ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11460605

BEAM-CLIP: MULTIMODAL ALIGNMENT AND MMWAVE BEAM PATTERN REPRESENTATION LEARNING

Andrew El Kommos, Saba Mohammadhosseini, and Nazanin Rahnavard

Department of Electrical and Computer Engineering
University of Central Florida, USA

ABSTRACT

This paper introduces **Beam-CLIP**, a contrastive learning framework for aligning mmWave beam "fingerprints" with multimodal sensor data. By adapting the Contrastive Language-Image Pre-training (CLIP) framework, Beam-CLIP maps camera, radar, LiDAR, and GPS observations into a joint embedding space with mmWave channel power vectors, enabling cross-modal retrieval of beam patterns. A soft retrieval mechanism identifies beams with similar spatial power distributions, producing a continuous channel power state estimate across all 64 beams. This approach improves beam prediction particularly under challenging non-line-of-sight scenarios.

***Index Terms*—** multimodal contrastive learning, mmWave beam prediction, sensor fusion, soft retrieval, CLIP

1. INTRODUCTION

The evolution toward 6G wireless systems introduces new challenges for base station intelligence, particularly in reliably predicting mmWave beams under dynamic, environment-dependent conditions. While mmWave links offer large bandwidth gains, they suffer from high path loss and blockage sensitivity, requiring precise beam steering and environmental awareness [1]. Recent datasets such as DeepSense 6G [2] pair mmWave measurements with multimodal sensors (camera, LiDAR, radar, GPS), enabling machine learning to model channel behavior and improve beam selection.

In mmWave beam prediction (BP), machine learning addresses the high cost of exhaustive beam search, with DeepSense providing a multimodal testbed. Early work used single-modality networks [3, 4, 5, 6], followed by stage-wise and hierarchical fusion [7, 8], and pseudo-labeling from semantic features [9], though most relied on fixed sensor setups and handcrafted pipelines. Transformers [10] have since become central for multimodal fusion, with masked variants such as Zorro [11] using learnable routing masks to preserve modality-specific information and remain robust under missing sensors.

Contrastive learning has proven effective for aligning diverse modalities in shared embedding spaces. CLIP [12] pioneered text-image alignment with InfoNCE loss [13], inspiring extensions to audio-visual [14], video [15], geo-location [16], and radar [17].

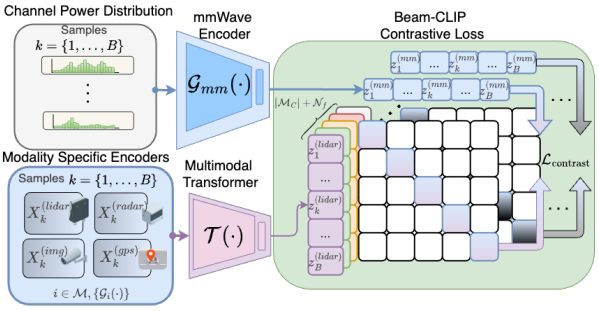

Fig. 1. **Beam-CLIP pairwise contrastive loss.** A multimodal transformer $\mathcal{T}(\cdot)$ produces modality-specific tokens $z^{(s)}$, and the mmWave encoder $\mathcal{G}_{\text{mm}}(\cdot)$ outputs $z^{(\text{mm})}$, both defined over a batch of size B. Pairwise token cycling is performed across the batch over all modality-to-modality and modality-to-fusion pairs, with similarity scores (5) being used for directional InfoNCE losses (6) which are aggregated into the total objective (7).

In this paper, we propose **Beam-CLIP**, a *cross-modal alignment framework* that learns a shared embedding space between sensor observations and mmWave channel power vectors. Treating the full channel power vector as a compact RF "fingerprint" of the environment, Beam-CLIP aligns multimodal sensor inputs with the entire beam distribution, enabling retrieval of the pattern most consistent with the sensed scene.

