Brasil 6G

Projeto Brasil 6G
Fase III

Atividade 3.1 - Aplicações de Visão Computacional para Verticais Estratégicas - Parte 2

31/07/2026

Histórico de Atualizações:

Versão	Data	Autor(es)	Notas
1	15/07/2026	Alcides Tomas - Eldorado Aldebaro Klautau – UFPA Cristiano Bonato Both – UNISINOS Felipe A. P. de Figueiredo – Inatel Francine Cassia de Oliveira - Eldorado Matheus Ferreira Silva – Inatel João Borges – UFPA José Ferreira de Rezende – UFRJ Wesley L. Passos – UFRJ	Elaboração de conteúdo
2	31/07/2026	Aldebaro Klautau – UFPA Luciano Leonel Mendes – Inatel Vanessa Mendes Rennó – Inatel	Revisão de texto

Lista de Figuras

1 Grade angular do *codebook* adotado, com os índices dos feixes representados no plano azimute e elevação. 4
2 Representação geométrica do vetor unitário de direção associado a um feixe, destacando sua projeção no plano (XY), sua componente vertical e as relações angulares definidas pelo azimute (ϕ) e pela elevação (θ). 6
3 Cenário experimental interno, incluindo as dimensões da sala, o sistema de coordenadas adotado, a posição da BS e a trajetória seguida pela plataforma robótica ao longo das posições do Rx. 7
4 Comparação dos erros de localização cartesianos nas componentes x e y para os métodos Top-1, baseado em subconjuntos e baseado em subconjuntos assistido por imagem. 10
5 Posições reais do Rx no cenário, juntamente com a localização do Tx e as regiões de cobertura azimutal da UPA e da câmera RGB. 11
6 Amostras de *frames* RGB da trajetória exp6, capturadas pela câmera Nerian Ruby RGB-D acoplada à estação base. O UE (Turtlebot 4) aparece em diferentes posições ao longo da trajetória. 14
7 Trajetórias exp6, exp7 e exp8, com as posições 3D do UE em relação à estação base. 14
8 Arquitetura proposta. 15
9 Estrutura geral do *framework* de predição de feixes baseado em LLM multimodal. 22

Lista de Tabelas

1 Desempenho de localização 2D para os métodos avaliados 9
2 Resultados de MAE_{3D} e $RMSE_{3D}$ por *fold holdout* e média ± desvio padrão, inter-trajetória (LOTO). 17
3 Comparação dos modelos avaliados no conjunto Raymobtime S009. 19
4 Acurácia Top-1: validação e teste. 23
5 Ablação das modalidades sensoriais (acurácia Top-1 de validação). 24

Acrônimos

5G Quinta Geração de Rede Móvel Celular
5G-NR 5G-*New Radio*
6G Sexta Geração de Rede Móvel Celular
AoD *Angle of Departure*
B5G *Beyond*-5G
BEV *Bird's Eye View*
BGAM *Beam-Guided Attention Masking*
BS *Base Station*
CARLA *Car Learning to Act*
CCT2 CONVERGE *Challenge Task 2*
CNN *Convolutional Neural Network*
CQI *Channel Quality Indicator*
CSI *Channel State Information*
FiLM *Feature-wise Linear Modulation*
FLOP *Floating Point Operation*
FoV *Field of View*
GMU *Gated Multimodal Unit*
GNSS *Global Navigation Satellite System*
GNN *Graph Neural Network*
GPS *Global Positioning System*
GPT *Generative Pre-trained Transformer*
GPU *Graphics Processing Unit*
IA Inteligência Artificial
ISAC *Integrated Sensing and Communications*
LiDAR *Light Detection and Ranging*
LLM *Large Language Model*
LoS *Line-of-Sight*
LOTO *Leave-One-Trajectory-Out*
MAE *Mean Absolute Error*
MIMO *Multiple Input Multiple Output*
MISO *Multiple Input Single Output*
MLP *Multilayer Perceptron*
mmWave *Millimeter Wave*
MOTPE *Multi-Objective Tree-structured Parzen Estimator*

NAS *Neural Architecture Search*

NLoS *Non-Line-of-Sight*

O-RAN *Open Radio Access Network*

PTP *Picture Transfer Protocol*

RF Radiofrequência

RGB *Red, Green, and Blue*

RGB-D RGB-*Depth*

RMSE *Root Mean Square Error*

RSRP *Reference Signal Received Power*

RSRQ *Reference Signal Received Quality*

RSS *Received Signal Strength*

RSSI *Received Signal Strength Indicator*

RSU *Road Side Unit*

Rx *Receiver*

SINR *Signal-to-Interference-plus-Noise Ratio*

SRS *Sounding Reference Signal*

SUMO *Simulation of Urban Mobility*

TPE *Tree-structured Parzen Estimator*

Tx *Transmitter*

UE *User Equipment*

UPA *Uniform Planar Array*

YOLO *You Only Look Once*

Sumário

1 Introdução **1**

2 Estimação do Posicionamento de Receptor de Rádio Baseada em RSSI e Imagens RGB em Ambientes Indoor Usando Equipamentos de Baixo Custo **2**
2.1 Modelagem do Sistema . . . 3
2.1.1 Sistema de Comunicação . . . 3
2.1.2 Mapeamento das *codewords* para as direções dos lóbulos principais . . . 3
2.1.3 Modelagem angular baseada em imagem . . . 4
2.1.4 Estimação geométrica da posição . . . 5
2.1.5 Estratégia de combinação por subconjuntos . . . 5
2.2 Descrição do Cenário e do Conjunto de Dados . . . 7
2.3 Método Proposto . . . 8
2.4 Resultados e Discussões . . . 8
2.4.1 Protocolo experimental e métricas de avaliação . . . 9
2.4.2 *Baseline* Top-1 . . . 9
2.4.3 Localização baseada em subconjuntos de feixes . . . 9
2.4.4 Localização por subconjuntos assistida por imagem . . . 10
2.5 Conclusões . . . 12

3 Localização 3D de Equipamentos de Usuário via Fusão Multimodal Rádio-Visual em Redes mmWave **13**
3.1 Metodologia . . . 13
3.1.1 Conjunto de dados . . . 13
3.1.2 Arquitetura proposta . . . 14
3.2 Resultados e Discussões . . . 16
3.3 Conclusões e Direções Futuras . . . 16

4 *Neural Architecture Search* Aplicada à Seleção Top-k de Feixes no Conjunto de Dados Raymobtime S008 e S009 **18**
4.1 Contextualização . . . 18
4.2 Metodologia . . . 18
4.3 Comparação dos Modelos . . . 19
4.4 Trabalhos Futuros . . . 20

5 Predição de Feixes com LLM Multimodal **21**
5.1 *Framework* de Predição de Feixes Baseado em LLM Multimodal . . . 21
5.1.1 Trabalhos relacionados . . . 21
5.1.2 Arquitetura implementada . . . 21
5.2 Resultados Preliminares . . . 23
5.2.1 Progressão do treinamento . . . 23
5.2.2 Análise dos resultados intermediários . . . 23
5.3 Proposta de Contribuição . . . 24

6 Conclusão **25**

1 Introdução

A Sexta Geração de Rede Móvel Celular (6G) deverá integrar comunicação, sensoriamento e Inteligência Artificial (IA) de forma nativa, permitindo a oferta de serviços avançados com elevados requisitos de confiabilidade, baixa latência e alta eficiência espectral. Nesse contexto, dados provenientes de diferentes modalidades sensoriais, como sinais de rádio, câmeras, *Light Detection and Ranging* (LiDAR) e outros dispositivos de percepção do ambiente, desempenham papel central no desenvolvimento de algoritmos capazes de aumentar a eficiência operacional da rede e viabilizar novas aplicações para verticais estratégicas, incluindo indústria, logística, cidades inteligentes, segurança pública, saúde e transporte autônomo.

Entre as diversas tecnologias habilitadoras da 6G, a visão computacional destaca-se por fornecer informações semânticas e geométricas complementares às obtidas pelos sistemas de comunicação. Quando combinadas por meio de técnicas de fusão multimodal e aprendizado de máquina, essas modalidades permitem aprimorar tarefas como localização de dispositivos, seleção de feixes em comunicações por *Millimeter Wave*s (mmWaves), monitoramento do ambiente e tomada de decisão baseada em contexto. Paralelamente, os avanços em arquiteturas profundas, métodos de otimização automática de modelos e modelos fundacionais de linguagem multimodais ampliam significativamente o potencial dessas aplicações, possibilitando soluções mais robustas, adaptáveis e eficientes.

Neste segundo relatório da Atividade 3.1 da Fase III do Projeto Brasil 6G são apresentados os avanços obtidos no desenvolvimento de técnicas de IA aplicadas à integração entre visão computacional e sistemas de comunicação sem fio. Os estudos realizados abordam diferentes aplicações dessa integração, incluindo métodos de localização baseados na fusão de dados de rádio e imagem, arquiteturas multimodais para localização tridimensional de equipamentos de usuário, técnicas de *Neural Architecture Search* (NAS) para seleção de feixes e investigações preliminares sobre o emprego de *Large Language Model*s (LLMs) multimodais na predição de feixes em sistemas mmWave.

Em conjunto, essas atividades investigam estratégias capazes de explorar simultaneamente informações provenientes da interface de rádio e de sensores visuais, utilizando equipamentos comerciais de baixo custo e dados experimentais obtidos em ambientes reais. Além disso, são avaliadas arquiteturas de aprendizado profundo para problemas de seleção de feixes em comunicações mmWave, considerando não apenas o desempenho em termos de acurácia, mas também aspectos relevantes para futuras implementações práticas, como complexidade computacional, latência e eficiência energética. Também são exploradas abordagens baseadas em modelos multimodais de grande porte, com o objetivo de avaliar seu potencial para incorporar conhecimento contextual e múltiplas fontes de informação ao processo de tomada de decisão da rede. Os capítulos seguintes apresentam os avanços obtidos em cada uma dessas frentes de pesquisa, destacando as metodologias empregadas, os principais resultados alcançados e as perspectivas de continuidade.

2 Estimação do Posicionamento de Receptor de Rádio Baseada em RSSI e Imagens RGB em Ambientes Indoor Usando Equipamentos de Baixo Custo

As capacidades do Wi-Fi [1] e da Quinta Geração de Rede Móvel Celular (5G) [2] permitem explorar casos de uso relacionados ao posicionamento de usuários em ambientes internos, especialmente devido à alta diretividade da propagação em mmWave. Além disso, essa estimação de posição baseada em sinais de rádio pode ser aprimorada por meio da fusão de informações multimodais, tais como imagens [3]. No entanto, muitos trabalhos dependem de equipamentos sofisticados, como radares e câmeras especiais, bem como da disponibilidade de dados oriundos de sistemas de comunicação, como por exemplo, a *Channel State Information* (CSI), o que nem sempre está disponível em equipamentos comerciais e de baixo custo, aumentando o custo operacional desses sistemas.

A literatura acerca de localização usando ondas de rádio é vasta. Por exemplo, em [3], Zhang et al. utilizam radares mmWave especializados para realizar a localização e a refinam com dados sintéticos de Radiofrequência (RF) obtidos a partir das formas de objetos detectados por meio de imagens de câmeras e visão computacional. Bai et al. [1] também utilizam *Received Signal Strength* (RSS) para localização, mas consideram apenas Wi-Fi tradicional (2,4 GHz/5 GHz) e refinam a localização com o uso de imagens capturadas no lado do *Receiver* (Rx) por uma câmera mais sofisticada, a Kinect v2.

