ENWAR 2.0: An Agentic Multimodal Wireless LLM Framework with Reasoning, Situation-Aware Explainability and Beam Tracking

Ahmad M. Nazar, *Graduate Student Member, IEEE*, Abdulkadir Celik, *Senior Member, IEEE*, Mohamed Y. Selim, *Senior Member, IEEE*, Asmaa Abdallah, *Member, IEEE*, Daji Qiao, *Senior Member, IEEE*, and Ahmed M. Eltawil, *Senior Member, IEEE*

***Abstract*—The evolution of next-generation wireless networks demands intelligent, adaptive, and explainable decision-making for robust communication in dynamic environments. This paper presents ENWAR 2.0, the first agentic large language model (LLM) framework integrating adaptive retrieval-augmented generation (RAG) and chain-of-thought (CoT) reasoning into situation-aware and explainable wireless network management. ENWAR 2.0 introduces two specialized agents: a transformer-fusion (TransFusion)-based beam prediction agent and an environment perception agent, both of which fuse multi-modal sensory inputs—including camera, LiDAR, radar, and GPS—from the DeepSense6G dataset. The beam prediction agent enables infrastructure-to-vehicle (I2V) target-in-the-loop beam tracking and real-time adaptation based on dynamic environmental conditions. In contrast, the environment perception agent provides situation-aware reasoning and justifications for beam decisions. Unlike its predecessor, ENWAR 1.0, which relied on static knowledge bases (KBs) and text-only LLMs, ENWAR 2.0 is designed for CoT reasoning, leverages LLaMa3.2-3B/LLaMa3.1-8B/LLaMa3.3-70B for text-generation, the multi-modal capabilities of LLaMa 3.2, and employs LlamaIndex for fine-grained, dynamic context retrieval, eliminating retrieval ambiguities and enhancing response relevance. Numerical results show that the beam prediction agent achieves up to 90.0% Top-3 accuracy at $t+3$, effectively predicting optimal beam selections three time steps ahead. Overall, ENWAR 2.0 achieves state-of-the-art performance, with up to 89.7%/83.5% interpretation/perception correctness, 81.6%/80.9% faithfulness, and 89.9%/88.2% relevancy. In comparison, the baseline pretrained LLaMa3 models without adaptive RAG achieves up to 80.3%/77.3% correctness, and the baseline without RAG performs significantly worse at 67.1%/64.8%. Additionally, ENWAR 2.0 reduces processing time by over 100% relative to the baseline, while its adaptive RAG improves performance by up to 13.7% compared to static RAG.**

I. Introduction

GENERATIVE artificial intelligence (AI) is poised to transform 6G and future wireless networks by enabling systems that can produce, adapt, and interpret vast amounts of data [2]. Central to this shift are large language models (LLMs), which are transformer-based architectures proficient in diverse tasks, from natural language understanding to decision support [3]. LLMs operate by learning statistical patterns from massive text corpora and can generate, summarize, or analyze language, code, or structured data in response to user prompts. They predict the following, most likely words or tokens in a sequence, enabling them to answer questions, provide explanations, and integrate diverse sources of knowledge. Their scalability and adaptability make them well-suited for next-generation networks' complex, dynamic environments, supporting zero-touch network and service management (ZSM) through rapid decision-making and resource optimization [4].

However, future wireless systems face challenges absent in legacy networks. Higher frequencies and massive antenna arrays demand rapid beam alignment, proactive blockage mitigation, and seamless handovers to maintain reliable links. Traditional cloud-centric or heuristic methods often fail to meet these stringent needs, particularly in dense urban settings.

A crucial enabler for addressing these challenges is multi-modal integrated sensing and communication (ISAC), which fuses inputs from sensors such as cameras, GPS, LiDAR, and radar to build fine-grained, real-time environment views. This fusion supports envisioning digital twins (DTs), near-real-time digital network state replicas that enable optimized decision-making and advance situational awareness and ZSM goals [5].

LLMs offer strong potential to bridge multi-modal sensing with wireless tasks, but conventional text-focused models often struggle with real-world data [6]. Simply extending them to multi-modal inputs does not guarantee domain-specific accuracy, as generic correlations and outdated knowledge can cause hallucinations or imprecise responses. Two primary strategies address these issues: *fine-tuning* and retrieval-augmented generation (RAG). While fine-tuning is resource-intensive, recent *parameter efficient fine-tuning (PEFT)* methods reduce adaptation overhead [7].