2. METHODOLOGY

Beam-CLIP projects all sensor modalities into a shared d-dimensional space and is trained with a pairwise contrastive loss objective (Fig. 1). A masked multimodal transformer with learnable fusion tokens [11] fuses modality embeddings for alignment. At inference, BP is performed as similarity search in the joint embedding space, retrieving the mmWave pattern most consistent with the sensed context (Fig. 2).

2.1. Beam-CLIP Overview

Beam-CLIP is designed to handle a set of sensor modalities $\mathcal{M} = \{\text{img}, \text{radar}, \text{lidar}, \text{gps}, \text{mm}\}$, where each element corresponds to a distinct raw sensor domain of image frames, radar returns, LiDAR point clouds, GPS readings, and mmWave channel features. For each modality $i \in \mathcal{M}$,

let $X^{(i)}$ denote its raw input. Each $X^{(i)}$ is encoded by a modality-specific encoder $\mathcal{G}_i(\cdot)$ into a shared d-dimensional space (with batch size B): $e^{(i)} = \mathcal{G}_i\big(X^{(i)}\big) \in \mathbb{R}^{B\times d}$. During training, the encoder outputs of non-mmWave modalities $\mathcal{M}_C = \mathcal{M}\backslash\{\text{mm}\}$ are concatenated with N_f learnable fusion tokens $U \in \mathbb{R}^{N_f\times d}$ to form the transformer input context:

$$C \in \mathbb{R}^{B\times(|\mathcal{M}_C|+N_f)\times d}. \tag{1}$$

The multimodal transformer encoder $\mathcal{T}(C) \in \mathbb{R}^{B\times(|\mathcal{M}_C|+N_f)\times d}$ maps the input context C to modality tokens $h^{(m)} \in \mathbb{R}^{B\times d},\ m \in \mathcal{M}_C$ and fusion tokens $h_1^{(\text{fus})},\ldots,h_{N_f}^{(\text{fus})} \in \mathbb{R}^{B\times d}$:

$$\mathcal{T}(C) = [\underbrace{h^{(m)} : m \in \mathcal{M}_C}_{\text{modality tokens}}, \underbrace{h_1^{(\text{fus})},\ldots,h_{N_f}^{(\text{fus})}}_{\text{fusion tokens}}]. \tag{2}$$

A single fused representation is obtained by averaging:

$$h^{(\text{fus})} = \frac{1}{N_f}\sum_{r=1}^{N_f} h_r^{(\text{fus})}. \tag{3}$$

Additionally, all transformer output tokens ($h^{(s)}$ for $s \in \mathcal{M}_C \cup \{\text{fus}\}$) are ℓ_2-normalized to unit vectors row-wise to form $z^{(s)}$. Similarly, $e^{(\text{mm})}$, the output of mmWave encoder $\mathcal{G}_{\text{mm}}(\cdot)$ is normalized to $z^{(\text{mm})}$, and the full set $\{z^{(s)}\}_{s\in\mathcal{M}\cup\{\text{fus}\}}$ is used for multimodal alignment (Sec. 2.4).

2.2. Modality-Specific Encoders

Each modality $i \in \mathcal{M}$ produces a raw input $X^{(i)}$, which is mapped into the shared embedding space using a modality-specific encoder $\mathcal{G}_i(\cdot)$. The encoders process each modality as follows:

- **mmWave:** The received power vector over 64 beams, $X^{(\text{mm})} \in \mathbb{R}^{64}$, is processed by a multilayer perceptron (MLP) to capture spatial power distributions.
- **Camera:** The vector of image frames, $X^{(\text{img})}$, is encoded by a ResNet-18 backbone (classification head removed) and the extracted visual features are projected into the embedding space via a MLP.
- **Radar:** Range-angle maps $X^{(\text{radar})} \in \mathbb{R}^{256\times 8}$ are processed with 2D convolutions [1].
- **LiDAR:** Point clouds are projected into a bird's-eye view grid [7], $X^{(\text{lidar})} \in \mathbb{R}^{100\times 100}$, then encoded via MLP.
- **GPS:** $X^{(\text{gps})} = (x, y, \sqrt{x^2+y^2}, \theta, \sin\theta, \cos\theta) \in \mathbb{R}^6$, with $\theta = \arctan 2(y, x)$, projected via MLP.