Em [4], Liu et al. investigam o rastreamento de veículos em condições adversas por meio da fusão de radar mmWave com *bounding boxes* obtidas por detecção de objetos. Esse estudo utiliza equipamentos de radar especializados e dados simulados provenientes dos simuladores *Simulation of Urban Mobility* (SUMO) e *Car Learning to Act* (CARLA), sem medições físicas. Em [5], Sunami et al. concentram-se em eventos de bloqueio dinâmico e utilizam a frequência mais baixa de 5,22 GHz em seus experimentos, enquanto neste trabalho utilizamos 60 GHz, explorando, portanto, feixes consideravelmente mais estreitos, além de fornecer localização independente de eventos de bloqueio dinâmico. Saikko et al. [6] utilizam medições de fase da portadora e Doppler, juntamente com técnicas de filtragem de Kalman e de partículas, para localizar precisamente a posição 3D do Rx em simulações numéricas de cenários avaliados em frequências de até 28 GHz.

Huan et al. [7] combinam *fingerprinting* de CSI em 60 GHz com imagens *Red, Green, and Blue* (RGB), ambos provenientes do conjunto de dados sintético ViWi [8], e apresentam experimentos voltados ao posicionamento de veículos em ambientes externos. Por fim, no trabalho de Mihara et al. [9], o *Received Signal Strength Indicator* (RSSI) é utilizado em conjunto com imagens de câmeras estéreo para fornecer o posicionamento do *User Equipment* (UE) com o objetivo de prever bloqueios, dependendo, portanto, de *hardware* de imageamento sofisticado.

Diferentemente das abordagens anteriores, neste trabalho não utilizamos dados simulados nem equipamentos especializados de radar para localização. Em vez disso, baseamo-nos em dados reais de rádio e imagem, obtidos a partir de equipamentos comerciais, bem como em um procedimento simples de varredura de feixes para realizar a localização. Nesse contexto, as principais contribuições deste trabalho são:

O conteúdo deste capítulo foi desenvolvido pelos pesquisadores *João Borges* e *Aldebaro Klautau*.

1. investigar o uso de *hardware* de baixo custo para a localização do Rx, considerando tanto dados de rádio quanto dados de imagens;

2. realizar essa investigação considerando dados de rádio em mmWave;

3. explorar a contribuição de imagens RGB obtidas no lado da infraestrutura para a tarefa de estimação de posição, isto é, a partir de uma câmera posicionada no *Transmitter* (Tx), em vez de câmeras posicionadas nos Rxs;

4. comparar o método baseado na fusão com imagens RGB com dois níveis distintos de implementação usados como *baselines*: o método que utiliza apenas a direção estimada a partir do feixe de maior potência e o método que utiliza a mediana das estimativas fornecidas por subgrupos de feixes selecionados a partir do Top-K.

2.1 Modelagem do Sistema

Esta seção apresenta a estrutura proposta para a determinação da localização do Rx, a qual combina informações geométricas com medidas de rádio e imagens. A Subseção 2.1.1 descreve a configuração *Multiple Input Single Output* (MISO) considerada e o procedimento de varredura de feixes. A Subseção 2.1.2 apresenta o mapeamento dos feixes para suas respectivas direções angulares. A Subseção 2.1.3 introduz o modelo angular baseado em imagem. A Subseção 2.1.4 detalha a estimação da posição por meio de relações geométricas, enquanto a Subseção 2.1.5 descreve a estratégia baseada em subconjuntos utilizada para melhorar a robustez da metodologia.

2.1.1 Sistema de Comunicação

A configuração considerada consiste em uma *Base Station* (BS) fixa equipada com uma *Uniform Planar Array* (UPA) e um UE móvel com uma única antena, formando um sistema MISO operando na frequência da portadora f_c. A BS emprega uma UPA com M_t elementos de antena, sendo M_y colunas e M_z linhas. O canal entre a BS e o UE, para cada posição do Rx, é representado por um vetor complexo de dimensão $M_t \times 1$.

O *beamforming* é realizado a partir de um *codebook* $\mathcal{W}$, contendo W *codewords*, em que cada *codeword* w direciona o feixe transmitido para uma direção específica. Durante a operação, a BS executa a varredura de feixes sobre todas as *codewords* e, para cada direção de apontamento, o UE mede o RSSI. Os valores correspondentes de RSS, em dBm, são então obtidos por meio do mapeamento linear descrito em [10], dado por

$$\mathrm{RSS} = 0.0651 \cdot \mathrm{RSSI} - 74.3875. \tag{1}$$

2.1.2 Mapeamento das *codewords* para as direções dos lóbulos principais

Cada *codeword* $w \in \mathcal{W}$ gera um padrão de radiação direcional na UPA. Considerando elementos de antena isotrópicos dispostos ao longo dos eixos y e z, o fator de arranjo associado a w é dado por [11]

$$AF(\theta, \phi) = \sum_{m_y=1}^{M_y} \sum_{m_z=1}^{M_z} w_{m_y m_z} e^{j(m_y-1)\Psi_y} e^{j(m_z-1)\Psi_z}, \tag{2}$$

com

$$\Psi_y = kd_y \cos\theta \sin\phi \tag{3}$$

$$\Psi_z = kd_z \sin\theta \tag{4}$$

em que d_y e d_z representam os espaçamentos entre elementos, enquanto θ, ϕ e k denotam, respectivamente, os ângulos de elevação, azimute e o número de onda.

Devido à simetria do padrão da UPA, $AF(\theta, \phi)$ é avaliado sobre uma grade esférica discreta com resolução angular a_s, em que $\theta \in [-\pi/2, \pi/2)$ e $\phi \in [-\pi/2, \pi/2)$. A direção do lóbulo principal de cada *codeword* w é definida como o ponto da grade que maximiza a magnitude do fator de arranjo, dado por

$$(\theta_w, \phi_w) = \underset{\theta_i \in \Theta, \phi_j \in \Phi}{\arg\max} |AF_w(\theta_i, \phi_j)|, \tag{5}$$

em que $\Theta = \theta | \theta = -\frac{\pi}{2} + ka_s, k \in \mathbb{Z} \cap [-\frac{\pi}{2}, \frac{\pi}{2})$, sendo Φ definido de forma análoga para o ângulo de azimute. A repetição desse procedimento para todas as *codewords* fornece o mapeamento direcional do *codebook*, isto é, $(\theta_w^*, \phi_w^*)_{w \in \mathcal{W}}$, o qual associa cada vetor de apontamento ao centro de seu respectivo lóbulo principal. A Figura 1 ilustra a distribuição dos feixes disponíveis na grade angular.

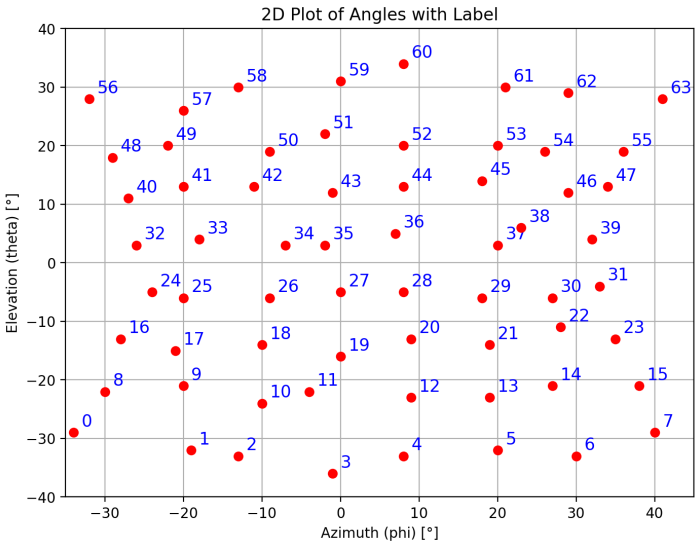

Figura 1: Grade angular do *codebook* adotado, com os índices dos feixes representados no plano azimute e elevação.

2.1.3 Modelagem angular baseada em imagem

Como fonte auxiliar de informação, imagens RGB capturadas a partir do ponto de vista do Tx são incorporadas à estrutura de localização proposta. Para isso, um detector baseado em *You Only Look Once* (YOLO) é utilizado para identificar o Rx em cada imagem, e o centro da caixa delimitadora estimada, denotado por (u, v), é adotado como ponto de referência visual.

Esse ponto é então mapeado para ângulos locais da câmera a partir da resolução da imagem e dos campos de visão horizontal e vertical adotados. A conversão segue o modelo de câmera *pinhole*, no qual as coordenadas do pixel são inicialmente transformadas em coordenadas normalizadas no plano da imagem, associadas a um raio de visualização e, por fim, convertidas em ângulos de elevação e azimute [12]. Além disso, uma correção de *pitch* é aplicada para compensar a inclinação vertical da câmera.

O par angular resultante, $(\theta_{\text{img}}, \phi_{\text{img}})$, é interpretado como um feixe artificial, fornecendo uma observação direcional adicional que pode ser combinada com as estimativas baseadas nos feixes de rádio.

2.1.4 Estimação geométrica da posição

A posição do Rx é estimada geometricamente a partir da localização conhecida do Tx, da altura conhecida do Rx e da informação angular associada a cada feixe selecionado. Seja a posição do Tx denotada por

$$\mathbf{p}_{tx} = \begin{bmatrix} x_{tx} & y_{tx} & z_{tx} \end{bmatrix}^T, \tag{6}$$

e considere que o Rx está localizado sobre um plano horizontal de altura conhecida z_{rx}. Para um dado feixe, sua direção é representada pelo vetor unitário

$$\mathbf{u} = \begin{bmatrix} \cos(\theta)\cos(\phi) & \cos(\theta)\sin(\phi) & \sin(\theta) \end{bmatrix}^T, \tag{7}$$

em que θ e ϕ denotam, respectivamente, os ângulos de elevação e azimute. Esse vetor corresponde à decomposição da direção do feixe em suas componentes horizontal e vertical, conforme ilustrado na Figura 2.

A partir da posição do Tx e seguindo a direção do feixe, o caminho de propagação pode ser descrito pela reta paramétrica

$$\mathbf{p}(t) = \mathbf{p}_{tx} + t\mathbf{u}, \qquad t \in \mathbb{R}. \tag{8}$$

Como o Rx está restrito ao plano $(z = z_{rx})$, a posição estimada é obtida pela interseção entre essa reta e o plano do Rx. Considerando a componente vertical da equação paramétrica, impõe-se

$$z_{tx} + tu_z = z_{rx}, \tag{9}$$

a partir da qual o parâmetro correspondente é obtido como

$$t^\star = \frac{z_{rx} - z_{tx}}{u_z}. \tag{10}$$

Substituindo $(t^\star)$, em (10), na equação da reta em (9), obtém-se a estimativa cartesiana da posição do Rx, dado por

$$\hat{\mathbf{p}}_{rx} = \mathbf{p}_{tx} + t^\star\mathbf{u}. \tag{11}$$

Dessa forma, a localização do Rx é formulada como um problema de interseção entre reta e plano, no qual a direção do feixe define a reta e a altura conhecida do Rx define o plano. Essa interpretação também é útil para identificar feixes geometricamente inconsistentes, como aqueles aproximadamente paralelos ao plano do Rx ou apontando na direção oposta, os quais podem ser descartados para reduzir o impacto de *outliers* e efeitos associados a múltiplos percursos.