By contrast, RAG provides a cost-effective solution by integrating targeted knowledge bases (KBs) into the generation process. RAGs enhance contextual relevance, reduces token overhead, and improves inference efficiency. Retrieving only relevant context avoids repetitive preprocessing, lowers energy costs, and ensures concise, coherent responses without unnecessary token consumption [8].

A. M. Nazar, M. Y. Selim, and D. Qiao are with the Department of Electrical and Computer Engineering, Iowa State University of Science and Technology, Ames, IA, 50014, USA.

A. Celik, is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.

A. Abdallah, and A. M. Eltawil are with Computer, Electrical, and Mathematical Sciences & Engineering (CEMSE) Division at King Abdullah University of Science and Technology (KAUST), Thuwal, 23955 KSA.

A conference version of this work focusing on the beam prediction agent is submitted to IEEE ICC'26 [1].

With these motivations, we previously introduced ENWAR 1.0, an ENvironment-aWARe RAG-based multi-modal LLM framework [9]. While it successfully integrated domain-specific retrieval with raw sensory data to provide environment-aware insights, it remained primarily a retrieval system unable to perform downstream wireless tasks or offer reasoning-driven decision support.

This gap underscores the need for novel *agentic, reasoning-driven* LLM frameworks for 6G networks. Embedding specialized agents, such as analytical models, ML components, or optimization algorithms, into the LLM pipeline enables dynamic orchestration of complex wireless functions [10]. Reasoning is equally essential for trust and transparency: when paired with multi-modal sensing, agentic LLMs can interpret rich environmental data (e.g., traffic density, obstacle locations, mobility patterns), infer network states, and explain their actions in human-interpretable terms. Even as ZSM reduces human involvement, explainability is vital for fairness, accountability, and human oversight [11].

Motivated by these needs, we present ENWAR 2.0[1][2], an advanced framework that extends LLM-based intelligence with *agentic* capabilities and chain-of-thought (CoT) reasoning. It introduces two specialized agents: a *beam prediction agent*, enabling target-in-the-loop beam tracking between vehicles and the roadside unit (RSU), and an *environment perception agent*, which provides real-time reasoning and justifications for beam decisions. Together, these agents exemplify how ENWAR 2.0 bridges AI-driven automation with human interpretability, positioning it as a foundation for trustworthy, explainable AI in future 6G networks.

II. RELATED WORK AND MAIN CONTRIBUTIONS

This section provides an overview of related work within wireless networks and LLMs, and outlines the main contributions of our proposed approach. Additionally, for a complete tabular summary of related works, please refer to App. A[3].

A. Related Work

Recent research has explored diverse LLM-based frameworks for wireless communications, including RAG, question-answering (Q&A) training, instruction tuning, and multi-agent collaboration. RAG enhances LLM outputs by retrieving domain-specific knowledge from curated KBs, improving response accuracy [9], [12]–[14]. Q&A training fine-tunes LLMs with telecom-specific datasets to aid spectrum management and protocol understanding [15]. Instruction tuning further refines responses for telecom tasks but demands large curated datasets and significant computation [16], [17]. Multi-agent LLM systems have also emerged for task decomposition and decision-making in wireless networks [18]–[21]. However, most of these methods focus on query-based interactions rather than real-time decision-making.

Beyond static knowledge, recent work explores retrieval-augmented strategies for real-time, bandwidth-efficient communication in multi-agent and multi-modal vehicular networks [22], [23]. These approaches combine multi-modal data, semantic retrieval, and reinforcement learning to improve task efficiency, lower bandwidth use, and support user interactions.

Despite recent advancements, existing LLM frameworks for wireless networks face several key limitations. A major challenge is the lack of structured reasoning and explainability. Current LLMs function as black-box models, generating responses based on statistical correlations rather than explicit inference. This opacity is problematic in wireless decision-making, where network operators require interpretable justifications for AI-driven actions [11], [24], [25]. Without clear reasoning mechanisms, LLMs risk producing unreliable or suboptimal decisions in dynamic network environments.

Recent work has integrated LLMs and foundation models into wireless systems to boost adaptability and intelligence. For example, the Large Wireless Model, a transformer pre-trained on wireless channel data, generates contextualized embeddings that improve various communication and sensing tasks, especially in data-scarce scenarios [26]. In security, Gao et al. develop a BERT-based RF fingerprinting framework for 6G IoT networks, using self-supervised pretraining and distillation to robustly authenticate devices even under challenging conditions like multipath fading and Doppler shifts [27].

