

UE Status as a Query in End-to-End Vision-Aided Beam Prediction for Mobile Millimeter-Wave Links

Rui Geng
Graduate School of
Information Science
and Technology,
The University of Osaka
Osaka, Japan
r-geng@ist.osaka-u.ac.jp

Daichi Kominami
Institute for Open
and Transdisciplinary
Research Initiatives,
The University of Osaka
Osaka, Japan
d-kominami@ist.osaka-u.ac.jp

Hideyuki Shimonishi
D3 Center,
The University of Osaka
Osaka, Japan
shimonishi.cmc@osaka-u.ac.jp

Masayuki Murata
Graduate School of
Information Science
and Technology,
The University of Osaka
Osaka, Japan
murata@ist.osaka-u.ac.jp

Abstract—In this paper, we propose UTOPIA, an end-to-end framework for vision-aided millimeter-wave beam prediction with UE status as a Query. Existing vision-aided beam prediction methods often predict the beam index from full-frame visual observations, but they do not explicitly maintain the status of the currently served UE when multiple visual instances appear in the scene. UTOPIA initializes the UE status from the initial target bounding box and updates it over consecutive RGB frames through cross-attention with full-frame visual tokens. The updated status is then used to predict a beam distribution over a predefined codebook. To guide target-related visual aggregation and beam supervision, UTOPIA combines instance-aligned attention with a soft-label beam prediction loss based on the received-power distribution. Experiments on the DeepSense 6G dataset show that UTOPIA achieves 0.759 Top-1 accuracy in the standard scenario. In multi-object scenarios, UTOPIA improves Top-1 accuracy from 0.196 to 0.848 and reaches Top-2 and Top-3 accuracies of 0.994 and 0.999, respectively.

Index Terms—millimeter-wave, mobile communications, beam prediction, visual sensing, object tracking, deep learning

I. INTRODUCTION

In recent years, millimeter-wave (mmWave) communication has been widely studied for Beyond 5G/6G mobile networks, because its large bandwidth can support high data rates [1]. However, mmWave links suffer from severe propagation loss, blockage, and rapid beam variation under mobility. Since exhaustive codebook-based beam sweeping or beam training introduces control overhead and access delay [2], predicting the optimal beam for the currently served user equipment (UE) with external sensing information is an important problem for beam management [3]–[5].

Infrastructure-side RGB cameras can provide information about UE position, road layout, and the surrounding environment at low cost by reusing existing visual infrastructure, such as traffic monitoring and roadside sensing systems. Existing vision-aided beam prediction studies usually formulate this task as classification from an image or an image sequence to the optimal beam index [4], [6], [7]. This full-frame formulation uses global scene information, but it usually does not maintain an explicit state for the served UE.

In real environments, an infrastructure-side camera often observes multiple pedestrians, vehicles, and other visual in-

stances. The propagation condition of an mmWave link depends on the served UE, nearby blockers and reflectors, and surrounding structures. If prediction uses only whole-image features without specifying the target UE, non-target instances may be used as evidence for the current link. Prior multi-candidate studies have shown that identifying the relevant user affects vision-aided beam prediction [8]. In the online beam maintenance setting considered in this paper, the target UE is specified at the initial time instant, and its state must be maintained over subsequent visual observations.

Based on these observations, this paper formulates vision-aided mmWave beam prediction as an online prediction problem with UE status as a Query. The UE status is a compact representation of the target UE, which is initialized from its bounding box and updated over time. For each frame, the Query attends to full-frame visual tokens and gathers features related to the target instance. The updated status is then used as a target-conditioned representation to predict the beam distribution. This formulation makes target initialization, online state update, and beam prediction parts of a single end-to-end model.

In this paper, we propose UTOPIA (UE-Targeted Online Beam Prediction with Instance-Aligned Attention), an end-to-end framework for vision-aided mmWave beam prediction. UTOPIA first constructs the UE status from the initial target bounding box. In subsequent frames, cross-attention updates this status using full-frame visual tokens, so that the identity and temporal continuity of the served UE are maintained during prediction. To guide the Query to the target instance more consistently, UTOPIA introduces instance-aligned attention. It also uses a soft-label beam prediction loss to learn the received-power structure among candidate beams, because neighboring beams may have similar received powers and should not be treated in the same way as distant beams. Fig. 1 shows the overall structure of UTOPIA.