2.1.5 Estratégia de combinação por subconjuntos

Após a definição do conjunto Top-K, composto pelos K feixes com maior potência, considera-se a formação de subconjuntos de cardinalidade b, com $b \leq K$. Seja $\mathcal{B}_K$ o conjunto contendo

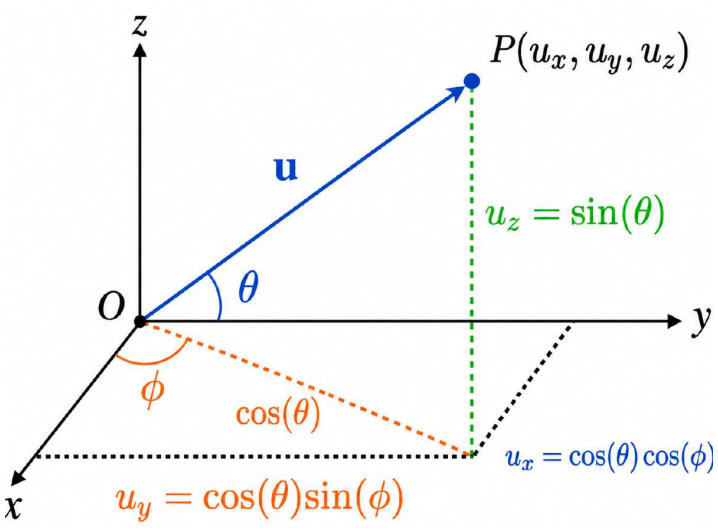

Figura 2: Representação geométrica do vetor unitário de direção associado a um feixe, destacando sua projeção no plano (XY), sua componente vertical e as relações angulares definidas pelo azimute (ϕ) e pela elevação (θ).

esses K feixes. A partir dele, define-se a família de todos os subconjuntos distintos de tamanho b, denotada por $\mathcal{S}$. Assim, o número total de combinações possíveis é dado por

$$N_{K,b} = \binom{K}{b} = \frac{K!}{b!(K-b)!}. \tag{12}$$

Cada subconjunto $S_j \in \mathcal{S}$, com $j \in 1, 2, \ldots, N_{K,b}$, induz uma estimativa parcial da posição do Rx a partir das estimativas individuais associadas aos feixes que o compõem. Seja $\hat{\mathbf{p}}_i$ a estimativa cartesiana produzida pelo i-ésimo feixe pertencente ao subconjunto S_j. A estimativa parcial associada a esse subconjunto é então definida como

$$\hat{\mathbf{p}}_{S_j} = \frac{1}{b} \sum_{i=1}^{b} \hat{\mathbf{p}} * i, \tag{13}$$

em que b representa o número de elementos em cada subconjunto. Assim, cada subconjunto gera uma estimativa parcial, e o conjunto de todas essas estimativas pode ser escrito como

$$\mathcal{P} = \left\{ \hat{\mathbf{p}}_{S_1}, \hat{\mathbf{p}}_{S_2}, \ldots, \hat{\mathbf{p}}_{S_{N_{K,b}}} \right\}. \tag{14}$$

A estimativa final da posição é então obtida por meio de uma operação de *consenso*, baseada na mediana das estimativas médias, aplicada sobre $\mathcal{P}$. Neste trabalho, a mediana é adotada como estatística de agregação devido à sua maior robustez na presença de estimativas parciais discrepantes, as quais podem surgir em decorrência de feixes *outliers* ou trajetórias associadas a múltiplos percursos, sendo dada por

$$\hat{\mathbf{p}}_{\text{final}} = \text{median}\left(\mathcal{P}\right). \tag{15}$$

2.2 Descrição do Cenário e do Conjunto de Dados

Os experimentos foram conduzidos em um ambiente laboratorial interno utilizando uma BS fixa e uma plataforma robótica móvel atuando como UE. A BS foi equipada com uma UPA 6×6, isto é, $M_y = 6$ e $M_z = 6$, com os quatro elementos dos cantos desabilitados. Cada elemento de antena consistia em um dipolo cruzado, e as medições foram coletadas em $f_c = 60.48$ GHz, utilizando um *codebook* cuja cobertura angular é aproximadamente limitada a $[-40°, 40°]$, tanto em elevação quanto em azimute.

A plataforma robótica, equipada com um rádio mmWave com um único elemento de antena ativo e uma câmera RGB, seguiu uma rota predefinida dentro do ambiente. Ao longo dessa trajetória, foram consideradas 58 posições válidas do Rx, cada uma contendo tanto informações de canal quanto a imagem RGB correspondente. Para a etapa de localização, foi utilizada apenas a primeira captura de RSS de cada posição [10]. Nesse sentido, foi adotado um sistema de coordenadas bidimensional no plano da sala, com o canto superior direito definido como a origem $(0, 0)$. Nessa convenção, o eixo x positivo aponta para cima, enquanto o eixo y positivo aponta para a esquerda. As dimensões da sala, a posição da BS, a trajetória e as posições do Rx são apresentadas na Figura 3.

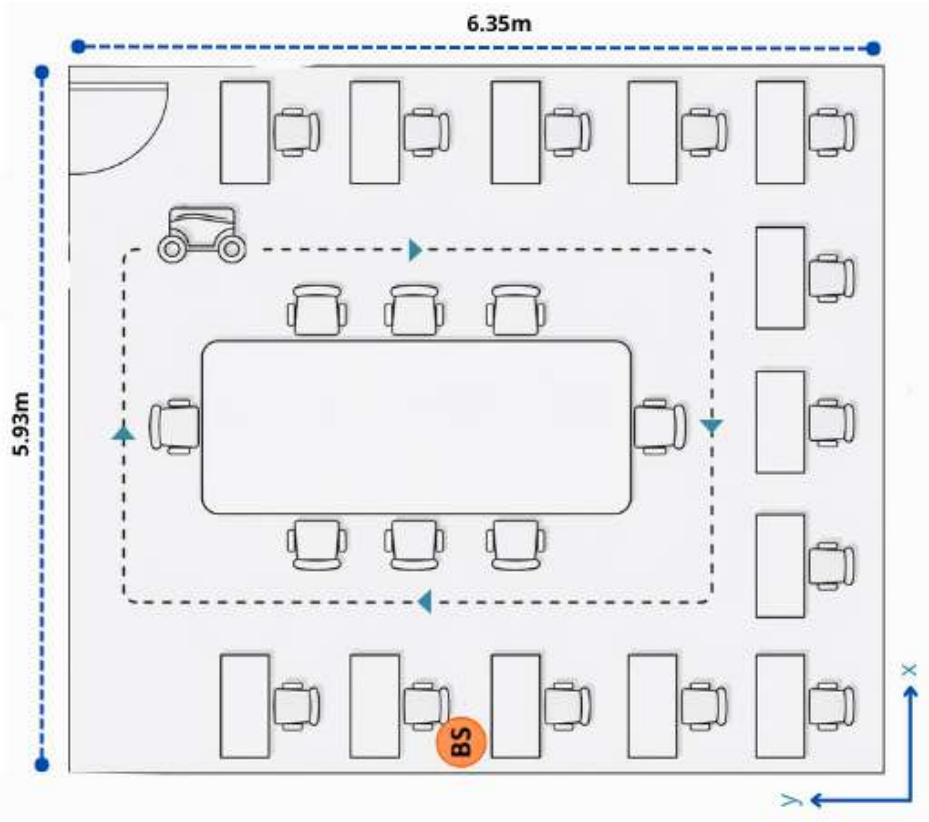

Figura 3: Cenário experimental interno, incluindo as dimensões da sala, o sistema de coordenadas adotado, a posição da BS e a trajetória seguida pela plataforma robótica ao longo das posições do Rx.

As posições analisadas incluem condições *on-grid* e *off-grid* em relação à discretização angular do *codebook*. No primeiro caso, a direção do Rx está alinhada com a grade angular e pode ser melhor representada pelos feixes disponíveis. No segundo caso, a direção do Rx não coincide exatamente com as direções discretizadas, resultando em uma condição de localização *off-grid* mais desafiadora.

2.3 Método Proposto

O método proposto estima a posição cartesiana do Rx por meio da combinação de informações geométricas prévias com medidas de rádio obtidas a partir da varredura de feixes em mmWave. O método assume que a posição do Tx é conhecida, que o Rx está localizado sobre um plano de altura conhecida e que cada índice de feixe pode ser associado a uma direção angular por meio do *codebook* adotado e do modelo de antena. Na versão assistida por imagem, imagens RGB capturadas a partir do ponto de vista do Tx também são utilizadas como uma fonte adicional de informação angular.

O processo inicia-se a partir dos valores de RSSI associados a todos os feixes em uma dada posição do Rx. Os Top-K feixes com maior potência recebida são selecionados sob a hipótese de que, em condições de linha de visada, feixes mais fortes possuem maior probabilidade de apontar na direção real do Rx. Cada feixe selecionado é então mapeado para seu respectivo par angular (θ, ϕ), representando elevação e azimute. Essa associação é obtida a partir do padrão de radiação do feixe, em que a direção do feixe é definida como a localização angular de máxima potência transmitida, considerando uma aproximação com elementos de antena isotrópicos. No caso assistido por imagem, o Rx é detectado na imagem RGB, e o centro da caixa delimitadora estimada é mapeado para um par angular local $(\theta_{\text{img}}, \phi_{\text{img}})$. Essa direção é interpretada como um feixe artificial e adicionada ao conjunto de observações angulares baseadas em rádio.

Uma vez disponíveis as informações angulares, cada feixe de rádio e, quando aplicável, o feixe artificial derivado da imagem são convertidos em vetores unitários de direção em 3D e projetados sobre o plano do Rx, produzindo estimativas cartesianas individuais. Em vez de depender de um único feixe, o método explora todos os subconjuntos de cardinalidade b formados a partir do conjunto Top-K, conforme definido em (12). Na versão assistida por imagem, a estimativa derivada da imagem também é incluída na formação dos subconjuntos. Em seguida, cada subconjunto produz uma estimativa parcial de posição por meio da média aritmética das estimativas cartesianas de seus elementos. Por fim, todas as estimativas parciais são agregadas por meio da mediana calculada coordenada a coordenada, conforme apresentado na Equação (15). Portanto, a estimativa derivada da imagem contribui tanto para as estimativas parciais dos subconjuntos quanto para o resultado final de localização baseado em consenso.

2.4 Resultados e Discussões

Esta seção apresenta os resultados de localização obtidos para as 58 posições do Rx presentes no conjunto de dados. Todas as estratégias avaliadas são baseadas na estimação geométrica por feixe descrita na Subseção 2.3 e diferem quanto à forma como as informações direcionais são selecionadas e agregadas. Três abordagens foram consideradas: a primeira é um *baseline* Top-1, baseado apenas no feixe de maior potência; a segunda é uma estratégia baseada em subconjuntos, utilizando os feixes Top-K com consenso por mediana; e a terceira é uma versão assistida por imagem, na qual a informação angular derivada de imagens RGB é incorporada como uma observação adicional. Salvo indicação em contrário, os experimentos baseados em subconjuntos foram realizados com Top-$K = 5$ e cardinalidade de subconjunto $b = 4$, uma vez que essa configuração apresentou o melhor compromisso entre acurácia e robustez.

2.4.1 Protocolo experimental e métricas de avaliação

O desempenho foi avaliado por meio dos erros nas coordenadas cartesianas e do erro planar de localização, definido como

$$e_{2D} = \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2}, \tag{16}$$

em que (x, y) e $(\hat{x}, \hat{y})$ denotam, respectivamente, as posições real e estimada do Rx. Essas métricas permitem avaliar tanto o comportamento por coordenada quanto a acurácia geral da localização. Além disso, gráficos espaciais de erro foram utilizados para análise qualitativa, enquanto os principais indicadores estatísticos derivados do erro de localização 2D são resumidos na Tabela 1.

Tabela 1: Desempenho de localização 2D para os métodos avaliados

Método	**Média [m]**	**Var. [m^2]**	**DP [m]**
Top-1	1.1394	0.4933	0.7024
Subconjuntos	0.7025	0.1774	0.4212
Subconjuntos+imagem	0.6374	0.1861	0.4314

2.4.2 *Baseline* Top-1

A primeira estratégia avaliada utiliza apenas o feixe de maior potência, de modo que a posição do Rx é estimada diretamente a partir da informação angular associada a uma única observação direcional. Essa configuração é utilizada como *baseline*, pois representa o uso mais direto da formulação geométrica proposta, sem qualquer mecanismo adicional de combinação ou consenso.

