Multimodal Cost-Aware DRL for Beam Realignment in mmWave V2X Communications

Nuri Choi*, Taesik Nam†, Han-Shin Jo‡
Dept. Automotive Engineering (Automotive–Computer Convergence) of Hanyang University, Seoul, Republic of Korea *
Dept. Automotive Engineering of Hanyang University, Seoul, Republic of Korea (Corresponding Author) ‡
Dept. Electrical and Electronic Engineering of Yonsei University, Seoul, Republic of Korea †
Email: {angela6nuri, hsjo23}@hanyang.ac.kr, ts.nam@yonsei.ac.kr

***Abstract*—Millimeter-wave (mmWave) bands offer large bandwidth but suffer from blockage, making rapid beam re-alignment critical for V2X communication. Conventional beam management relies on exhaustive codebook search, causing high time-domain overhead. Vision-aided studies reduce this cost by predicting the best beam index, yet they still overlook when re-alignment should occur. To address this gap, we propose a DRL-based policy that learns optimal re-alignment timing using sector probability, temporal stability, link fluctuations, and switching rate. The vision-aided model shows stronger proactive behavior than non-vision baselines, demonstrating the importance of visual cues for early degradation detection and efficient beam management.**

***Index Terms*—Beamforming, Millimeter-Wave Communication, Multi-Modality, Deep Reinforcement Learning**

I. Introduction

Millimeter wave (mmWave) bands, including 28GHz and 57-71GHz bands, provide abundant spectral resources. Therefore they are increasingly regarded as a major band of next-generation wireless systems. However, due to strong straightness of radio waves, it is highly vulnerable to unexpected obstacles, which can make it significantly challenge when performing Vehicle-to-Everything communication (V2X) communication in mmWave band.

Conventional beam management performs exhaustive codebook search, resulting in high time-domain overhead. To address this issue, several recent works have utilized various modalities such as computer vision. [1] has utilized position, visual, and point cloud features to provide richer contextual information. Using these multimodal features, their method improves optimal beam prediction and reduces search overhead. [2] extracts position or motion information from images to improve mmWave beam management, and reinforcement learning is applied to continuous beam estimation. This work has alleviated the limitations of codebook-dependent classification.

However, prior vision-aided mmWave beam prediction methods mainly focus on predicting the optimal beam index without considering when to perform beam realignment. In addition, these methods do not incorporate sweeping overhead, realignment costs into the optimization process. To address these limitations, we proposed new methods which determines the optimal trigger timing based on sector probability, link quality variations, temporal stability features and trigger costs by deep reinforcement learning. This approach enables us to aware cost such as incur time and overhead costs from beam measurements and switching overhead generated from beam switching and adaptive beam realignment scheduling, providing a novel direction for practical V2X mmWave beam management.

II. System Model

A. Beam management

We consider beam management defined in 3rd Generation Partnership Project-New Radio (3GPP NR) at mmWave frequencies for the V2X system composed of next-generation NodeB (gNB) and vehicle User Equipment (UE). The Standalone-Downlink Scheme has 4 phases: beam sweep, beam measurement, beam determination, and beam reporting. In beam sweeping phase, gNB periodically transmit Synchronization Signal Block (SS Block) to predefined angles. As a result of exhaustive search of codebook beam, UE detect synchronization and reference signals. In beam measurement phase, the UE measures the quality of the SS blocks and the UE selects the beam which shows the maximum Signal-to-Noise ratio (SNR) above a predefined threshold in beam determination step. Lastly in beam reporting, the UE reports the selected beam using a Random Access Channel (RACH) preamble.

B. Channel model

In mmWave systems, the link quality is characterized by the received power of reference signals (SSB/CSI-RS) across the beam directions. Following the 3GPP NR modeling framework, the received power for beam b at time t is expressed as

$$P_t(b) = |h_t(b)|^2 P_{\text{tx}},$$

where $h_t(b)$ denotes the complex channel coefficient of beam b, and P_{tx} is the transmit power. The UE then identifies the beam with the maximum received power as the representative link-quality metric:

$$y_t = \max_b P_t(b).$$

where y_t is the received power of the optimal beam pair.

C. IA and beam tracking for Beam management

For the experiment, we defined 2 assumptions to support the experimental setup. First, periodic SSB/CSI-RS bursts are assumed to be reliably received by the UE, and therefore signal acquisition failures are not considered in this work. Second, after successful Initial Access (IA), the UE is assumed to perform beam tracking based on continuous CSI-RS reception. In this tracking phase, beam measurements and beam switching are assumed to occur to maintain beam alignment.