The main contributions of this paper are as follows:

- 1) We formulate vision-aided mmWave beam prediction as an online prediction problem driven by the status of the currently served UE.

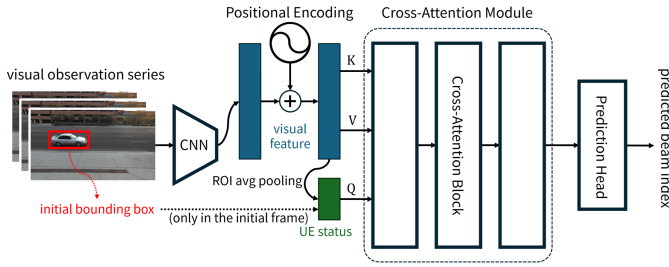


Fig. 1. Overall framework of UTOPIA.

- 2) We propose UTOPIA, an end-to-end architecture that uses UE status as a Query to aggregate target-related features from full-frame visual tokens for beam prediction.
- 3) We evaluate UTOPIA under standard conditions in Scenario 5, validate it in multi-object scenarios in Scenarios 10–12, and conduct an ablation study to assess the effect of soft-label supervision.

II. RELATED WORK

A. Vision-Aided mmWave Beam Prediction

Vision-aided mmWave beam prediction uses visual or scene-sensing information, such as RGB images, LiDAR measurements, or historical visual observations, to support beam management and infer the beam required by the current link. Camera-based studies showed that visual observations from base-station-side cameras can be used for mmWave beam prediction and blockage prediction [3], [9]. Datasets such as ViWi and DeepSense 6G provided aligned sensing, positional, and wireless measurements, enabling more consistent evaluation on real data [6], [7].

Based on these platforms, RGB-based beam prediction learns from image observations [10], and vision-position variants additionally use position information [4]. LiDAR-aided beam prediction uses geometric scene measurements to infer current or future beams [11], while computer-vision-aided beam tracking uses consecutive visual observations to maintain beam alignment [5]. Environment semantic aided communication further extracts communication-related semantic information from images [12]. These studies show that scene information is useful for beam management, but their modeling focus is often still the mapping from visual, geometric, multi-modal, or temporal inputs to the beam index. In many settings, the target link or UE is already fixed by the dataset design, input sample, or external position information. Therefore, the model does not have to explicitly decide which visual instance corresponds to the currently served UE. The status of this UE, and its continuity over visual observations, are usually not maintained as internal variables of the model. When multiple visual instances appear in the scene, the model has limited constraints on how visual evidence should be organized around the target UE.

B. Multi-Candidate Scenarios and Query-Based Representations

Multi-candidate beam prediction studies deal with scenes where multiple visual candidates appear in the camera view. A representative approach detects candidate objects in each frame, identifies the transmitter with position information, and predicts the beam based on the identified user [8]. This direction indicates that user identification is an important factor for vision-aided beam prediction when several possible users exist in the visual scene. Its main focus, however, is frame-wise transmitter identification from detected candidates.

In the online beam maintenance setting considered here, the target UE is specified at the initial time instant. The task is then to maintain its status over continuous visual observations and use the maintained status to guide subsequent beam prediction.

Query-based representations are useful for modeling this problem. In Transformer models, a Query retrieves relevant information from Key-Value tokens [13]. In visual models such as DETR, object queries are used to extract target-level information from image features [14]. However, these queries are usually representations designed for object detection or sequence modeling, and they do not directly contain the status of a UE in a wireless link. In the proposed formulation, the Query is specified as UE status, which is initialized from the target UE and updated over time. This Query-based representation is used for online UE status maintenance in end-to-end vision-aided mmWave beam prediction.