A Figura 4 apresenta, em azul, os erros de localização nas componentes x e y para a configuração Top-1. Embora estimativas válidas sejam obtidas para todas as posições, os maiores erros ocorrem principalmente nas posições 1-16 e 50-58 para a componente x, e em torno das posições 16, 44-46 e 54 para a componente y, conforme ilustrado na Figura 5. Essas regiões são mais sensíveis a incompatibilidades angulares, especialmente quando o Rx está mais distante do Tx ou quando sua direção não é bem representada por um único feixe do *codebook*. Como resultado, a discretização angular finita do *codebook* se traduz diretamente em maiores erros cartesianos de localização.

2.4.3 Localização baseada em subconjuntos de feixes

A segunda estratégia explora a diversidade angular dos feixes Top-K. Conforme mostrado na Figura 4, a abordagem baseada em subconjuntos apresenta uma distribuição de erro mais equilibrada do que a abordagem Top-1, reduzindo o erro médio 2D de 1.1394 m para 0.7025, m, isto é, uma melhoria de 38.34%. Esse ganho resulta da combinação de múltiplos feixes fortes, o que reduz a dependência de uma única direção de apontamento e mitiga parcialmente a discretização angular do *codebook*. Além disso, a etapa de consenso baseada na mediana favorece estimativas de subconjuntos geometricamente consistentes, reduzindo o efeito de soluções ruidosas ou discrepantes.

Também é importante observar que o mapeamento de coordenadas esféricas para cartesianas induz uma distribuição espacial não uniforme das direções estimadas. Como resultado, a grade

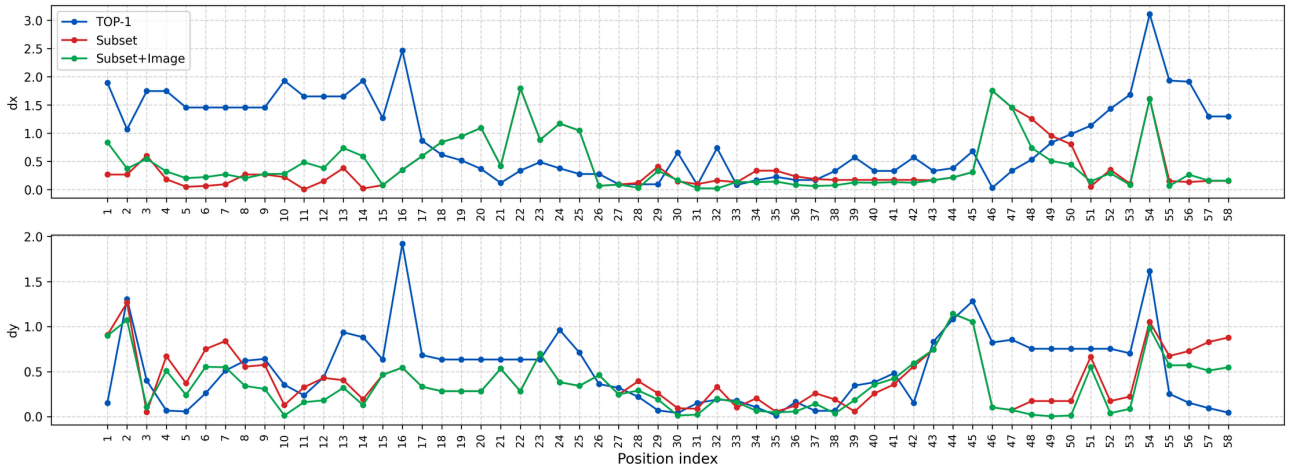

Figura 4: Comparação dos erros de localização cartesianos nas componentes x e y para os métodos Top-1, baseado em subconjuntos e baseado em subconjuntos assistido por imagem.

angular torna-se efetivamente mais densa próxima ao Tx e mais esparsa em maiores distâncias. Isso ajuda a explicar por que a estratégia baseada em subconjuntos é particularmente benéfica em posições mais distantes do Rx, nas quais a estimação por feixe único é mais fortemente afetada pelos efeitos da discretização. Como cada estimativa de subconjunto é obtida pela média das projeções associadas aos feixes, a solução final pode apresentar um leve viés em direção à região central da área mapeada. Além disso, a formulação assume comportamento dominante em *Line-of-Sight* (LoS); sob condições de *Non-Line-of-Sight* (NLoS), trajetórias refletidas podem deslocar as direções aparentes dos feixes e, consequentemente, enviesar a posição estimada.

2.4.4 Localização por subconjuntos assistida por imagem

A última estratégia avaliada estende o método de localização baseado em subconjuntos de feixes por meio da incorporação de informação angular derivada da imagem. Mais especificamente, a localização do rádio na imagem RGB é utilizada para gerar uma observação direcional adicional a partir do ponto de vista do Tx. Essa contribuição visual é então combinada com as estimativas baseadas em feixes, a fim de restringir ainda mais a posição do Rx, especialmente em condições *off-grid*.

A Figura 4 mostra que a estratégia assistida por imagem resulta em um perfil de erro mais concentrado do que a abordagem baseada apenas em subconjuntos de feixes. Essa melhoria também é confirmada pelos resultados quantitativos: o erro médio 2D é reduzido para 0.6374 m, correspondendo a uma melhoria de aproximadamente 44.06% em relação ao *baseline* Top-1. O benefício da contribuição derivada da imagem não é uniforme ao longo da trajetória. Ele é mais evidente nas posições 4-14 e 48-51, nas quais a solução assistida por imagem reduz os erros nas componentes x ou y de forma mais consistente do que o método baseado apenas em subconjuntos de feixes. Essas regiões correspondem a porções externas da área coberta, nas quais o Rx está mais distante do Tx ou mais próximo dos limites da cobertura angular. Nesses casos, a observação visual atua como um feixe artificial complementar, refinando a inferência direcional e tornando a localização final menos sensível à discretização do *codebook* e aos efeitos da resolução angular.

Esse ganho foi obtido mesmo na presença de limitações práticas que afetam a modelagem angular baseada em imagem, incluindo distorções mais pronunciadas longe do centro da ima-

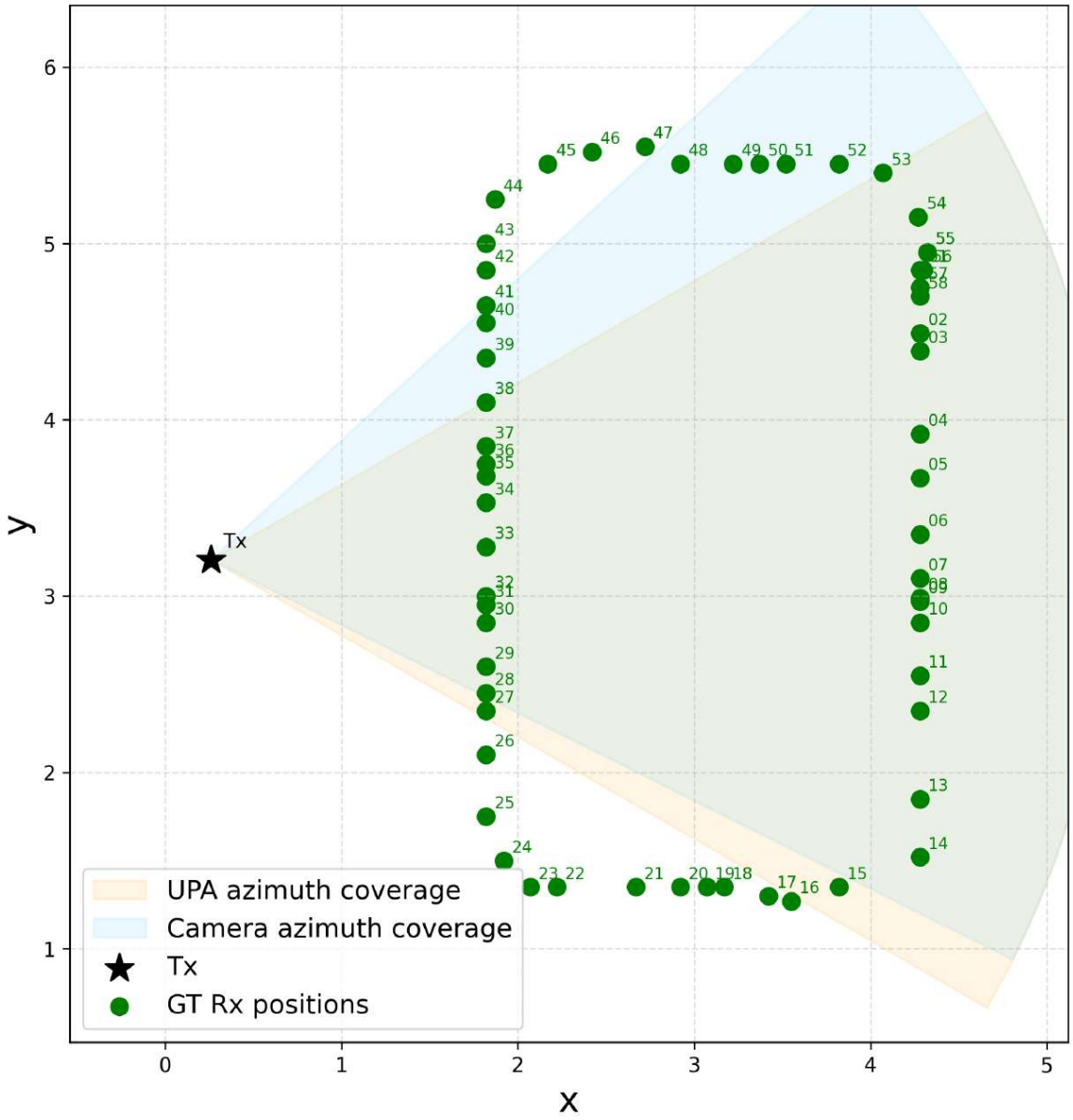

Figura 5: Posições reais do Rx no cenário, juntamente com a localização do Tx e as regiões de cobertura azimutal da UPA e da câmera RGB.

gem, a adoção de um modelo simplificado de câmera pinhole e possíveis inclinações verticais e horizontais da câmera, fatores que introduzem incertezas no processo de mapeamento de pixel para ângulo. Ainda assim, a contribuição visual permaneceu suficientemente informativa para melhorar a consistência geométrica da estimativa final e complementar a estratégia baseada em subconjuntos de feixes.

2.5 Conclusões

Nesta frente de pesquisa, utilizamos uma configuração mmWave de baixo custo em conjunto com uma câmera RGB para realizar a estimação da posição do Rx em ambiente interno por meio da fusão multimodal de dados de rádio e imagem. A abordagem proposta explora informações de rádio facilmente acessíveis, obtidas a partir de medidas de RSS coletadas durante a varredura de feixes realizada pelo Tx sobre um *codebook* composto por 64 feixes direcionais.

As orientações desses feixes, juntamente com informações prévias sobre a posição do Tx e a altura do Rx, permitem obter uma estimativa inicial da localização do Rx para cada feixe. Essa estimativa é posteriormente refinada por meio da formação de subconjuntos de feixes, nos quais são calculadas estimativas parciais a partir da média das coordenadas cartesianas associadas a cada subconjunto. Em seguida, a mediana dessas estimativas parciais é utilizada para produzir uma estimativa final mais robusta.

Por fim, foi empregada uma abordagem baseada em detecção de objetos para refinar ainda mais a estimação, explorando caixas delimitadoras geradas pelo YOLO sobre o Rx detectado em imagens RGB capturadas a partir do ponto de vista do Tx. Essa contribuição visual, incorporada à abordagem baseada em subconjuntos, forneceu a estimativa final, alcançando erro médio de localização inferior a 1 metro para o cenário estudado.