Another critical limitation lies in handling large-scale optimization problems. While LLMs are effective in pattern recognition, they struggle with NP-hard tasks such as beam alignment and resource allocation [12], [18]–[21]. Standard LLM approaches lack the mathematical rigor required for optimal solutions, necessitating hybrid models that combine LLM-driven insights with specialized optimization solvers. One such approach introduced an LLM-driven wireless communication paradigm that translates user natural language requests into structured queries and optimization tasks, enabling adaptive, user-centric system behavior through a prototype semantic communication system [28].

Most LLM architectures remain fundamentally text-based, limiting their ability to handle multi-modal wireless tasks that demand real-time analysis of sensory data like LiDAR, radar, and camera feeds [12], [20], [29], [30]. Although early multi-modal LLM efforts exist, practical implementations are still rare and lack support for real-time, data-driven network optimization. Additionally, RAG-based LLMs rely on static KBs that quickly become outdated, risking inaccurate recommendations in dynamic wireless environments [9]. Dynamically updating these KBs is crucial to ensure decisions reflect the latest standards and policies. These challenges underscore the need for new LLM frameworks with integrated reasoning, optimization, and multi-modal capabilities, motivating the development of ENWAR 2.0.

B. Main Contributions

To the best of the authors' knowledge, ENWAR 2.0 is the first agentic LLM framework to integrate RAG, CoT reasoning, and multi-modal ISAC for situation-aware wireless

[1] Enwar is a common name in Turkic and Arabic cultures, meaning enlightened, insightful, and intellectual; herein referring to a multi-modal LLM providing deep situational and contextual insights into the environment.

[2] ENWAR 2.0's code is available at https://github.com/anazar99/Enwar2.0

[3] Appendices are in the supplemental file provided with the submission.

network management. Unlike prior studies that either (i) applied generic LLMs with static knowledge bases or (ii) used multi-modal fusion solely for beam prediction, ENWAR 2.0 uniquely integrates adaptive RAG, CoT reasoning, and agentic multi-modal perception within a single framework. This combination enables real-time high-accuracy beam prediction and situation-aware justifications, going beyond accuracy-only baselines to provide interpretable, reasoning-driven wireless decision support.

Additionally, unlike ENWAR 1.0, which relied on single-modality LLMs and pre-embedded multi-modal data with static explanations in an offline KB, ENWAR 2.0 dynamically extracts relevant data and structures prompts in real-time, eliminating retrieval ambiguities and redundant explanations caused by overlapping textual data. It leverages models from the LLaMa 3 family—including LLaMa3.2-3B, 3.1-8B, and 3.3-70B—to enable flexible deployment across various resource budgets. LLaMa3.2-3B is a lightweight, efficient variant of LLaMa 3.1, while LLaMa3.3-70B offers enhanced capabilities and reasoning over earlier 70B versions. LLaMa3.2-Vision-11B, an inherently multi-modal model, also handles image-to-text conversion, supporting seamless visual input processing. These choices allow ENWAR 2.0 to balance performance, scalability, and computational efficiency without relying on external models for non-textual data. ENWAR 2.0's key improvements and main contributions are as follows:

- ✓ **Wireless Multi-Modal, Agentic LLM:** ENWAR 2.0 introduces specialized multi-modal agents (beam prediction and perception) natively integrated into the LLM pipeline for reasoning-driven decision support for situation-aware perception and grounded beam prediction.
- ✓ **CoT Reasoning for Explainability:** ENWAR 2.0 employs CoT reasoning to refine responses using real-time multi-modal sensory data iteratively. This structured reasoning process breaks down complex network scenarios into logical steps, ensuring well-justified and interpretable decisions without storing ground truths in the KB.
- ✓ **Efficient RAG with LlamaIndex:** Replacing ENWAR 1.0's LangChain[4]-based retrieval, Enwar 2.0 employs adaptive[5] for finer-grained indexing, chunking, and adaptive retrieval. This granularity optimizes context selection and ensures that only the most relevant information informs response generation, improving efficiency and contextual accuracy.
- ✓ **Adaptive KB for Dynamic Environments:** LlamaIndex also enables real-time KB updates, eliminating reliance on static, pre-curated knowledge. This change enables ENWAR 2.0 to remain responsive to evolving network conditions, ensuring up-to-date and contextually relevant decision-making.