III. METHOD

A. Problem Formulation

We consider a millimeter-wave link between a single base station and one currently served UE. The single UE in this formulation denotes the served UE of the communication link, while the infrastructure-side camera may observe multiple visual instances in the same scene. Given the initially specified target UE, the model predicts the mmWave beam required by the current link from consecutive RGB visual observations.

The system has a predefined beam codebook $\mathcal{B} = \{b_1, b_2, \dots, b_{N_b}\}$, where each beam corresponds to a fixed spatial steering pattern. Given an RGB image sequence $\mathbf{I}_{1:t} = \{\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_t\}$ and the bounding box of the target UE at the initial time instant, the model outputs a predicted distribution $\hat{\mathbf{p}}_t \in \Delta^{N_b-1}$ over the beam codebook at each time t . Here, $\hat{p}_{t,i}$ denotes the predicted probability of selecting beam b_i . Thus, the task is formulated as online beam distribution prediction over a predefined discrete beam set. The final beam index decision is given by the beam prediction head described later.

B. Proposed UTOPIA Framework

The proposed UTOPIA framework maintains the status of the currently served UE and uses it as a Query for online beam prediction. Each RGB frame is first encoded by the visual backbone into full-frame visual tokens. At the initial time instant, the target bounding box is used to construct the initial UE status from the first-frame tokens. At later time instants, the previous UE status is used as the Query, while the current

visual tokens are used as Key and Value in cross-attention. The updated UE status is then passed to the beam prediction head to produce the beam distribution over the predefined codebook.

1) *Visual Token Representation*: For an input image $\mathbf{I}_t \in \mathbb{R}^{3 \times H \times W}$ at time t , the model first uses a pretrained backbone $f_{bb}(\cdot)$ to extract a feature map:

$$\mathbf{F}_t = f_{bb}(\mathbf{I}_t) \in \mathbb{R}^{C \times H_f \times W_f}, \quad (1)$$

where C is the number of output channels, and (H_f, W_f) is the spatial resolution of the feature map.

Then, a 1×1 convolution $\phi(\cdot)$ projects the channel dimension to the hidden dimension d :

$$\tilde{\mathbf{F}}_t = \phi(\mathbf{F}_t) \in \mathbb{R}^{d \times H_f \times W_f}. \quad (2)$$

To preserve spatial information in the visual tokens, a two-dimensional sine-cosine positional encoding [14] $\mathbf{P} \in \mathbb{R}^{d \times H_f \times W_f}$ is added to the feature map:

$$\hat{\mathbf{F}}_t = \tilde{\mathbf{F}}_t + \mathbf{P} \in \mathbb{R}^{d \times H_f \times W_f}. \quad (3)$$

Finally, the two-dimensional feature map is flattened into a token sequence:

$$\mathbf{X}_t = \text{Flatten}(\hat{\mathbf{F}}_t) \in \mathbb{R}^{N \times d}, \quad N = H_f W_f. \quad (4)$$

Here, $\mathbf{X}_t[n]$ denotes the visual token at the n -th spatial location in the current frame. This token sequence is used as the candidate Key and Value set in the following cross-attention operation.

2) *UE Status Initialization*: To specify the target UE for beam prediction, the model uses the bounding box of the target UE at the initial time instant, denoted by $\mathbf{b}_1 = (x_1, y_1, x_2, y_2)$. Based on this bounding box, the initial UE status $\mathbf{z}_1 \in \mathbb{R}^d$ is constructed from the token sequence $\mathbf{X}_1$ of the first frame. Fig. 2 illustrates this initialization process.

Specifically, the bounding box is mapped from the input image coordinate system to the feature map coordinate system, and the token index set inside the target region is defined as $\Omega_1 \subset \{1, \dots, N\}$. The model then applies average pooling over the visual tokens in this region to obtain a visual aggregation vector:

$$\mathbf{r}_1 = \frac{1}{|\Omega_1|} \sum_{n \in \Omega_1} \mathbf{X}_1[n] \in \mathbb{R}^d. \quad (5)$$

To include the spatial location of the target region in the initial status, the model also uses the normalized bounding box geometry:

$$\mathbf{g}_1 = \left[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H} \right]^\top \in \mathbb{R}^4. \quad (6)$$

Here, H and W denote the height and width of the input image, respectively. The visual aggregation vector and the geometry vector are concatenated and mapped by a learnable function $\phi_{\text{init}}(\cdot)$. The initial UE status is obtained as

$$\mathbf{z}_1 = \phi_{\text{init}}([\mathbf{r}_1; \mathbf{g}_1]) \in \mathbb{R}^d. \quad (7)$$

The vector $\mathbf{z}_1$ is the initial status representation of the target UE, and it is also the initial Query for the following query-based status update.