Como trabalhos futuros, pretende-se empregar um modelo de antena de dipolo cruzado mais preciso para aprimorar a caracterização dos feixes, utilizar modelos de câmera mais refinados para reduzir erros no mapeamento de pixels para ângulos e desenvolver estratégias adaptativas para escolher parâmetros como Top-K, cardinalidade dos subconjuntos e uso da informação visual de acordo com a confiança das estimativas individuais baseadas em rádio e imagem. Além disso, cenários mais desafiadores, com múltiplos Rxs, obstáculos e condições de NLoS, também podem ser considerados.

3 Localização 3D de Equipamentos de Usuário via Fusão Multimodal Rádio-Visual em Redes mmWave

A localização precisa de UE em ambientes *indoor* é um requisito importante para serviços de comunicação, onde a ausência de sinal de *Global Navigation Satellite System* (GNSS) torna necessário o desenvolvimento de métodos alternativos [13]. Sistemas de rádio mmWave com arquitetura *Open Radio Access Network* (O-RAN), como os empregados na 5G-*New Radio* (5G-NR), disponibilizam métricas detalhadas de canal pela interface E2, entre eles *Reference Signal Received Power* (RSRP), *Reference Signal Received Quality* (RSRQ), *Signal-to-Interference-plus-Noise Ratio* (SINR), RSSI e *Channel Quality Indicator* (CQI), que carregam informação implícita sobre a geometria da propagação e, portanto, sobre a posição do UE. Paralelamente, câmeras RGB-*Depth* (RGB-D) integradas às estações base fornecem contexto visual rico do ambiente, complementando a resolução espacial limitada das medidas de rádio. A fusão dessas duas fontes de informação constitui uma direção promissora para localização de alta precisão, alinhada à visão de *Integrated Sensing and Communications* (ISAC) em redes de próxima geração [14].

3.1 Metodologia

A integração de câmeras RGB-D com sinais de rádio para estimação de posição 3D de UE é explorada por meio da proposta de uma arquitetura de fusão rádio-visual treinável de ponta a ponta (*end-to-end*). São avaliadas três estratégias de fusão multimodal, incluindo mecanismos nos quais as *features* de rádio modulam adaptativamente as representações visuais. Além disso, são empregados um mapa de disparidade (RGB-D) e um prior geométrico obtido por meio de um detector de objetos, com o objetivo de enriquecer a representação visual com informações estruturais e semânticas. Por fim, é proposto um protocolo de avaliação robusto para avaliar a capacidade de generalização a trajetórias desconhecidas. Os modelos são treinados e avaliados no conjunto de dados multimodal disponibilizado no CONVERGE *Challenge Task 2* (CCT2) [14].

3.1.1 Conjunto de dados

O conjunto de dados do CCT2 [14] é coletado em um ambiente *indoor* com uma estação base 5G mmWave (LiteOn FR2, pilha OAI) montada sobre um braço robótico UR-10e e um UE Quectel FR2 embarcado num robô Turtlebot 4. A Figura 6 ilustra amostras do ambiente experimental capturadas pela câmera da estação base.

O conjunto de dados contém três trajetórias (exp6, exp7, exp8) gravadas de forma sincronizada via *Picture Transfer Protocol* (PTP) com: (i) câmera *stereo* Nerian Ruby RGB-D acoplada à estação base, fornecendo quadros RGB e mapas de disparidade a 7 fps (142 ms por quadro); (ii) métricas de interface E2 do protocolo O-RAN (como RSRP, RSRQ, SINR, RSSI e CQI) amostradas a cada 50 ms; e (iii) posição 3D de referência do UE obtida pelo sistema de captura de movimento Qualisys com precisão sub-milimétrica (*ground truth* em mm). Nas trajetórias exp6 e exp7, o UE permanece dentro do *Field of View* (FoV) da câmera, e na trajetória exp8 o robô inicia e termina fora do FoV. A Figura 7 apresenta as trajetórias exp6, exp7 e exp8, com as posições 3D do UE em relação à estação base.

O conteúdo deste capítulo foi desenvolvido pelos pesquisadores *Wesley L. Passos* e *José F. Rezende*.

Figura 6: Amostras de *frames* RGB da trajetória exp6, capturadas pela câmera Nerian Ruby RGB-D acoplada à estação base. O UE (Turtlebot 4) aparece em diferentes posições ao longo da trajetória.

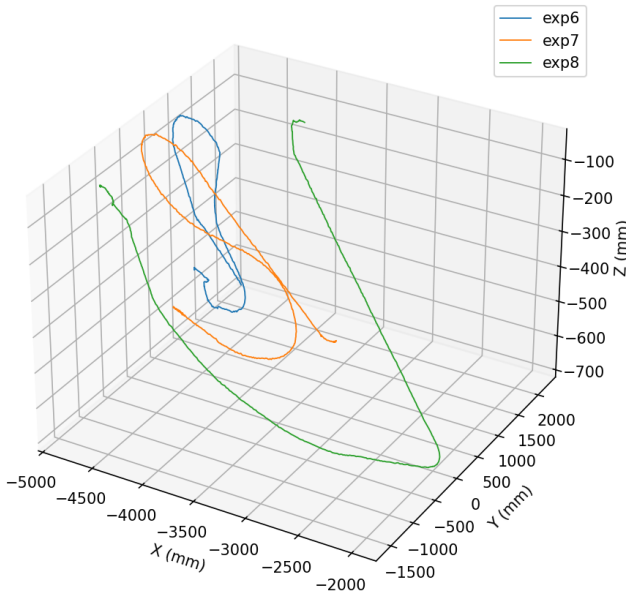

Figura 7: Trajetórias exp6, exp7 e exp8, com as posições 3D do UE em relação à estação base. As trajetórias exp6 e exp7 permanecem dentro do campo de visão da câmera, enquanto a exp8 inicia e termina fora do FoV.

3.1.2 Arquitetura proposta

A Figura 8 ilustra a arquitetura proposta, que é composta por três módulos: extração de *features*; fusão de *features*; e regressão.

A extração de *features* é dividida em 3 ramos: (i) visuais, (ii) rádio (E2) e (iii) geométricas. Para extração de *features* visuais é utilizado um *backbone* ResNet-18 [15], com pesos pré-treinados na base ImageNet. Esse *backbone* processa o quadro de entrada (RGB, RGB-D ou disparidade), com a primeira camada convolucional adaptada ao número de canais da entrada. No caso do canal de disparidade, a camada é inicializada como a média dos pesos RGB para aproveitar as representações pré-treinadas. A camada de classificação é substituída por uma identidade, produzindo um vetor de *features* visuais $\mathbf{v} \in \mathbb{R}^{512}$.

As *features* de rádio são extraídas das seguintes medidas da interface E2: RSRP, RSRQ, SINR, RSSI, e CQI, as quais são normalizadas via z-*score* e passam por uma rede *Multilayer Perceptron* (MLP) de dois estágios (5→64→64, ReLU, e *dropout* com probabilidade p_d=0,1),

produzindo uma representação densa $\mathbf{r} \in \mathbb{R}^{64}$.

Para injetar um prior geométrico explícito, é anexado um descritor de detecção compacto extraído de um modelo YOLOv8 [16] pré-treinado que detecta a posição do robô na imagem. O resultado de cada quadro é compactado em um vetor $\mathbf{b} = [x_c; y_c; w; h; s; d] \in \mathbb{R}^6$, onde x_c e y_c são as coordenadas do centro da *bounding box* na imagem, w e h a largura e altura da *bounding box*, $s \in [0, 1]$ é o nível de confiança da detecção pelo modelo YOLO e d um indicador binário de detecção válida. Quando não há detecção, $(x_c; y_c)$ assume o valor padrão (0,5; 0,5) e o indicador $d = 0$.

Dadas as representações $\mathbf{v} \in \mathbb{R}^{512}$, $\mathbf{r} \in \mathbb{R}^{64}$ e $\mathbf{b} \in \mathbb{R}^6$, três estratégias de fusão são avaliadas: concatenação simples (concat.), *Gated Multimodal Unit* (GMU) [17] e *Feature-wise Linear Modulation* (FiLM) [18].

Após a fusão das representações por algum dos mecanismos descritos, uma rede MLP com duas camadas ocultas (512 e 256 unidades, ReLU, *dropout* p=0,1) realiza a regressão para estimar as coordenadas 3D (x, y, z).

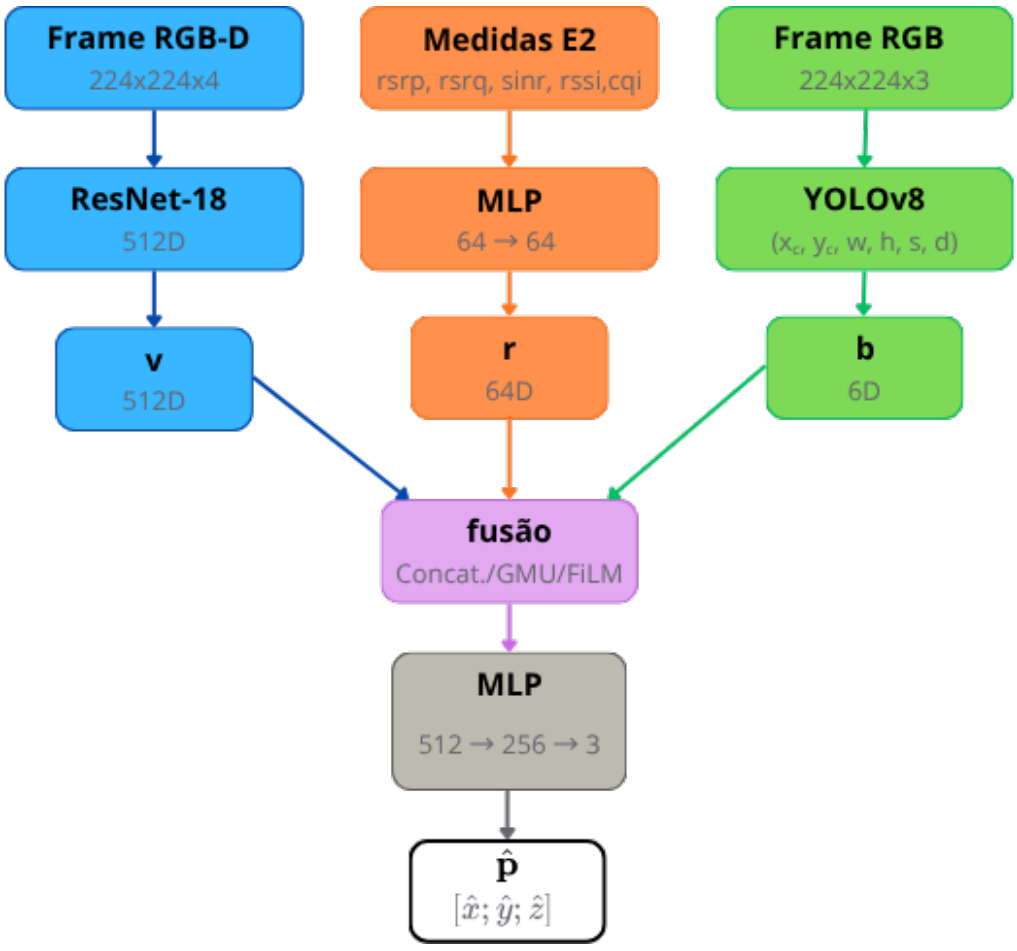

Figura 8: Arquitetura proposta. Três ramos extraem representações complementares: (i) visual: *backbone* ResNet-18 adaptado para entrada RGB-D produz $\mathbf{v} \in \mathbb{R}^{512}$; (ii) rádio: MLP aplicada às cinco medidas E2 produz $\mathbf{r} \in \mathbb{R}^{64}$; e (iii) geométrico: detecção do modelo YOLOv8 compactado em $\mathbf{b} \in \mathbb{R}^6$ (centro, dimensões, confiança e validade da *bounding box*). As representações são fundidas por um dos mecanismos avaliados (Concat., GMU ou FiLM) e regredidas para a posição 3D $\hat{\mathbf{p}} = [\hat{x}; \hat{y}; \hat{z}]$ em mm.