Modality-specific embeddings $e^{(i)}$ are concatenated with learnable fusion tokens and passed through $\mathcal{T}(\cdot)$ for contrastive training.

2.3. Multimodal Transformer

Beam-CLIP employs a multimodal transformer based on a prior architecture that leverages the specialized fusion tokens and masking technique [11]. The design has been adapted to our input modalities, and to operate over mmWave to sensor alignment mechanism.

2.3.1. Positional Ordering

A fixed token ordering is used when constructing the input context sequence C to the transformer, $C = [e^{(\text{img})}, e^{(\text{radar})}, e^{(\text{lidar})}, e^{(\text{gps})}, h^{(\text{fus},1)}, \ldots, h^{(\text{fus},N_f)}]$, with corresponding positional embeddings added before entering the transformer. This ensures consistent token indexing and consistent attention masking across samples.

2.3.2. Attention Mask

During training, modality tokens attend only to themselves and fusion tokens, while fusion tokens attend to all tokens. This is enforced with a binary mask:

$$A = [a_{rc}], \quad a_{rc} = \begin{cases} 1 & \text{if token } r \text{ can attend to token } c, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where $r, c \in \{1,\ldots,|\mathcal{M}_C| + N_f\}$. Applied to the $QK^\top$ term in self-attention, this implements the Zorro [11] scheme, ensuring that cross-modal exchange occurs only through fusion tokens. Combined with structured modality dropout (Sec. 2.4.1) and fixed token ordering (Sec. 2.3.1), the mask guarantees consistent interaction patterns even when sensor modalities are randomly dropped.

2.4. Beam-CLIP Pre-Training

Beam-CLIP has a training paradigm that is centered around contrastive losses. We first discuss a technique that aims to make the multimodal transformer $\mathcal{T}(\cdot)$ robust to modality dropout and then we go into details on the contrastive learning technique for embedding alignment.

2.4.1. Modality Dropout

To improve robustness, Beam-CLIP applies structured modality dropout before the transformer $\mathcal{T}(\cdot)$. During training, each modality in $\mathcal{M}_C$ is independently dropped with a fixed probability p_d, and its encoder output in the context window C is replaced with zeros. The number of active modalities therefore follows a binomial distribution with retention probability $(1-p_d)$. Fusion tokens remain active across all samples, ensuring the transformer always forms a consistent global context even when modalities are missing.

2.4.2. Contrastive Learning

Beam-CLIP (Fig. 1) adapts the contrastive framework of CLIP [12] to multimodal sensing data. Normalized embeddings from each modality and the pooled fusion representation are compared across pairs, with a directional InfoNCE loss applied to encourage cross-modal alignment while separating mismatched samples. This objective is introduced below, beginning with the definition of similarity scores.

Similarity scores. For each pair of modalities $(i, j) \in \mathcal{P}$, we compute a similarity matrix $S^{(i,j)} \in \mathbb{R}^{B\times B}$ with its (k, l) entry is given by:

$$S^{(i,j)}[k,l] = \frac{z_k^{(i)} \cdot z_l^{(j)}}{\tau_c}, \qquad k, l = 1,\ldots,B, \tag{5}$$

where $z_k^{(i)}$ and $z_l^{(j)}$ are the ℓ_2-normalized embeddings of the k-th and l-th samples from modalities i and j, respectively,

τ_c is the contrastive temperature, and $\cdot$ denotes the Euclidean dot product. Since all embeddings are unit-normalized, this dot product is equivalent to cosine similarity.