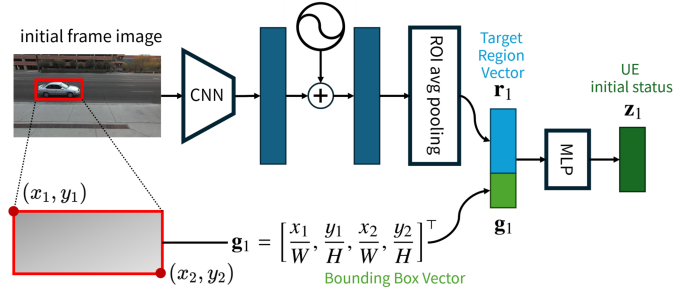


Fig. 2. UE status initialization from the initial target bounding box.

3) *Query-Based UE Status Update*: At time $t \geq 2$, the model has the UE status $\mathbf{z}_{t-1} \in \mathbb{R}^d$ from the previous time instant and the visual token sequence $\mathbf{X}_t \in \mathbb{R}^{N \times d}$ extracted from the current frame. UTOPIA uses $\mathbf{z}_{t-1}$ as the Query and uses $\mathbf{X}_t$ as Key and Value. Through cross-attention [13], the model aggregates visual evidence related to the currently served UE from full-frame visual tokens and updates the UE status.

For simplicity of notation, let h denote the number of attention heads, and let d_k and d_v denote the projected dimensions of Key and Value, respectively. The ℓ -th cross-attention block has learnable projection parameters $\{\mathbf{W}_Q^{(\ell)}, \mathbf{W}_K^{(\ell)}, \mathbf{W}_V^{(\ell)}, \mathbf{W}_O^{(\ell)}\}$. Let the input state of the block be $\mathbf{z}_t^{(\ell-1)} \in \mathbb{R}^d$, where $\mathbf{z}_t^{(0)} = \mathbf{z}_{t-1}$.

For the m -th attention head, the Query, Key, and Value are first computed as

$$\begin{aligned} \mathbf{q}_{t,m}^{(\ell)} &= \mathbf{z}_t^{(\ell-1)} \mathbf{W}_{Q,m}^{(\ell)} \in \mathbb{R}^{d_k}, \\ \mathbf{K}_{t,m}^{(\ell)} &= \mathbf{X}_t \mathbf{W}_{K,m}^{(\ell)} \in \mathbb{R}^{N \times d_k}, \\ \mathbf{V}_{t,m}^{(\ell)} &= \mathbf{X}_t \mathbf{W}_{V,m}^{(\ell)} \in \mathbb{R}^{N \times d_v}. \end{aligned} \quad (8)$$

The attention weight and context vector of this head are then calculated as

$$\begin{aligned} \alpha_{t,m}^{(\ell)} &= \text{Softmax} \left(\frac{\mathbf{K}_{t,m}^{(\ell)} \mathbf{q}_{t,m}^{(\ell)}}{\sqrt{d_k}} \right) \in \mathbb{R}^N, \\ \mathbf{c}_{t,m}^{(\ell)} &= \alpha_{t,m}^{(\ell)\top} \mathbf{V}_{t,m}^{(\ell)} \in \mathbb{R}^{d_v}. \end{aligned} \quad (9)$$

The outputs of all heads are concatenated and projected to obtain the output of the cross-attention sublayer:

$$\mathbf{c}_t^{(\ell)} = \text{Concat}(\mathbf{c}_{t,1}^{(\ell)}, \dots, \mathbf{c}_{t,h}^{(\ell)}) \mathbf{W}_O^{(\ell)} \in \mathbb{R}^d. \quad (10)$$