3.2 Resultados e Discussões

O erro de localização 3D é definido como a distância euclidiana 3D entre as posições previstas e as medidas. As métricas de avaliação são o erro absoluto médio, *Mean Absolute Error* (MAE), e a raiz do erro quadrático médio, *Root Mean Square Error* (RMSE), calculadas respectivamente como

$$\text{MAE}_{\text{3D}} = \frac{1}{N}\sum_{i=1}^{N}||\hat{\mathbf{p}}_i - \mathbf{p}_i|| \quad \text{e} \tag{17}$$

$$\text{RMSE}_{\text{3D}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}||\hat{\mathbf{p}}_i - \mathbf{p}_i||^2}, \tag{18}$$

onde $\hat{\mathbf{p}} = [\hat{x}; \hat{y}; \hat{z}]$ e $\mathbf{p} = [x; y; z]$ são, respectivamente, a predição e a anotação da posição 3D do UE.

Um protocolo de avaliação inter-trajetória, *Leave-One-Trajectory-Out* (LOTO), é empregado. Neste protocolo, um modelo é treinado com os dados de duas trajetórias e avaliado nos dados da trajetória restante (*holdout* 3-*fold*). Dessa forma, é possível estimar a generalização dos modelos para trajetórias não conhecidas durante o treinamento.

Na Tabela 2, são apresentados os resultados do protocolo de avaliação inter-trajetória (LOTO). O modelo 'FiLM RGBD+E2+YOLO' obteve o menor MAE_{3D} médio, com desempenho consistente em exp6 (172,7 mm) e exp7 (169,4 mm). A modulação FiLM [18], que condiciona as ativações visuais ao vetor de rádio por escalonamento e translação afim, favorece a generalização ao adaptar a representação visual à condição de canal observada.

O 'GMU RGBD+E2+YOLO' apresentou o menor MAE em exp6 (166,5 mm), porém alta variância entre *folds* (372,2 mm em exp7), com MAE médio de 515 mm. Esse mecanismo não generaliza para trajetórias desconhecidas.

A variante 'concat. Disparidade+E2+YOLO' alcançou 922,6 mm de MAE_{3D} em exp8, reduzindo o erro em ≈118 mm frente ao FiLM e apresentando o menor desvio padrão médio (325 mm vs. ≈400 mm dos demais). A disparidade codifica a estrutura geométrica do ambiente ao redor da última posição detectada do robô, oferecendo pistas de localização residuais mesmo quando a aparência visual do objeto está ausente.

Especificamente na trajetória exp8, o robô sai do campo visual da câmera em 14,7% dos *frames*. Todos os métodos predominantemente visuais colapsam, apresentando erros que superam em mais de uma ordem de grandeza os obtidos em exp6 e exp7. A degradação na trajetória exp8 reflete uma limitação estrutural inerente ao conjunto de dados. Posições fora do FoV correspondem a pontos de operação fora da distribuição de treino, independentemente da trajetória de *holdout*. Uma direção promissora é a fusão adaptativa de *features*, de modo que, quando a confiança de detecção do modelo YOLO estiver próxima de zero, o sistema poderia atribuir maior peso à modalidade de rádio, que permanece sempre disponível.

3.3 Conclusões e Direções Futuras

Nesta seção foi apresentada a proposta de uma arquitetura de fusão rádio-visual treinável de ponta a ponta para estimação de posição 3D de UE, avaliada no conjunto de dados do CONVERGE *Challenge Task* 2. A fusão FiLM com *backbone* ResNet-18 e prior geométrico de detecção via YOLOv8 obteve MAE_{3D} médio de 461 mm no protocolo LOTO. O mecanismo de fusão FiLM mostrou maior consistência inter-trajetória que o GMU, e o congelamento parcial

Tabela 2: Resultados de MAE_{3D} e $RMSE_{3D}$ por *fold holdout* e média ± desvio padrão, inter-trajetória (LOTO). Valores em mm. Linhas em itálico: referências sem ajuste fino end-to-end. Demais: ResNet-18 treinável. †Em *exp8* o robô sai do FoV em 14,7% dos *frames*.

Método	*exp6*		*exp7*		*exp8*†		Média ± std	
	MAE_{3D}	$RMSE_{3D}$	MAE_{3D}	$RMSE_{3D}$	MAE_{3D}	$RMSE_{3D}$	MAE_{3D}	$RMSE_{3D}$
Concat. RGBD+E2	178,9	216,4	179,4	195,2	1042,8	1163,4	467±407	525±452
Concat. RGBD+E2+YOLO	221,5	240,4	154,0	171,7	1035,1	1153,6	470±400	522±448
Concat. RGB+E2+YOLO	204,3	233,9	194,5	224,0	1053,2	1223,2	484±403	560±469
GMU RGBD+E2+YOLO	**166,5**	**184,9**	372,2	380,7	1006,0	1141,6	515±357	569±413
FiLM RGBD+E2+YOLO	172,7	194,1	**169,4**	**185,4**	1040,7	1183,8	**461±410**	**521±469**
Concat. Disparidade+E2+YOLO	237,0	266,9	230,2	279,4	**922,6**	**1002,3**	463±325	516±344
Apenas E2 (rádio)	759,5	867,1	956,7	1019,1	3581,8	3861,1	1766±1287	1916±1377

do *backbone* mostrou-se a melhor estratégia para o conjunto de dados reduzido. A principal limitação é a degradação quando o UE sai do campo de visão, cenário em que a variante com disparidade mostrou-se a mais robusta. Como trabalhos futuros, destacam-se a fusão adaptativa de modalidades, a agregação de informação temporal entre amostras consecutivas e a avaliação com *Sounding Reference Signal* (SRS).

4 *Neural Architecture Search* Aplicada à Seleção Top-k de Feixes no Conjunto de Dados Raymobtime S008 e S009

Esta seção apresenta uma avaliação preliminar de técnicas de NAS aplicadas à seleção Top-k de feixes em enlaces veiculares operando em mmWave. O estudo foi conduzido no contexto de redes *Beyond*-5G (B5G) e 6G, nas quais a elevada diretividade dos enlaces impõe desafios importantes para o alinhamento entre transmissor e receptor. O objetivo principal desta etapa é comparar diferentes famílias de modelos neurais sob um mesmo protocolo experimental, utilizando dados sincronizados de *Global Positioning System* (GPS) e LiDAR dos cenários Raymobtime S008 e S009. Neste relatório, são apresentados apenas os resultados comparativos de modelos, mantendo-se a descrição completa da contribuição e das análises adicionais para uma etapa posterior.

4.1 Contextualização

A seleção de feixes em comunicações mmWave busca identificar, com baixa latência, os pares de feixes Tx–Rx mais adequados para estabelecer o enlace de comunicação. A abordagem tradicional baseada em varredura exaustiva avalia todos os pares possíveis de feixes, o que pode resultar em elevada sobrecarga de sinalização e aumento do tempo de descoberta do feixe. Esse problema torna-se ainda mais crítico em cenários veiculares, nos quais a mobilidade, os bloqueios e as rápidas variações do canal podem degradar a confiabilidade da comunicação.

Nesse contexto, informações sensoriais externas à interface de rádio, como coordenadas de GPS, imagens e nuvens de pontos LiDAR, vêm sendo exploradas para reduzir a dependência da varredura exaustiva. Trabalhos recentes demonstram que modelos de aprendizado de máquina podem utilizar tais informações para estimar os melhores feixes candidatos antes da etapa de sondagem do canal, reduzindo a busca a um subconjunto Top-k de alternativas promissoras [19, 20, 21, 22]. No entanto, a escolha da arquitetura neural ainda representa uma decisão sensível, pois ganhos modestos de acurácia podem não compensar aumentos expressivos de complexidade computacional, latência de inferência ou custo de implantação.

4.2 Metodologia

A metodologia adotada considera os subconjuntos Raymobtime S008 e S009, que representam cenários urbanos veiculares com comunicação entre infraestrutura e usuários móveis. O conjunto S008 é utilizado para treinamento e validação dos modelos, enquanto o conjunto S009 é reservado para teste. O S008 contém 11.194 amostras, das quais 6.482 são classificadas como LoS e 4.712 como NLoS. Por sua vez, o S009 contém 9.638 amostras, sendo 1.473 LoS e 8.165 NLoS. Essa diferença torna o conjunto S009 particularmente desafiador, pois a maior proporção de enlaces NLoS aumenta a ambiguidade da tarefa de seleção de feixes.

A etapa de pré-processamento utiliza uma representação conjunta baseada em GPS e LiDAR. Para cada amostra, a nuvem de pontos LiDAR é limitada a um alcance máximo de 100 m e quantizada em um histograma tridimensional de ocupação com dimensão $(20, 200, 10)$. Em seguida, essa representação é agregada em quatro indicadores semânticos por célula da grade, associados à estação rádio-base, veículo, obstáculo e espaço livre, resultando em um

O conteúdo deste capítulo foi desenvolvido pelos pesquisadores *Matheus F. Silva e Felipe A. P. de Figueiredo.*

tensor de dimensão $(20, 200, 4)$. Para as famílias baseadas em sequências, essa grade é reorganizada em 4000 *tokens*, e as coordenadas GPS, após padronização com base no conjunto de treinamento, são replicadas e concatenadas a cada *token*, formando uma entrada conjunta de dimensão $(4000, 6)$. Para a família 2D *Convolutional Neural Network* (CNN), a estrutura espacial da grade é preservada, e as coordenadas GPS são incorporadas como canais adicionais.

A busca de arquiteturas foi realizada com a biblioteca Optuna, empregando o algoritmo *Tree-structured Parzen Estimator* (TPE) para otimização dos hiperparâmetros [23]. Foram avaliadas famílias de modelos baseadas em 1D CNN, 2D CNN, modelos do tipo *transformer*, *Graph Neural Network* (GNN) e arquiteturas híbridas 1D CNN+GNN. Todas as famílias foram comparadas sob o mesmo particionamento de treino e validação em S008, o mesmo protocolo de teste em S009 e a mesma tarefa de classificação com 256 classes, correspondentes aos pares de feixes disponíveis no *codebook*. Também foi avaliada uma versão destilada cuja busca de hiperparâmetros foi conduzida utilizando o algoritmo *Multi-Objective Tree-structured Parzen Estimator* (MOTPE), com o objetivo de reduzir o custo computacional mantendo desempenho próximo ao dos modelos de maior capacidade.

4.3 Comparação dos Modelos

A Tabela 3 apresenta a comparação dos principais modelos avaliados no conjunto S009. As métricas consideradas incluem o número de parâmetros, o custo computacional em *Floating Point Operations* (FLOPs), a latência média de inferência e as acurácias Top-1 e Top-10. Os valores de latência foram obtidos com lote unitário em uma *Graphics Processing Unit* (GPU) NVIDIA RTX 4090, após passagens iniciais de aquecimento, utilizando a média de execuções sincronizadas.

Tabela 3: Comparação dos modelos avaliados no conjunto Raymobtime S009.