Combining these advancements with specialized agents yields significant performance gains:

- ✓ **TransFusion-Based Beam Prediction Agent:** ENWAR 2.0 fully exploits multi-modal sensory data through a robust beam prediction agent that uses specialized feature extraction, pre-fusion techniques, and transformers. Numerical results show the agent achieves a Top-3 prediction accuracy of up to 90.0% and an average power loss (APL) of -0.009552 dB at $t+3$, enabling precise beam selection three steps ahead in dynamic environments.
- ✓ **Situation-Aware Grounding, Reasoning, and Explanation:** ENWAR 2.0 advances interpretability through its environment perception agent, delivering situation-aware reasoning and justifications for beam decisions. To quantify this, we introduce two novel key performance indicators (KPIs): (1) *Interpretation*, assessing how effectively ENWAR 2.0 explains beam choices based on multi-modal data and dynamic vehicular contexts, and (2) *Perception*, evaluating its analysis of environmental factors like potential object overlap and obstructions, and traffic density.

  We evaluate retrieval strategies by comparing adaptive RAG, static RAG, and full context injection without RAG across LLaMa3.2-3B, 3.1-8B, and 3.3-70B models. Adaptive RAG updates the KB with new sensory data for relevant retrieval, while static RAG relies on a fixed KB with full preemptive knowledge. Full context injection simultaneously processes all data in the prompt, leading to high computational load and reduced accuracy.

  Numerical results highlight the benefits of adaptive RAG: ENWAR 2.0 achieves up to 89.7%/83.5% interpretation/perception correctness, 81.6%/80.9% faithfulness, and 89.9%/88.2% relevancy, outperforming static and non-adaptive RAG by as much as 13.7%. Baseline pretrained LLaMa3 models without RAG achieve only up to 68.6% interpretation and 67.2% perception correctness. Full prompt injection without RAG suffers from higher latency and reduced interpretability, confirming the inefficiencies of excessive context injection. Additionally, ENWAR 2.0 improves efficiency, producing responses in 1.26 seconds compared to 2.67 seconds for the baseline without RAG.

C. Paper Organizations

The rest of the paper is organized as follows: Sec. III provides a high-level overview of ENWAR 2.0's framework; Sec. IV and Sec. V provides a breakdown components and functionality of the agents; Sec. VI and VII describe the offline and online pipeline of ENWAR 2.0, respectively; Sec. VIII evaluates the performance of ENWAR 2.0, and Sec. IX concludes the paper with a few remarks.

III. AN OVERVIEW OF ENWAR 2.0 FRAMEWORK

As illustrated in Fig. 1, this work considers a millimeter wave (mmWave) infrastructure-to-vehicle (I2V) scenario that consists of two main units: The former unit is a vehicle equipped with a uniform linear array (ULA) comprising M antennas and a GPS receiver. The latter unit is a RSU (i.e., base station (BS)) with a single receive antenna and multiple sensing capabilities, including a camera, radar, and LiDAR. Let $\mathbf{h}(t) \in \mathbb{C}^{M\times 1}$ denote the channel between the vehicle and the RSU at time t. The ULA is associated with a predefined

[4] https://python.langchain.com/docs/

[5] https://docs.llamaindex.ai/en/stable/

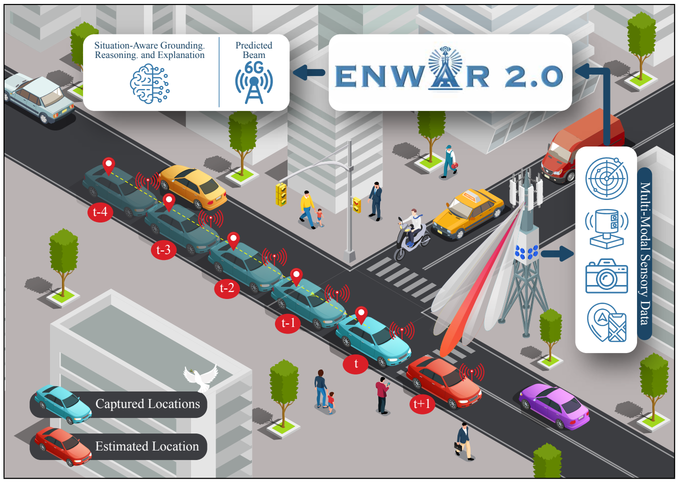


Fig. 1. Illustration of the system model.