After the cross-attention output is obtained, the intermediate status is updated by a residual connection and Layer Normalization:

$$\tilde{\mathbf{z}}_t^{(\ell)} = \text{LN} \left(\mathbf{z}_t^{(\ell-1)} + \mathbf{c}_t^{(\ell)} \right) \in \mathbb{R}^d. \quad (11)$$

Then, a feed-forward network further updates the status:

$$\mathbf{z}_t^{(\ell)} = \text{LN} \left(\tilde{\mathbf{z}}_t^{(\ell)} + \text{FFN}^{(\ell)}(\tilde{\mathbf{z}}_t^{(\ell)}) \right) \in \mathbb{R}^d. \quad (12)$$

After stacking L cross-attention blocks within the same frame, the final status is obtained as $\mathbf{z}_t = \mathbf{z}_t^{(L)}$.

For attention supervision during training, the attention distribution is defined by averaging the final-layer attention weights over heads:

$$\alpha_t \triangleq \frac{1}{h} \sum_{m=1}^h \alpha_{t,m}^{(L)} \in \mathbb{R}^N. \quad (13)$$

Since the Query comes from the UE status at the previous time instant, the update process maintains the target UE over consecutive visual observations through recurrent status update and target-conditioned retrieval.

4) *Beam Distribution Prediction*: After obtaining the UE status $\mathbf{z}_t \in \mathbb{R}^d$ at time t , the model maps it to classification outputs over the predefined beam codebook $\mathcal{B} = \{b_1, \dots, b_{N_b}\}$ by the beam prediction head. Specifically, let $h_{\text{beam}}(\cdot)$ denote a learnable mapping. The logit vector is generated as

$$\mathbf{s}_t = h_{\text{beam}}(\mathbf{z}_t) \in \mathbb{R}^{N_b}, \quad (14)$$

where $s_{t,i}$ represents the confidence of selecting beam b_i . A Softmax operation is then applied to obtain the beam distribution:

$$\hat{\mathbf{p}}_t = \text{Softmax}(\mathbf{s}_t) \in \Delta^{N_b-1}, \quad \hat{p}_{t,i} = \frac{\exp(s_{t,i})}{\sum_{j=1}^{N_b} \exp(s_{t,j})}. \quad (15)$$

During online inference, the model outputs the beam index with the largest probability:

$$\hat{k}_t = \arg \max_{i \in \{1, \dots, N_b\}} \hat{p}_{t,i}. \quad (16)$$

C. Training Objective

During training, the model optimizes a soft-label beam prediction loss and an instance-aligned attention loss. The training samples are fixed-length sequences of length T , and the model outputs a beam probability distribution $\hat{\mathbf{p}}_t \in \Delta^{N_b-1}$ and an attention distribution α_t for each frame.

1) *Soft-label Beam Prediction Loss*: In mmWave beam prediction, candidate beams near the optimal beam may have similar received powers. If only one-hot labels are used, all non-optimal beams are treated as equally incorrect, and the relative received quality among candidate beams is ignored. Therefore, UTOPIA uses a soft-label beam distribution as the supervision signal.

Let the ground-truth beam distribution at time t be $\mathbf{p}_t \in \Delta^{N_b-1}$, and let the model prediction be $\hat{\mathbf{p}}_t$. The beam prediction loss is defined as the time-averaged cross entropy:

$$\mathcal{L}_{\text{beam}} = -\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^{N_b} p_{t,i} \log \hat{p}_{t,i}. \quad (17)$$

When $\mathbf{p}_t$ degenerates into a one-hot distribution, this loss is equivalent to the standard multi-class cross entropy.