Modelo	**Parâmetros**	**FLOPs**	$t_{\mathbf{inf}}$	**Top-1**	**Top-10**
1D CNN conjunta	230k	321,9M	2,30 ms	61,16%	92,08%
2D CNN preservando grade	4,87M	16,28G	28,99 ms	57,62%	92,33%
Transformer	705k	536,9M	2,16 ms	57,12%	90,07%
GNN	1,47M	60,94M	35,27 ms	60,82%	92,14%
Híbrido 1D CNN+GNN	15,91M	76,72M	45,94 ms	59,66%	90,52%
Ensemble por empilhamento	–	–	3,60 s	62,59%	93,68%
1D CNN destilada com TPE	230k	321,9M	2,30 ms	61,59%	93,43%
1D CNN destilada com MOTPE	157k	40,7M	0,49 ms	61,23%	92,30%

Observa-se que a 1D CNN conjunta apresenta o melhor compromisso inicial entre acurácia e custo entre os modelos isolados avaliados. Essa arquitetura alcança 61,16% de acurácia Top-1 e 92,08% de acurácia Top-10, com 230 mil parâmetros e latência média de 2,30 ms. Embora a 2D CNN que preserva a grade alcance desempenho Top-10 ligeiramente superior, seu custo computacional é substancialmente maior, atingindo 16,28G FLOPs e 28,99 ms de latência. Portanto, sua adoção pode ser menos atrativa em cenários de borda com restrições de tempo de resposta.

Os modelos baseados em *transformer*, GNN e arquitetura híbrida também apresentam desempenhos competitivos, mas não superam a relação entre acurácia, latência e complexidade

obtida pela família 1D CNN. A GNN, por exemplo, atinge 60,82% de Top-1 e 92,14% de Top-10, porém apresenta latência de 35,27 ms. Já o modelo híbrido 1D CNN+GNN possui maior número de parâmetros e maior latência, sem apresentar ganho proporcional de acurácia. Esses resultados indicam que arquiteturas mais complexas não necessariamente oferecem melhor desempenho prático quando aplicadas à representação GPS–LiDAR considerada.

O *ensemble* por empilhamento apresenta o maior desempenho absoluto, alcançando 62,59% de Top-1 e 93,68% de Top-10. Entretanto, sua latência média de 3,60 s torna essa alternativa menos adequada para aplicações em tempo real. Por outro lado, a 1D CNN destilada com MOTPE reduz o custo computacional para 40,7M FLOPs e a latência para 0,49 ms, mantendo acurácia Top-1 de 61,23% e Top-10 de 92,30%. Assim, a etapa de destilação indica um caminho promissor para obter modelos mais leves, preservando desempenho próximo ao das arquiteturas de maior custo.

4.4 Trabalhos Futuros

Como próximos passos, pretende-se ampliar a análise para incluir métricas de desempenho no nível de comunicação, tais como razão de *throughput* retido após a sondagem do subconjunto Top-k. Também serão avaliados cenários com erros de sincronização entre sensores, imprecisões nas coordenadas de GPS, ruído nas nuvens de pontos LiDAR e variações na distribuição de tráfego. Essas análises são importantes para verificar a robustez dos modelos em condições mais próximas de implantação real.

5 Predição de Feixes com LLM Multimodal

O gerenciamento eficiente de feixes (*beam management*) em sistemas *Multiple Input Multiple Output* (MIMO) massivo operando em frequências de mmWave é um desafio central para as redes 5G e 6G. Métodos convencionais baseados em varredura exaustiva do espaço de feixes impõem elevado *overhead* de sinalização, tornando-se inviáveis em cenários de alta mobilidade.

Nesse contexto, abordagens baseadas em aprendizado de máquina têm sido propostas para realizar a predição proativa de feixes a partir de dados sensoriais contextuais, como imagens de câmera RGB, nuvens de pontos LiDAR, eliminando a necessidade de conhecimento explícito do ângulo de partida, *Angle of Departure* (AoD). Em particular, o trabalho de Mao *et al.* [24] propõe um *framework* inovador baseado em modelos de linguagem de grande porte, os LLMs, para predição multimodal de feixes, atingindo 80,8% de acurácia Top-1.

Este capítulo apresenta os resultados preliminares da reprodução do *framework* descrito neste trabalho [24], utilizando o *dataset Multimodal-Wireless* [25], com foco na implementação e validação da arquitetura proposta, na integração do pré-processador PointPillars [26] pré-treinado, e na análise dos resultados de treinamento obtidos.

5.1 *Framework* de Predição de Feixes Baseado em LLM Multimodal

5.1.1 Trabalhos relacionados

Mao *et al.* [24] propõem um *framework* que integra três modalidades de entrada, índices históricos de feixes, nuvens de pontos LiDAR e imagens de câmera RGB, com um *backbone* LLM pré-treinado (*Generative Pre-trained Transformer* (GPT)-2 Large) congelado, utilizando a técnica de *reprogramming* para traduzir representações multimodais ao espaço semântico do modelo de linguagem.

A arquitetura distingue-se de abordagens convencionais de *fine-tuning* por manter os pesos do LLM fixos durante o treinamento, atualizando apenas módulos codificadores leves (∼10M parâmetros de um total de ∼777M). Esse *design* permite explorar o conhecimento temporal pré-treinado do GPT-2 para capturar padrões de mobilidade veicular, sem incorrer no custo computacional de ajuste fino de um modelo de grande porte.

Adicionalmente, o trabalho introduz dois mecanismos originais:

- *Beam-Guided Attention Masking* (BGAM): injeta viés indutivo físico no codificador LiDAR, mascarando regiões do mapa *Bird's Eye View* (BEV) irrelevantes para o setor angular do feixe atual, melhorando a desambiguação espacial em cenários com múltiplos veículos;
- Alinhamento temporal de alta frequência: ao invés de subamostragem, os dados sensoriais de menor frequência (10 Hz) são replicados para correspondência com a sequência de índices de feixes de 100 Hz, preservando a resolução temporal.

5.1.2 Arquitetura implementada

A Figura 9 apresenta a estrutura geral do *framework* reproduzido. A implementação completa foi realizada em PyTorch, seguindo fielmente a descrição arquitetural do artigo original.

O conteúdo deste capítulo foi desenvolvido pelos pesquisadores *Alcides Tomás* e *Francine Cassia de Oliveira*.

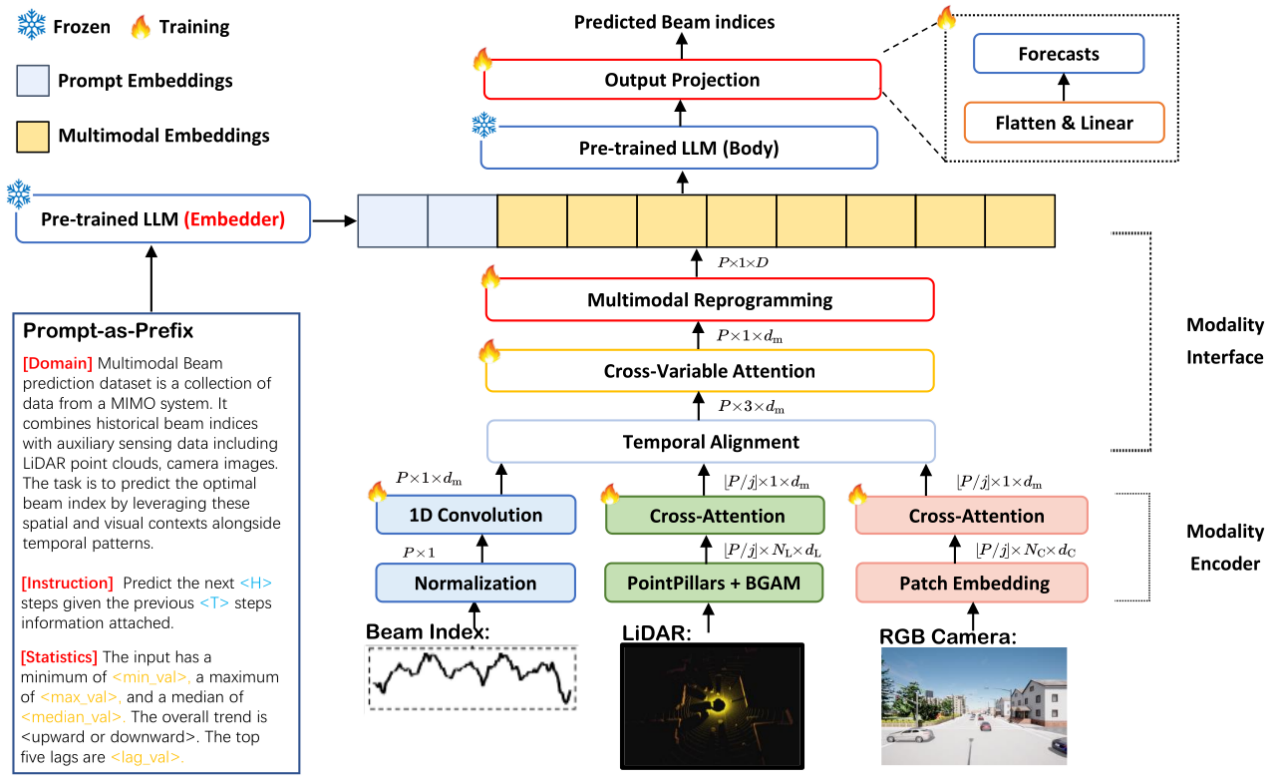

Figura 9: Estrutura geral do *framework* de predição de feixes baseado em LLM multimodal.

Os componentes principais implementados, conforme descrito em [24], são:

- ***BeamIndexEncoder***: Responsável por codificar o histórico de índices de feixes observados nos instantes anteriores, transformando informações discretas de *beamforming* em representações vetoriais adequadas para processamento pelo modelo multimodal.
- ***FrozenPointPillars***: Rede baseada no detector PointPillars pré-treinado e congelado durante o treinamento, utilizada para extrair características espaciais dos dados LiDAR sem atualizar seus pesos.
- **BGAM**: Mecanismo de atenção utilizado para modelar a correlação entre diferentes modalidades de entrada, permitindo destacar informações mais relevantes para a tarefa de predição de feixes.
- ***CameraEncoder***: Extrator de características visuais responsável por converter sequências de imagens capturadas pelas câmeras em representações de alta dimensão que descrevem o contexto do ambiente veicular.
- ***ModalityReprogramming***: Módulo que projeta as características extraídas das diferentes modalidades para o espaço de entrada do modelo de linguagem, permitindo sua integração ao LLM sem a necessidade de modificar sua arquitetura interna.
- ***Prompt* estruturado**: Estrutura textual utilizada para organizar e fornecer ao LLM as informações provenientes das modalidades disponíveis, incluindo histórico de feixes, características visuais e dados LiDAR, permitindo que o modelo realize a predição dos feixes futuros.

5.2 Resultados Preliminares

5.2.1 Progressão do treinamento

A Tabela 4 compara a acurácia Top-1 obtida na reprodução do *framework* com os valores reportados nos trabalhos de referência [24, 27]. Os resultados são apresentados separadamente para os conjuntos de validação e de teste. Os trabalhos de referência reportam apenas os resultados no conjunto de teste, por este refletir a capacidade de generalização do modelo em cenários fora da distribuição de treinamento. Conforme descrito em [24], o conjunto de dados é particionado em 50 trajetórias, das quais 40 são utilizadas para treinamento, 5 para validação e 5 para teste.

Tabela 4: Acurácia Top-1: validação e teste.

Configuração	**val_top1**	**test_top1**
Index-only [27]	—	72,8%
Reprodução	75,5%	59,5%
Multimodal [24]	—	80,8%
Reprodução	66,1%	39,2%

5.2.2 Análise dos resultados intermediários

O método de predição de feixes baseado em LLM foi originalmente proposto em [27] utilizando o conjunto de dados *DeepMIMO*. Neste trabalho, reproduzimos esse método na configuração *index-only*, considerando apenas o histórico de índices de feixes, e o avaliamos no conjunto de dados *Multimodal-Wireless*. Como referência de comparação, utilizamos o *baseline index-only* reportado em [24] (Tabela III), treinado e avaliado no mesmo conjunto de dados, que alcança 72,8% de acurácia Top-1 no conjunto de teste. Nossa reprodução obteve 59,5% de acurácia Top-1 nesse mesmo conjunto. Essa diferença pode estar associada, entre outros fatores, ao particionamento das trajetórias utilizado durante o treinamento e a avaliação, cuja composição não é especificada no artigo original. No conjunto de validação, a acurácia alcançou 75,5%, indicando que a arquitetura é capaz de atingir desempenho próximo ao reportado na literatura quando avaliada em trajetórias com características semelhantes às observadas durante o treinamento.