Directional InfoNCE losses. The contrastive loss is applied in both directions:

$$\mathcal{L}_{i\to j} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp(S^{(i,j)}[k,k])}{\sum_{l=1}^{B}\exp(S^{(i,j)}[k,l])}. \tag{6}$$

Pairwise contrastive loss. Beam-CLIP employs a pairwise contrastive loss across all relevant pairs, including modality-to-modality and modality-to-fusion alignments. Averaging over all pairs $(i,j)\in\mathcal{P}$ gives:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}(\mathcal{L}_{i\to j}+\mathcal{L}_{j\to i}). \tag{7}$$

Reconstruction loss. Fusion tokens $h_1^{(\text{fus})},\ldots,h_{N_f}^{(\text{fus})}$ are regressed to predict the mmWave beam pattern $\hat{y}_k$ for each sample k, with supervision from the ground-truth vector y_k:

$$\mathcal{L}_{\text{recon}} = \tfrac{1}{B}\sum_{k=1}^{B}\|\hat{y}_k - y_k\|_2^2. \tag{8}$$

Total objective. The final pre-training loss combines alignment and reconstruction given as

$$\mathcal{L}_{\text{total}} = \lambda_c\mathcal{L}_{\text{contrast}} + \lambda_r\mathcal{L}_{\text{recon}}, \tag{9}$$

where λ_c and λ_r are weighting coefficients.

2.5. Multimodal Beam Retrieval Mechanism

The retrieval objective is to return the most semantically relevant mmWave beam pattern given a query containing any subset of modalities $\mathcal{M}_C$ (Fig. 2). To this end, we first construct a retrieval database from the training set. For each mmWave beam power vector $y_i\in\mathbb{R}^{64}, i=1,\ldots,N_{\text{train}}$, we obtain a normalized mmWave embedding using the trained mmWave encoder $\mathcal{G}_{mm}(\cdot)$. The database is then defined as $\mathcal{D}=\{(z_i^{(\text{mm})}, y_i)\}_{i=1}^{N_{\text{train}}}$.

Query encoding. Given a test query x_q composed of input modalities $\mathcal{M}_q\subseteq\mathcal{M}$, we encode it with the trained $\mathcal{T}(\cdot)$ to obtain N_f fusion token outputs $\{h_j^{(\text{fus})}\}_{j=1}^{N_f}$. The fusion tokens are averaged into a pooled embedding (3), ℓ_2-normalized to yield the fusion representation f_q, which serves as the retrieval query embedding.

Retrieval similarity. For a query embedding f_q, the similarity with each database element $z_i^{(\text{mm})}$ is computed as

$$S^{(q,\text{mm})}[1,i] = \frac{f_q\cdot z_i^{(\text{mm})}}{\tau_r}, \qquad i=1,\ldots,N_{\text{train}}, \tag{10}$$

where $S^{(q,\text{mm})}\in\mathbb{R}^{1\times N_{\text{train}}}$ collects the query-database similarities, and τ_r is a retrieval-specific temperature parameter controlling the sharpness of similarity-based weighting.

Soft retrieval. Given similarities $S^{(q,\text{mm})}[1,i]$ from (10), we select the top-K entries and form a weighted reconstruction:

$$\hat{y}_q = \sum_{i=1}^{K} w_i y_i, \qquad w_i = \frac{\exp\big(S^{(q,\text{mm})}[1,i]\big)}{\sum_{j=1}^{K}\exp\big(S^{(q,\text{mm})}[1,j]\big)}. \tag{11}$$

The top-K matches are denoted as $\{S_1,S_2,\ldots,S_K\}$, corresponding to the highest similarity scores $S^{(q,\text{mm})}[1,i]$. The soft weighting yields the final continuous estimate $\hat{y}_q$.