2) *Instance-Aligned Attention Loss*: To guide cross-attention toward the target UE region, an instance-aligned attention loss is introduced. Let the token index set corresponding to the target UE region at time t be $\Omega_t \subset \{1, \dots, N\}$. A

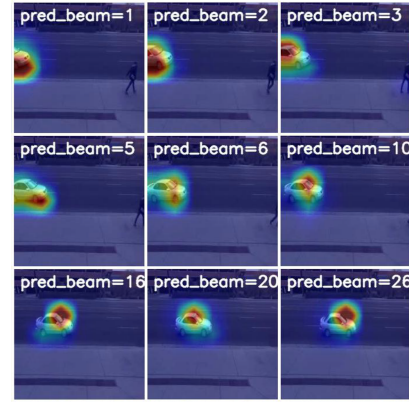


Fig. 3. Visualization of target-related attention under different predicted beam indices.

normalized mask distribution $\mathbf{m}_t \in \Delta^{N-1}$ is constructed from this region:

$$m_{t,n} = \begin{cases} \frac{1}{|\Omega_t|}, & n \in \Omega_t, \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

The instance-aligned attention loss is defined as the cross entropy from the target mask distribution to the attention distribution:

$$\mathcal{L}_{\text{attn}} = -\frac{1}{T} \sum_{t=1}^T \sum_{n=1}^N m_{t,n} \log(\alpha_{t,n} + \epsilon), \quad (19)$$

where ϵ is a small constant for numerical stability.

This loss encourages the attention distribution to align with the target mask distribution. Fig. 3 visualizes target-related attention produced by UTOPIA.

3) *Total Loss*: The final training objective is the weighted sum of the beam prediction loss and the instance-aligned attention loss:

$$\mathcal{L} = \mathcal{L}_{\text{beam}} + \lambda \mathcal{L}_{\text{attn}}, \quad (20)$$

where $\lambda \geq 0$ controls the strength of the attention alignment constraint. By optimizing these two losses jointly, the model learns the received-power structure among candidate beams and is guided to use spatial visual evidence related to the target UE for prediction.

IV. EVALUATION

A. Experimental Setup

1) *Dataset and Beam Codebook*: We evaluate UTOPIA using the DeepSense 6G dataset [7]. Each sample consists of an infrastructure-side RGB image and the received-power vector over the mmWave beam codebook at the same time instant. The RGB images have an original resolution of 960×540 and are recorded at 30 FPS. The mmWave system operates at the 60 GHz band. The receiver uses a 16-element phased array and provides received-power measurements over a 64-beam codebook.

To keep the evaluation setting consistent with prior vision-aided beam prediction studies, we adopt a beam classification setting with $N_b = 64/\delta$ classes. The original 64-beam

codebook is uniformly downsampled into a sub-codebook. For hard-label evaluation, the beam with the maximum received power is treated as the ground-truth optimal beam. For soft-label training, the beam supervision is constructed from the received-power distribution over the sub-codebook.

2) *Evaluation Scenarios*: The evaluation consists of a standard scenario and multi-object scenarios. The standard scenario uses Scenario 5 of DeepSense 6G and provides a common setting for comparing UTOPIA with the baseline under regular vision-aided beam prediction conditions.

The multi-object scenarios use Scenarios 10–12 of DeepSense 6G. In these scenarios, multiple visual instances may appear in the camera view, and the beam prediction target is the currently served UE. This setting is used to evaluate the effect of target-conditioned status update when multiple visual candidates are observed in the same scene.

3) *Training and Online Inference Protocol*: Training and test data are divided at the sequence level with a ratio of 7:3. During training, fixed-length windows of $T = 16$ frames are generated from each training sequence. The model takes an RGB frame sequence and the initial target bounding box as input, and outputs the beam distribution at each time instant.

All evaluations are conducted under the online inference protocol. At the first frame where the target UE appears, the model receives its initial bounding box and uses it to initialize the UE status. After initialization, RGB images are fed to the model sequentially in temporal order. The prediction at each time instant is based only on the current and past visual observations. The target region mask is used as supervision for the instance-aligned attention loss during training. During online inference, the model requires only the initial bounding box and RGB images.

B. Implementation Details

UTOPIA uses ResNet-50 [15] as the visual backbone and converts each RGB image into full-frame visual tokens. The UE status update module is composed of stacked cross-attention blocks. The model is trained using the AdamW optimizer [16], and gradient clipping is applied to stabilize training.