Para a configuração multimodal completa (feixe + LiDAR + câmera), empregando o módulo *FrozenPointPillars*, o desempenho foi inferior ao da configuração *beam-only*, utilizada como referência na análise de ablação por corresponder à mesma arquitetura, porém sem a utilização das modalidades sensoriais. A acurácia Top-1 caiu de 77,4% para 66,1% na validação e de 61,8% para 39,2% no teste, o que corresponde a uma degradação de $-11{,}3$ pontos percentuais na validação e de $-22{,}6$ no teste. Em outras palavras, na configuração considerada, a incorporação das modalidades sensoriais não resultou em ganhos de desempenho; ao contrário, ocasionou uma degradação significativa da acurácia. Uma possível explicação é que o ruído introduzido pelos sensores afete mais os cenários fora da distribuição. Em ambos os conjuntos, o resultado permanece muito abaixo dos 80,8% reportados em [24].

Uma análise de ablação, realizada sobre a mesma arquitetura e o mesmo conjunto de dados, cujos resultados são apresentados na Tabela 5, indica que, no *setup* considerado, nem o LiDAR nem as imagens contribuíram para aumentar a acurácia do modelo.

Tabela 5: Ablação das modalidades sensoriais (acurácia Top-1 de validação).

Configuração	**val_top1**	**Δ**
beam-only (referência)	77,4%	—
+ câmera (ViT-B/16)	77,9%	+0,5 (neutro)
+ LiDAR (*PointPillars* treinável)	77,9%	+0,5 (neutro)
+ sensores (*FrozenPointPillars*)	66,1%	−11,3 (degrada)

A contribuição efetiva da câmera foi praticamente nula (+0,5 ponto, dentro da variação entre execuções). Isso se explica pelo campo de visão limitado da câmera do *Road Side Unit* (RSU), que enxerga o veículo em apenas parte de cada trajetória e, portanto, oferece informação espacial restrita. O resultado é consistente com a ablação de [24], que também aponta a câmera como a modalidade sensorial mais fraca.

Já a ausência de contribuição do LiDAR não é intrínseca ao sensor: decorre do extrator de BEV congelado (*FrozenPointPillars*), cujo mapa é quase constante entre cenas distintas. Prevê-se, portanto, um trabalho dedicado de pré-processamento e ajuste do *PointPillars* sobre o *Multimodal-Wireless*, com o objetivo de recuperar a contribuição efetiva dessa modalidade e, assim, reproduzir o ganho multimodal reportado no artigo.

Em síntese, a linha de base *index-only* foi reproduzida de forma consistente no conjunto de dados *Multimodal-Wireless*, ao passo que o ganho multimodal reportado no artigo original ainda não foi reproduzido no *setup* considerado. Os resultados obtidos indicam que essa diferença está predominantemente associada ao processamento da modalidade LiDAR, cuja implementação ainda demanda investigações adicionais para que seja possível reproduzir o ganho multimodal reportado na literatura.

5.3 Proposta de Contribuição

A principal contribuição proposta consiste em investigar o impacto da escolha do LLM na tarefa de predição multimodal de feixes, considerando simultaneamente o desempenho preditivo e a eficiência computacional dos modelos. Enquanto o trabalho original utiliza exclusivamente o GPT-2 Large, esta pesquisa propõe um estudo comparativo entre diferentes arquiteturas de LLMs, avaliando sua adequação para aplicações de comunicação sem fio assistidas por IA.

As próximas atividades planejadas incluem:

- **Comparação de *backbones* LLM alternativos**: investigando se arquiteturas LLM modernas superam o GPT-2 em predição de feixes com pesos congelados;
- **Avaliação em condições adversas**: utilizar dados de outras condições climáticas (chuva, nevoeiro) disponibilizados pelo *dataset Multimodal-Wireless* para avaliar a robustez do modelo treinado;
- **Análise energética**: quantificar o consumo energético durante a inferência dos diferentes modelos, permitindo investigar o compromisso entre desempenho e eficiência energética.

6 Conclusão

As atividades desenvolvidas neste segundo ano da Atividade 3.1 da Fase III do Projeto Brasil 6G consolidaram avanços importantes na aplicação de visão computacional e IA a problemas relevantes das futuras redes móveis. Em particular, os estudos realizados demonstram que a utilização conjunta de informações provenientes da interface de rádio e de sensores visuais constitui uma estratégia promissora para ampliar as capacidades de percepção do ambiente, localização de dispositivos e gerenciamento inteligente dos enlaces de comunicação em cenários 5G, B5G e 6G.

Ao longo deste relatório foram descritas diferentes abordagens. Inicialmente, foi explorada a estimação de posição em ambientes internos utilizando informações de rádio e imagens RGB obtidas por equipamentos comerciais de baixo custo, evidenciando que a incorporação de informações visuais pode aumentar a robustez dos métodos de localização baseados exclusivamente em sinais de rádio. Em seguida, foi proposta uma arquitetura multimodal para localização tridimensional de equipamentos de usuário por meio da fusão de informações rádio-visuais, explorando representações aprendidas de forma integrada. Complementarmente, foram avaliadas técnicas de NAS aplicadas ao problema de seleção de feixes em comunicações mmWave, permitindo comparar arquiteturas neurais sob diferentes critérios de desempenho, complexidade computacional e latência. Finalmente, foram iniciadas investigações sobre o emprego de LLMs multimodais na tarefa de predição de feixes, abrindo uma nova linha de pesquisa voltada à incorporação de modelos fundacionais em sistemas inteligentes de comunicação.

Os resultados obtidos também reforçam a importância da validação experimental utilizando plataformas reais e equipamentos comerciais, reduzindo a dependência exclusiva de cenários simulados e aproximando as soluções desenvolvidas das condições encontradas em futuras implantações. Além disso, a investigação simultânea de diferentes paradigmas de aprendizado de máquina amplia o conjunto de alternativas tecnológicas disponíveis para o desenvolvimento de aplicações inteligentes em redes móveis.

Como perspectivas para a continuidade da pesquisa, pretende-se expandir os experimentos para cenários mais desafiadores, contemplando ambientes externos, maior mobilidade, múltiplos usuários, condições severas de bloqueio e propagação sem linha de visada, bem como novas modalidades sensoriais e mecanismos mais sofisticados de fusão multimodal. Também serão investigadas estratégias para reduzir a complexidade computacional dos modelos, aumentar sua capacidade de generalização e permitir sua execução em plataformas de borda, aspecto fundamental para aplicações com requisitos rigorosos de latência e eficiência energética.

Referências

[1] N. Bai, G. Chen, R. Hou, and F. Ying, "A novel WiFi signal and RGB image fusion positioning method for manufacturing workshop," pp. 3229 – 3240, 2020.

[2] M. Abuyaghi, S. Si-Mohammed, G. Shaker, and C. Rosenberg, "Positioning in 5G Networks: Emerging Techniques, Use Cases, and Challenges," *IEEE Internet of Things Journal*, vol. 12, no. 2, pp. 1408–1427, 2025.

[3] D. Zhang, A. Prabhakara, S. Munir, A. C. Sankaranarayanan, and S. Kumar, "A Hybrid mmWave and Camera System for Long-Range Depth Imaging," 2021.

[4] H. Liu, J. Wan, P. Zhou, S. Ding, and W. Huang, "Augmented Millimeter Wave Radar and Vision Fusion Simulator for Roadside Perception," *Electronics*, vol. 13, no. 14, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/14/2729

[5] T. Sunami, S. Itahara, Y. Koda, T. Nishio, and K. Yamamoto, "Computer vision-assisted single-antenna and single-anchor RSSI localization harnessing dynamic blockage events," *arXiv preprint arXiv:2107.04770*, 2021.

[6] A. Saikko, J. Talvitie, and M. Valkama, "High-precision 3D location and orientation tracking: quaternion-based approach using cellular carrier phase and Doppler measurements," *Journal on Advances in Signal Processing*, 2026.

[7] O. Huan, T. Luo, and M. Chen, "Multi-Modal Image and Radio Frequency Fusion for Optimizing Vehicle Positioning," *IEEE Transactions on Mobile Computing*, vol. 24, no. 2, pp. 696–708, 2025.

[8] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "ViWi: A Deep Learning Dataset Framework for Vision-Aided Wireless Communications," in *IEEE Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.

[9] S. Mihara, S. Ito, T. Murakami, and H. Shinbo, "Positioning for User Equipment of a mmWave System Using RSSI and Stereo Camera Images," pp. 1–7, 2021.

[10] V. Conceição, F. Nunes, J. Borges, C. Nahum, I. Correa, S. Lins, and A. Klautau, "Low-Cost Experimental Platform for AI-Based Beam Management Using WiFi Radios," 01 2025.

[11] C. Balanis, *Antenna Theory: Analysis and Design*. Wiley, 2015. [Online]. Available: https://books.google.com.br/books?id=PTFcCwAAQBAJ

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004, pp. 153–156.

[13] B. Zholamanov, A. Saymbetov, M. Nurgaliyev, A. Bolatbek, G. Dosymbetova, N. Kuttybay, S. Orynbassar, A. Kapparova, N. Koshkarbay, and Ö. F. Beyca, "RSSI fingerprint-based indoor localization solutions using machine learning algorithms: A comprehensive review," *Smart Cities*, vol. 8, no. 5, p. 153, 2025.

[14] J. Chen, F. B. Teixeira, F. M. Ribeiro, A. Alkhateeb, M. Ricardo, L. M. Pessoa, and D. Slock, "Converge Challenge: Multimodal Learning for 6G Wireless Communications," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2026, pp. 21 799–21 801.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[16] J. Redmon, "You only look once: Unified, real-time object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[17] J. Arevalo, T. Solorio, M. M. y Gómez, and F. A. González, "Gated multimodal networks," *Neural Computing and Applications*, 2020.

[18] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conference on Artificial Intelligence*, 2018, pp. 3942–3951.

[19] A. Klautau, N. Gonzalez-Prelcic *et al.*, "LiDAR data for deep learning-based mmwave beam-selection," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 909–912, 2019.

[20] B. Salehi, G. Reus-Muns *et al.*, "Deep learning on multimodal sensor data at the wireless edge for vehicular network," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7639–7652, 2022.

[21] M. Zecchin, M. B. Mashhadi *et al.*, "LiDAR and position-aided mmwave beam selection with non-local CNNs and curriculum training," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 2979–2989, 2022.

[22] S. Gupta, N. Saxena, A. Roy, and A. Ravi, "Efficient mmwave beam selection using ViTs and GVEC: GPS-based virtual environment capture," in *International Conference on Computing Communication and Networking Technologies*, 2023.

[23] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *arXiv preprint arXiv:1907.10902*, 2019. [Online]. Available: https://arxiv.org/abs/1907.10902

[24] T. Mao, L. Liang, J. Yang, X. Li, and S. Jin, "Beam Prediction Based on Multimodal Large Language Models," *arXiv preprint arXiv:2603.15093*, Mar. 2026.

[25] T. Mao *et al.*, "Multimodal-Wireless: A Large-Scale Dataset for Sensing and Communication," *arXiv preprint arXiv:2511.03220*, Feb. 2026.

[26] A. H. Lang, S. Vora, H. Caesar, L. F. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast Encoders for Object Detection from Point Clouds," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[27] Y. Sheng, K. Huang, L. Liang, P. Liu, S. Jin, and G. Y. Li, "Beam prediction based on large language models," *IEEE Wireless Communications Letters*, 2025.