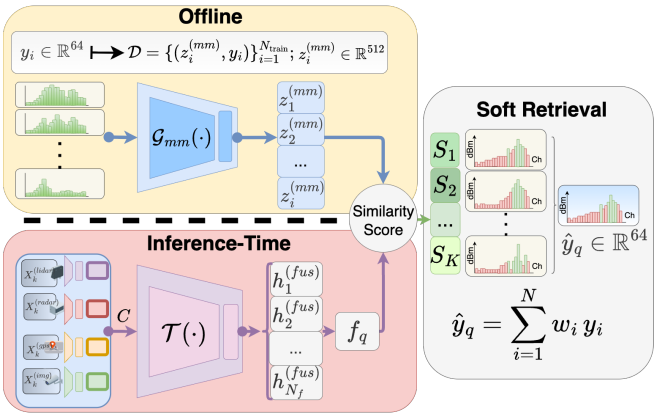

Fig. 2. **Beam-CLIP soft retrieval (inference).** The database $\mathcal{D}$ is prebuilt offline by encoding mmWave samples with $\mathcal{G}_{\text{mm}}(\cdot)$. At inference time, a query C (1) is processed by the multimodal transformer $\mathcal{T}(\cdot)$, producing fusion token embeddings (3). These are averaged into the query representation f_q, which is compared via retrieval similarity (10) against all mmWave embeddings z_i^{mm}. The top-K matches $\{S_1,S_2,\ldots,S_K\}$ are selected, and a soft weighting step yields the final continuous estimate $\hat{y}_q$ (11).

3. EXPERIMENTAL SETUP

Employing the soft retrieval framework, Beam-CLIP's experimental setup evaluates the pre-trained network's BP performance. We compare prior methods [7] that trained models for specific sensor modalities configurations with the best Beam-CLIP retrieval results to establish a baseline for both Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) performance. Beam-CLIP employs an L-layer multimodal transformer $\mathcal{T}(\cdot)$ with h attention heads, d-dimensional hidden states, and N_f learnable fusion tokens for cross-modal aggregation. Each attention layer uses a dropout rate r_d. Full architectural and training hyperparameters are summarized in Table 1.

Table 1. Experimental configuration hyper parameters.

Architecture $\mathcal{T}(\cdot)$		**Training**	
Hidden Dim (d)	512	Optimizer	AdamW
Heads (h)	8	Learning Rate (L_r)	1×10^{-4}
Layers (L)	6	Weight Decay (W_d)	0.01
Dropout (r_d)	0.1	Batch Size (B)	32
Fusion Tokens (N_f)	8	Epochs (e)	50
Total Params	72.3M	Model Size	275.9 MB
Loss Configuration		**Modality Dropout**	
Contrastive (λ_c)	1.0	Enabled	Yes
Reconstruction (λ_r)	1.0	Min Modalities ($m_{\min}$)	1
Temperature (τ_c)	0.07	Max Modalities ($m_{\max}$)	5
Loss Type	InfoNCE+MSE	Dropout Prob. (p_d)	0.2
Retrieval Configuration			
Top-K (K)	5	Similarity Metric	Cosine
Retrieval Temp. (τ_r)	0.07	Database D Sample Size	8,158

3.1. Dataset and Preprocessing

We evaluate our proposed scheme on three scenarios (32, 33, 34) from the DeepSense 6G dataset [2], totaling 11,656 samples with 70/15/15 train, validation, and test split, respectively. Samples are classified as LoS/NLoS using a GPS-based technique [7], resulting in 1618 LoS and 99 NLoS samples for our test set.

Table 2. Beam prediction performance: LoS vs NLoS comparison. Prior work implemented based on [7].