III. Proposed Method

A. Multimodal DRL-Based Beam Realignment Framework

Our multimodal DRL model determines optimal beam retrigger timing under link and environmental uncertainty.

a) Data preprocessing: Several preprocessing procedures were applied to the dataset before model training. First, beam power vectors are collapsed from 64 beams to 8 coarse sectors by adding their values, making a more stable directional representation. Second, power measurements are normalized using mean and standard deviation across scenarios. Third, using a sliding window, we made short temporal sequences to capture recent beam dynamics and mobility. Finally, we sampled the data in to reduce high-frequency measurements and to match the timescale of realistic beam dynamics.

b) YOLO-Based Visual Sector Prior: We detect the position of the UE in the input image and detected features are passed through the Multi-Layer Perceptron (MLP) to estimate the probability distribution of the object present in each of the 8 sectors. This feature shows the environmental changes and movement pattern which cannot be known by mmWave power vector, providing an important information to early predict the future link degradation. The 8-dimensional visual features are directly combined with the DRL policy, allowing the agent to make a proactive beam-switching decision considering environmental risk factors.

c) GRU-Based Temporal Stability Encoder: To model these temporal dynamics effectively, we utilized a Gated Recurrent Unit (GRU) as the time series encoder. The input sequence is consisted of power of each sectors and visual sector probabilities by YOLO. The hidden-state sequence H_t that summarizes changes across time is made as a result. This hidden state captures fluctuations in beam probability distribution, instantaneous instability, and direction of the movement. The final hidden state H_t is mapped into a scalar stability indicator. By capturing short-term temporal consistency in link dynamics, the GRU may also help alleviate reliance on visual inputs when visual sensing is uncertain.

d) State Fusion Model: This study designed a state fusion module that combines multiple features into a single integrated state vector so that the agent can decide the action using information obtained from various modalities. This module includes the normalized power value, recent power change, the distance between the current sector and the optimal sector, and the image model integrates visual features such as sector probability. In addition, the H_t extracted from the past power sequence, visual sector probabilities through GRU is included to reflect the temporal stability of the link. Layer normalization is applied to stably combine heterogeneous features of different scales, and the final state s_t is used as the input of the PPO policy network. This structure enables more reliable proactive decision-making by simultaneously utilizing spatial, visual, and temporal cues that are difficult to capture with a single modality.

e) PPO-Based Decision Policy: The policy is optimized using Proximal Policy Optimization (PPO). PPO uses a clipping mechanism to limit the amount of policy updates. This prevents excessive deviations from the previous policy and enables stable convergence with less samples. The policy network is composed of an actor head that makes action probabilities and a critic head that estimates the state value. Considering the strong temporal correlation in mmWave environment, we used a small batch size and multi-epoch updates.

TABLE I
Training hyperparameters used in proposed model

Parameter	Value
Learning rate	5×10^{-5}
Batch size	64
PPO epochs per update	5
Discount factor γ	0.98
Clipping ratio ϵ	0.25

IV. Multimodal DRL-Based Beam Realignment Algorithm Verification

A. DeepSense 6G For Model Verification

In this work, we employ the DeepSense 6G dataset for training and evaluating the proposed method. DeepSense 6G is real-world multimodal dataset which captures synchronized wireless, visual, and position data for 6G research. Especially, we use 4 V2I scenarios, 1, 6, 8, 9. These scenarios are different the conditions cover diverse LoS/NLoS, mobility and environmental conditions.

B. Performance Evaluation Metrics

We evaluate our model with 4 metrics. First, the Proactive Trigger Success rate (PTSR) measures whether the policy successfully initiates re-alignment before link degradation occurs, reflecting its predictive capability. Second, the Missed Degradation Rate (MDR) indicates insufficient sensitivity to early instability. Third, the False Trigger Rate (FTR) represents unnecessary re-alignments measuring decision overhead. Finally, the lead time indicates how many steps the model anticipates degradation in advance, directly capturing the degree of proactive behavior.

C. Model Verification Result

The proposed method is trained using PPO with a CNN-GRU architecture. Important hyperparameters, including learning rates, batch sizes, and PPO coefficients, are defined in

TABLE II
COMPARISON OF IMAGE-BASED MODEL AND NON-IMAGE ABLATION MODEL

Metric	Image	Non-image
Average Reward ($\overline{R}$)	-18.48	-28.90
Switch Rate Avg.	0.46	0.03
Average Realignment Interval ($\overline{\tau}$) [steps]	2.17	7.246
Proactive Trigger Metrics		
PTSR (Proactive Success Rate)	0.448	0.062
MDR (Missed Degradation Rate)	0.379	0.920
FTR (False Trigger Rate)	0.237	0.381
Average Lead Time [steps]	2.292	0.000