For UTOPIA, we evaluate two backbone training settings. In the frozen-backbone setting, the ResNet-50 backbone is fixed, and only the token projection, UE status update module, and beam prediction head are trained. In the fine-tuned-backbone setting, the later layers of ResNet-50 are unfrozen during training, so that the visual backbone can be further adapted to the beam prediction task. The two settings use the same input, model structure, loss functions, and online inference protocol, except for whether the visual backbone is fine-tuned.

C. Baselines

We compare UTOPIA with a reproduced implementation of a vision-aided mmWave beam prediction baseline [4]. The baseline formulates beam prediction as a multi-class classification problem based on the full RGB image. It uses

TABLE I
PREDICTION ACCURACY IN THE STANDARD SCENARIO ($\delta = 2$).

Method	Top-1	Top-2	Top-3
Baseline	0.745	0.912	0.974
UTOPIA – frozen backbone	0.739	0.931	0.998
UTOPIA – fine-tuned backbone	0.759	0.940	0.991

TABLE II
PREDICTION ACCURACY IN THE MULTI-OBJECT SCENARIOS ($\delta = 4$).

Method	Top-1	Top-2	Top-3
Baseline	0.196	0.392	0.449
UTOPIA – frozen backbone	0.754	0.976	0.996
UTOPIA – fine-tuned backbone	0.848	0.994	0.999

a pretrained ResNet-50 to extract visual features and predicts the beam class probability through a classification head.

We reproduce the baseline using the same sequence-level split ratio and the subsequences generated in our evaluation protocol. The baseline takes each full RGB frame in these subsequences as input and outputs the beam distribution at the current time instant. UTOPIA uses the same RGB inputs and additionally uses the initial target bounding box to construct the UE status.

D. Metrics

We use Top- k accuracy as the evaluation metric. At a given time instant, a prediction is counted as correct if the ground-truth optimal beam is included in the top k beam candidates with the highest predicted probabilities.

We report Top-1, Top-2, and Top-3 accuracy. Top-1 accuracy measures the ability to directly select the optimal beam. Top-2 and Top-3 accuracy measure whether the optimal beam can be included in a small candidate beam set.

E. Results

1) *Standard Scenario*: Table I shows the Top- k accuracy of the baseline and UTOPIA in the standard scenario. For UTOPIA, results are reported under both the frozen-backbone and fine-tuned-backbone settings.

With the frozen backbone, UTOPIA obtains a Top-1 accuracy of 0.739, which is close to the baseline result of 0.745. Its Top-2 and Top-3 accuracies are 0.931 and 0.998, both higher than the baseline results. This indicates that the frozen-backbone model can place the ground-truth beam in a small candidate set more reliably, while its direct Top-1 prediction is slightly lower than the baseline.

With the fine-tuned backbone, UTOPIA achieves the best Top-1 and Top-2 accuracies, reaching 0.759 and 0.940, respectively. The Top-3 accuracy is 0.991, which is higher than the baseline but lower than the frozen-backbone setting. These results show that fine-tuning the later layers of the visual backbone improves the direct prediction of the optimal beam in the standard scenario.

2) *Multi-Object Scenarios*: Table II summarizes the Top- k accuracy in the multi-object scenarios. The reported values are aggregated over Scenarios 10–12.

TABLE III
ABLATION RESULTS FOR SOFT-LABEL SUPERVISION ($\delta = 2$).

Variant	Top-1	Top-2	Top-3
UTOPIA	0.759	0.940	0.991
UTOPIA – w/o soft label	0.714	0.874	0.938

In the multi-object scenarios, multiple visual instances appear in the camera view, and beam prediction is performed for the currently served UE. The baseline predicts the beam distribution directly from the full RGB image. UTOPIA initializes the UE status from the initial target bounding box and updates this status over time through query-based visual aggregation.