Method	LoS						NLoS					
	Top-1A	Top-3A	DBA	Top-3B	Top-3,3BA	PR@3	Top-1A	Top-3A	DBA	Top-3B	Top-3,3BA	PR@3
Prior: LiDAR	0.37	0.71	0.81	0.70	0.88	0.97	0.03	0.14	0.18	0.25	0.43	0.93
Prior: Image	0.44	0.81	0.88	0.81	0.96	**0.99**	0.05	0.12	0.17	0.28	0.48	0.93
Prior: Radar	0.38	0.75	0.84	0.75	0.92	0.98	0.06	0.10	0.16	0.25	0.40	0.92
Prior: GPS	0.45	0.81	0.88	0.81	0.96	**0.99**	0.01	0.06	0.08	0.28	0.41	0.92
Prior: All	**0.46**	0.83	**0.89**	**0.84**	**0.97**	**0.99**	0.04	0.11	0.17	0.25	0.49	0.93
Beam-CLIP Retrieval (All)	0.44	**0.84**	**0.89**	0.82	0.96	**0.99**	0.37	**0.78**	**0.86**	0.77	**0.93**	**0.99**
Beam-CLIP Retrieval (G+I+L)	0.45	**0.84**	**0.89**	0.82	0.96	**0.99**	**0.40**	**0.78**	**0.86**	**0.78**	**0.93**	**0.99**
Beam-CLIP Retrieval (G+I+R)	0.36	0.68	0.76	0.67	0.80	0.94	0.28	0.56	0.69	0.53	0.71	0.92

3.2. Evaluation Metrics

We evaluate Beam-CLIP using the following BP metrics. *Top-K_A* Accuracy verifies whether the ground-truth optimal beam index is contained within the model's top-K predictions [2]. *Top-K_B Accuracy* checks if the beam predicted with the highest power lies within the top-K beams ranked by actual received power [7, 18]. *Top-(K_1, K_2) Beams Accuracy (Top-(K_1, K_2) BA)* measures how many of the top-K_2 predicted beams align with the top-K_1 beams from the ground-truth ranking [7, 18]. *Distance-Based Accuracy (DBA)* considers a prediction correct if the selected beam lies within an angular neighborhood of the ground-truth optimal beam [2]. *Power Ratio (PR@K)* computes the ratio between the maximum received power among the top-K predicted beams and the global maximum across all beams [7, 18].

4. RESULTS AND DISCUSSION

We evaluate Beam-CLIP across both LoS and NLoS scenarios using image (I), GPS (G), LiDAR (L), and radar (R) modalities to assess its BP performance (Table 2). All comparisons are performed against strong late-fusion baselines evaluated on the same benchmark [7], using their best-performing configurations. We observed that Beam-CLIP (All) and (G+I+L) are the top-performing variants and perform similarly, while (G+I+R) is lower on all metrics, suggesting LiDAR provides more reliable cues than radar under the current preprocessing and training setup.

4.1. LoS vs NLoS Performance

Beam-CLIP demonstrates exceptional performance improvements in challenging NLoS scenarios compared to prior methods [7]. While previous approaches achieve only 6% Top-1A and 14% Top-3A under NLoS conditions, Beam-CLIP retrieval mode reaches 40% and 78%, respectively.

4.2. Real-Time Performance and Deployment Viability

Beam-CLIP supports real-time inference with an average end-to-end latency of 17.5 ms (57 samples/s throughput), including 1.95 ms for retrieval over a 10,000 entry database, while operating within a < 1 GB VRAM budget. The model has 72.3M parameters (276 MB). Additional evaluation tests scalability using a synthetic embedding database of up to 1,000,000 entries, showing retrieval latency that increases linearly with database size without requiring changes to the model or inference pipeline. These results indicate Beam-CLIP's is suitable for deployment in resource-constrained and latency-sensitive edge environments.

4.3. Discussion

Beam-CLIP's effectiveness arises from three factors: (1) a masked multimodal transformer that enables flexible sensor fusion and robustness to missing modalities; (2) contrastive pre-training that yields semantically aligned embeddings transferable to downstream tasks; and (3) an architecture that frames BP as retrieval. The stronger gains under NLoS relative to LoS stem from the interaction between cross-modal contrastive alignment, modality dropout during training, and soft beam retrieval at inference. Beam-CLIP is trained on all modalities while randomly dropping subsets, explicitly encouraging robustness to occlusions and missing sensors common at test time. At inference, retrieving and softly aggregating beams from similar latent contexts captures shared multipath structure, producing larger benefits in NLoS settings where simpler predictors struggle. Although NLoS evaluation is limited by few samples, the model consistently improves NLoS performance over prior work on the same dataset [7].