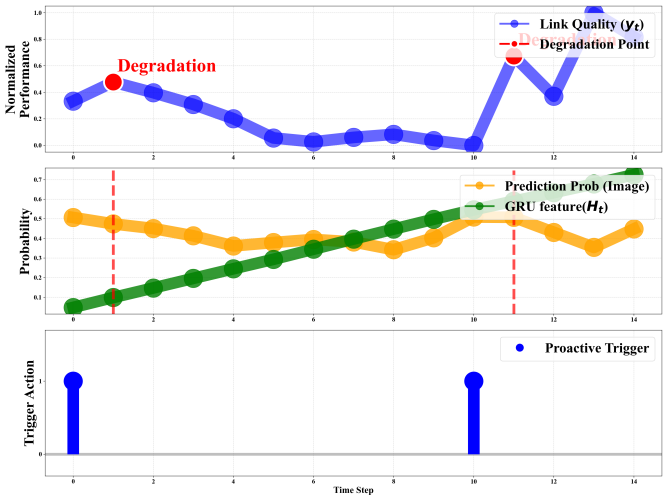

Fig. 1. The red dashed line marks the onset of link degradation, and the shaded region shows the degradation phase. Our policy triggers a switch before this point—indicated in blue—demonstrating successful proactive detection.

Table I. Fig. 1, 2 and Table II are the results of the proposed model (w/ vision) and the base model (w/o vision) using the 50 validation episode subset of the DeepSense 6G dataset.

The model we proposed had an average reward of -18.48, which was higher than the -28.90 of the non-image model, suggesting that visual features help the agent to infer future link states more accurately. Fig. 1 shows the characteristics of the image-based model. In the image-based model, the average re-alignment interval ($\overline{\tau}$) was 2.17 steps, which means that the policy responded more promptly to changes. The PTSR was 0.448 and the average lead time was 2.292 steps, confirming that when using images, the model can frequently anticipate future link deterioration and adjust the beam in advance. The image-based policy performed preemptive switching by detecting changes in stability embeddings (H_t) and probability distributions from images before the link quality decline.

Fig. 2 shows the characteristics of the non-image model. In the ablation model, the average re-alignment interval ($\overline{\tau}$) was exceptionally high, indicating that the agent failed to initiate switching even when the link quality was severely degraded. PTSR was 0.062 and the MDR was 0.920, showing that the non-image policy could not switch before degradation occurred in most cases. Visual information enabled proactive and early detection of degradation, while the non-image model

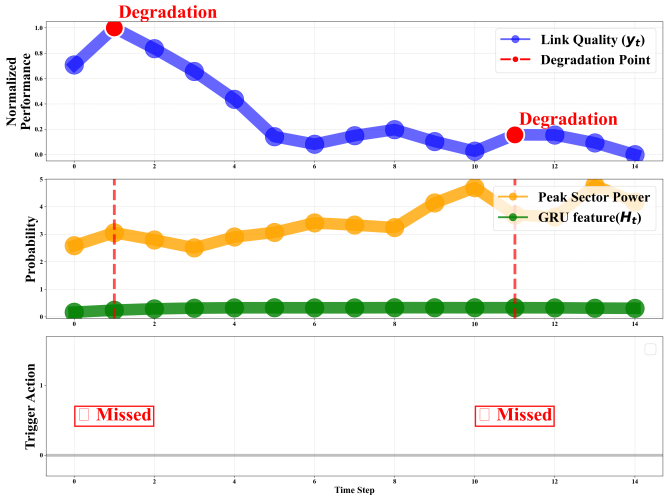

Fig. 2. Non-image model result on the same episode. The policy fails to trigger beam management before link degradation and reacts only after degradation begins, exhibiting purely reactive behavior without visual information.

failed to anticipate future degradation. These results show that visual information is important when predicting future degradation.

V. CONCLUSION

In this work, we proposed an DRL-based trigger policy that determines the optimal beam re-alignment timing using link variations and environmental uncertainty. The proposed model showed the performance of early detection of future degradation while reducing unnecessary switching when using state expression that includes visual features and power changes. In future studies, we plan to extend the policy to choose the best beam candidates that can maintain link stability for future intervals.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) Grants funded by Korean Government through Ministry of Science and ICT (MSIT) under Grant RS-2024-00409492 and Grant RS-2024-00346319.

REFERENCES

[1] M. B. Mollah, H. Wang, M. A. Karim and H. Fang, "Multi-Modality Sensing in mmWave Beamforming for Connected Vehicles Using Deep Learning," in IEEE Transactions on Cognitive Communications and Networking, 04 April 2025

[2] H. Wang, D. Yang and X. Xie, "A Deep-Reinforcement-Learning-Based Beam Prediction Scheme for Vision-Aided mmWave Wireless Communications," in IEEE Internet of Things Journal, vol. 12, no. 11, pp. 17869-17879, 1 June, 2025

[3] M. Giordani, M. Polese, A. Roy, D. Castor and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," in IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 173-196, Firstquarter 2019.,

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

[5] John Schulman, Sergey Levine, Pieter Abbeel, Michael I Jordan, and Philipp Moritz. Trust region policy optimization. In ICML, pages 1889–1897, 2015.