The baseline obtains Top-1, Top-2, and Top-3 accuracies of 0.196, 0.392, and 0.449, respectively. UTOPIA with the frozen backbone reaches 0.754, 0.976, and 0.996. With the fine-tuned backbone, UTOPIA further improves the results to 0.848, 0.994, and 0.999. These results show that the proposed method is effective when multiple visual instances are observed in the same scene. In UTOPIA, the initial target bounding box specifies the currently served UE, and the query-based update uses the UE status to aggregate target-related visual information from full-frame visual tokens.

F. Ablation Study

This experiment evaluates the effect of soft-label supervision on beam prediction. We compare UTOPIA with a hard-label variant trained using one-hot beam labels. Both models use the same network structure, input setting, and online inference protocol. The results are shown in Table III.

UTOPIA obtains Top-1, Top-2, and Top-3 accuracies of 0.759, 0.940, and 0.991, respectively. When soft-label supervision is removed, the accuracies decrease to 0.714, 0.874, and 0.938. The largest gap is observed in Top-2 accuracy, where the result decreases from 0.940 to 0.874. This indicates that the received-power distribution provides useful supervision for learning the ranking of candidate beams.

V. CONCLUSION

We presented UTOPIA, an end-to-end framework for vision-aided mmWave beam prediction with UE status as a Query. UTOPIA initializes the status of the currently served UE from its initial bounding box, updates this status over consecutive RGB frames through cross-attention with full-frame visual tokens, and predicts a beam distribution over a predefined codebook. The framework also uses instance-aligned attention and soft-label beam supervision to guide target-related visual aggregation and learn the received-power structure among candidate beams.

Experiments on DeepSense 6G show that UTOPIA provides clear advantages in multi-object scenarios, where several visual instances appear in the same camera view. In the standard scenario, the fine-tuned model achieves the best Top-1 and Top-2 accuracies, while the frozen-backbone model achieves the best Top-3 accuracy. In Scenarios 10–12, UTOPIA improves Top-1 accuracy from 0.196 to 0.848 and reaches Top-2 and Top-3 accuracies of 0.994 and 0.999. An ablation

study confirms that soft-label supervision improves all Top- k metrics.

Future work will extend UTOPIA to multi-camera and multi-modal sensing, and evaluate it in closed-loop beam management.

REFERENCES

- [1] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, J. S. Thompson, E. G. Larsson, M. Di Renzo, W. Tong, P. Zhu, X. Shen, H. V. Poor, and L. Hanzo, "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 905–974, 2023.
- [2] K. Ma, Z. Wang, W. Tian, S. Chen, and L. Hanzo, "Deep learning for mmWave beam-management: State-of-the-art, opportunities and challenges," *IEEE Wireless Communications*, vol. 30, no. 4, pp. 108–114, 2023.
- [3] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.
- [4] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, "Vision-position multi-modal beam prediction using real millimeter wave datasets," in *Proceedings of the 2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 2727–2731.
- [5] S. Jiang and A. Alkhateeb, "Computer vision aided beam tracking in a real-world millimeter wave deployment," in *Proceedings of the 2022 IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 142–147.
- [6] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "ViWi: A deep learning dataset framework for vision-aided wireless communications," in *Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.
- [7] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. 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.
- [8] G. Charan, M. Alrabeiah, T. Osman, and A. Alkhateeb, "Camera based mmWave beam prediction: Towards multi-candidate real-world scenarios," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 4, pp. 5897–5913, 2025.
- [9] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided 6G wireless communications: Blockage prediction and proactive handoff," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10193–10208, 2021.
- [10] H. Wang, B. Ou, X. Xie, and Y. Wang, "Vision-aided mmWave beam and blockage prediction in low-light environment," *IEEE Wireless Communications Letters*, vol. 14, no. 3, pp. 791–795, 2025.
- [11] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Communications Letters*, vol. 12, no. 2, pp. 212–216, 2023.
- [12] S. Imran, G. Charan, and A. Alkhateeb, "Environment semantic aided communication: A real world demonstration for beam prediction," in *Proceedings of the 2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2023, pp. 48–53.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008.
- [14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*. Springer, 2020, pp. 213–229.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: <https://openreview.net/forum?id=Bkg6RiCqY7>