings are initialized by compressing dataset-wide beamforming gain patterns into the latent space, followed by the pretraining of unimodal encoders with these initialized embeddings. Furthermore, the standard InfoNCE loss has been generalized to provide richer supervisory signals by incorporating high-relevance embedding pairs into the loss

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.

calculation. Simulation results have demonstrated that the proposed approach significantly outperforms its traditional counterparts across various missing modality scenarios while achieving substantially lower complexity. Extending the equal-weight averaging based fusion to a confidence-aware weighted counterpart for the scenarios with highly heterogeneous or rapidly varying modality qualities is worth investigating in the future.

REFERENCES

[1] M. Xiao et al., "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

[2] Q. Xue et al., "A survey of beam management for mmWave and THz communications towards 6G," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 1520–1559, 3rd Quart., 2024.

[3] T. S. Rappaport et al., "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[4] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[6] K. Ma, Z. Wang, W. Tian, S. Chen, and L. Hanzo, "Deep learning for mmWave beam-management: State-of-the-art, opportunities and challenges," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 108–114, Aug. 2023.

[7] C. Jeong, J. Park, and H. Yu, "Random access in millimeter-wave beamforming cellular networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 180–185, Jan. 2015.

[8] J. Wang et al., "Beam codebook based beamforming protocol for multi-gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.

[9] J. Wang et al., "Beamforming codebook design and performance evaluation for 60GHz wideband WPANs," in *Proc. VTC Fall*, Anchorage, AK, USA, 2009, pp. 1–6.

[10] F. Dai and J. Wu, "Efficient broadcasting in ad hoc wireless networks using directional antennas," *IEEE Trans. Parallel Distrib. Syst.*, vol. 17, no. 4, pp. 335–347, Apr. 2006.

[11] J. Kim and A. F. Molisch, "Fast millimeter-wave beam training with receive beamforming," *J. Commun. Netw.*, vol. 16, no. 5, pp. 512–522, Oct. 2014.

[12] C. Qi, Y. Wang, and G. Y. Li, "Deep learning for beam training in millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, early access, Sep. 22, 2020, doi: 10.1109/TWC.2020.3024279.

[13] K. Ma, D. He, H. Sun, Z. Wang, and S. Chen, "Deep learning assisted calibrated beam training for millimeter-wave communication systems," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6706–6721, Oct. 2021.

[14] K. Ma, S. Du, H. Zou, W. Tian, Z. Wang, and S. Chen, "Deep learning assisted mmWave beam prediction for heterogeneous networks: A dual-band fusion approach," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 115–130, Jan. 2023.

[15] J. Morais, A. Bchboodi, H. Pezeshki, and A. Alkhateeb, "Position-aided beam prediction in the real world: How useful GPS locations actually are?" in *Proc. ICC*, 2023, pp. 1824–1829.

[16] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 212–216, Feb. 2023.

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

[18] K. Zheng, H. Yang, Z. Ying, P. Wang, and L. Hanzo, "Vision-assisted millimeter-wave beam management for next-generation wireless systems: Concepts, solutions, and open challenges," *IEEE Veh. Technol. Mag.*, vol. 18, no. 3, pp. 58–68, Sep. 2023.

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

[20] 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, Mar. 2022.

[21] K. Patel and R. W. Heath Jr., "Harnessing multimodal sensing for multi-user beamforming in mmWave systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18725–18739, Dec. 2024.

[22] S. Tariq, B. E. Arfeto, U. Khalid, S. Kim, T. Q. Duong, and H. Shin, "Deep quantum-transformer networks for multimodal beam prediction in ISAC systems," *IEEE Internet Things J.*, vol. 11, no. 18, pp. 29387–29401, Sep. 2024.

[23] B. Shi, M. Li, M.-M. Zhao, M. Lei, and L. Li, "Multimodal deep learning empowered millimeter-wave beam prediction," in *Proc. IEEE 99th Veh. Technol. Conf. (VTC-Spring)*, Singapore, Jun. 2024, pp. 1–6.

[24] Q. Zhu, Y. Wang, W. Li, H. Huang, and G. Gui, "Advancing multimodal beam prediction with cross-modal feature enhancement and dynamic fusion mechanism," *IEEE Trans. Commun.*, vol. 73, no. 9, pp. 7931–7940, Sep. 2025.

[25] S. Lv et al., "Semantic-aware beam prediction approach with multimodal fusion," *IEEE Trans. Veh. Tech.*, early access, Oct. 1, 2025, doi: 10.1109/TVT.2025.3616031.

[26] J. Wang et al., "When vision-language model (VLM) meets beam prediction: A multimodal contrastive learning framework," *IEEE Trans. Veh. Tech.*, early access, Feb. 10, 2026, doi: 10.1109/TVT.2026.3663461.

[27] C. Zheng, J. He, C. G. Kang, G. Cai, Z. Yu, and M. Debbah, "M2BeamLLM: Multimodal sensing-empowered mmWave beam prediction with large language models," 2025, *arXiv:2506.14532*.

[28] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2302–2310.

[29] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 15878–15887.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[31] L. Yao, K. Ma, P. Wang, Y. Sang, and Z. Wang, "Robust multimodal beam prediction with missing modality," *IEEE Wireless Commun. Lett.*, vol. 14, no. 10, pp. 3264–3268, Oct. 2025.

[32] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[33] F. Liu, J. Pan, and R. Zhou, "Contrastive learning-based multimodal fusion model for automatic modulation recognition," *IEEE Commun. Lett.*, vol. 28, no. 1, pp. 78–82, Jan. 2024.

[34] X. Zhang, H. Zhang, C. Li, Y. Huang, and L. Yang, "Environment-specific beam training for extremely large-scale MIMO systems via contrastive learning," *IEEE Commun. Lett.*, vol. 27, no. 10, pp. 2638–2642, Oct. 2023.

[35] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2001, pp. 245–250.

[36] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[37] A. Alkhateeb et al., "DeepSense 6G: A large-scale real-world multimodal sensing and communication dataset," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, Sep. 2023.

[38] T. Sarlos, "Improved approximation algorithms for large matrices via random projections," in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2006, pp. 143–152.

[39] Y. Yang, F. Gao, C. Xing, J. An, and A. Alkhateeb, "Deep multimodal learning: Merging sensory data for massive MIMO channel prediction," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1885–1898, Jul. 2021.

[40] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed., Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[41] *Physical Channels and Modulation* (Release 18), document TS 38.211, 3GPP, 2024.

[42] *Physical Layer Procedures for Control (Release 18)*, document TS 38.213, 3GPP, 2025.

[43] M. Hou et al., "Beam orbital parameter prediction based on the deployment of cascaded neural networks at edge intelligence acceleration nodes," *Electronics*, vol. 13, no. 21, p. 4189, Oct. 2024.

[44] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Amer. Statistician*, vol. 42, no. 1, pp. 59–66, Feb. 1988.

Authorized licensed use limited to: UNIVERSITY OF SOUTHAMPTON. Downloaded on October 18,2026 at 17:04:03 UTC from IEEE Xplore. Restrictions apply.