Our approach emphasizes end-to-end system behavior and adopts established multimodal design practices, including fusion tokens, auxiliary reconstruction losses, and modality dropout. Future work will evaluate robustness, generalization to unseen scenarios, and scalability by testing on larger, more diverse mmWave datasets and broader NLoS conditions, including higher dataset sizes and compute budgets. Moreover, while we adopt a soft-retrieval mechanism for BP, alternative retrieval strategies may also prove effective.

5. CONCLUSION

This paper presents Beam-CLIP, a multimodal framework that extends CLIP to mmWave beam prediction by contrastively aligning heterogeneous sensor modalities with beam patterns. We showed that CLIP-style contrastive learning effectively aligns environmental sensor data with mmWave channels, enabling accurate retrieval and real-time beam prediction. Beam-CLIP achieves strong performance, particularly in challenging NLoS scenarios, underscoring its feasibility for practical deployment.

6. REFERENCES

[1] Gouranga Charan, Umut Demirhan, João Morais, Arash Behboodi, Hamed Pezeshki, and Ahmed Alkhateeb, "Multi-modal beam prediction challenge 2022: Towards generalization," *arXiv preprint arXiv:2209.07519*.

[2] Ahmed Alkhateeb, Gouranga Charan, Tawfik Osman, Andrew Hredzak, Joao Morais, Umut Demirhan, and Nikhil Srinivas, "Deepsense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 122–128, 2023.

[3] João Morais, Arash Bchboodi, Hamed Pezeshki, and Ahmed Alkhateeb, "Position-aided beam prediction in the real world: How useful gps locations actually are?," in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 1824–1829.

[4] Umut Demirhan and Ahmed Alkhateeb, "Radar aided 6G beam prediction: Deep learning algorithms and real-world demonstration," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 2655–2660.

[5] Gouranga Charan and Ahmed Alkhateeb, "Computer vision aided blockage prediction in real-world millimeter wave deployments," in *2022 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2022, pp. 1711–1716.

[6] Shuaifeng Jiang, Gouranga Charan, and Ahmed Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Communications Letters*, vol. 12, no. 2, pp. 212–216, 2022.

[7] Katarina Vuckovic, Saba M. Hosseini, and Nazanin Rahnavard, "Revisiting performance metrics for multimodal mmwave beam prediction using deep learning," in *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)*. IEEE, 2024, pp. 881–887.

[8] Yu Tian, Qiyang Zhao, Fouzi Boukhalfa, Kebin Wu, Faouzi Bader, et al., "Multimodal transformers for wireless communications: A case study in beam prediction," *arXiv preprint arXiv:2309.11811*, 2023.

[9] Shoaib Imran, Gouranga Charan, and Ahmed Alkhateeb, "Environment semantic aided communication: A real world demonstration for beam prediction," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, pp. 48–53.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] Adrià Recasens, Jason Lin, João Carreira, Drew Jaegle, Luyu Wang, Jean-baptiste Alayrac, Pauline Luc, Antoine Miech, Lucas Smaira, Ross Hemsley, et al., "Zorro: the masked multimodal transformer," *arXiv preprint arXiv:2301.09595*, 2023.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," in *arXiv preprint arXiv:1807.03748*, 2018.

[14] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "AudioCLIP: Extending CLIP to image, text and audio," 2021.

[15] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer, "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," in *arXiv preprint arXiv:2109.14084*, 2021.

[16] Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah, "GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization," in *Advances in Neural Information Processing Systems*, 2023.

[17] Yifan Xu et al., "PointCLIP: CLIP for 3D point cloud understanding," *arXiv preprint arXiv:2112.07630*, 2021.

[18] Saba M. Hosseini, Amirhossein Safari, and Nazanin Rahnavard, "Kolmogorov–arnold networks for multimodal feature fusion and mmwave beam prediction," in *MILCOM 2025 - 2025 IEEE Military Communications Conference (MILCOM)*, 2025, pp. 1341–1346.