beamforming codebook $\boldsymbol{\mathcal{F}} = \{\mathbf{f}_i\}_{i=1}^{Q}$, where $\mathbf{f}_i \in \mathbb{C}^{M\times 1}$ is the ith codeword (i.e., beam) and $Q = |\boldsymbol{\mathcal{F}}| = OM$ represents the codebook size after applying an oversampling factor O. Denoting $\imath(t)$ as the selected codeword/beam index for transmission of data symbol $x(t) \in \mathbb{C}$, then the received signal at time t is given by

$$y(t) = \mathbf{f}_{\imath(t)}^{H}\mathbf{h}(t)\,x(t) + n(t), \tag{1}$$

where $n(t) \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is the complex Gaussian noise. To achieve optimal connectivity, RSU aims to select the beamforming vector that maximizes the received power:

$$\overset{\star}{\imath}(t) = \underset{i\in[1,Q]}{\text{argmax}} |\mathbf{f}_i^{H}\mathbf{h}(t)|^2. \tag{2}$$

However, real-time computation of $\overset{\star}{\imath}(t)$ by exhaustive beam sweeping is impractical for large search spaces, especially in fast-changing vehicular environments with dynamic obstacles blocking the mmWave communications [31]. To address this challenge, the RSU functions as an LLM user, leveraging real-time multi-modal sensor measurements to construct prompts for the following wireless tasks:

1) *Target-in-the-Loop Beam Tracking:* The RSU continuously requests beam predictions, $\overset{\star}{\imath}(t+k)$, where k depends on the inference time, for a vehicle as it moves within the coverage area.
2) *Human-Interpretable Reasoning:* At certain intervals, a human-in-the-loop module may request situation-aware, grounded, and reasoned explanations for the selected beams.

ENWAR 2.0 achieves these objectives through two interlinked pipelines, as depicted in Fig. 2 and detailed below.

A. Offline Pipeline: KB Formation

The offline pipeline establishes a domain-specific KB by processing multi-modal data streams following below steps:

Ⓐ ***Agentic inference drawing*** employs two specialized agents that operate in parallel to convert multi-modal sensory data into semantically rich textual representations, simultaneously perceiving the wireless environment and predicting the best set of beams. Upon receiving new data, *raw data preprocessing* [c.f Sec. IV-A] takes place to ensure the integrity and consistency of timestamped multi-modal sensory data, which is vital for accurate beam prediction and reliable system inference. To that end, each modality is subjected to domain-specific preprocessing routines to maintain consistency across modalities and mitigate errors introduced by noisy measurements. After that, both agents initiate their respective processes in real-time, ensuring that perception and beam prediction occur concurrently for seamless, low-latency inference.

1) *The environment perception agent* [c.f. Sec. IV] receives data from diverse sensor modalities, each of which undergoes preprocessing suitable for its nature: LiDAR point clouds and radar signals are typically clustered to refine object detection and extract relevant features, while images are resized and masked to eliminate superfluous details before object detection models (e.g., You Only Look Once (YOLO)) identify various entities, e.g., vehicles, pedestrians, cyclers, etc. GPS data is also translated into textual form, providing spatial information that locates and contextualizes moving objects. The goal is to convert all sensor information into unified textual descriptions (e.g., "a vehicle stopped at a pedestrian crossing") that reflect the underlying environment in a form amenable to LLMs.
2) *The beam prediction agent* [c.f. Sec. V] employs a specialized transformer-fusion (TransFusion) model with cross-modality attention mechanisms to fuse information on object positions, trajectories, and physical obstacles, to provide a set of Top-k beam predictions. Each beam is characterized by key properties such as center angle and angular width. These properties are also converted into textual descriptions that are processed downstream.

🅑 ***Offline Knowledge Base Formation*** leverages LlamaIndex to arrange environmental and beam-related textual information for inclusion in the KB by following below steps:

- *Information combination* merges relevant descriptions of both environmental dynamics, e.g., detected objects, their velocities, and positions, and beam-related insights, e.g., beam center angles, and angular widths. We refer readers to App. B for an example of information combination.
- *Data chunking* segments these consolidated text into contextually coherent, and equal-sized chunks that conform to tokenization constraints. For instance, radar or LiDAR data can be segmented by object clusters, while GPS logs can be partitioned based on time intervals.
- *Embedding* vectorizes each data chunk using a general text embeddings (GTE) model to capture semantic relations. The resulting embeddings allow different modalities (e.g., radar- or camera-based descriptions) to be mapped into a unified semantic space, enabling ENWAR 2.0 to effectively cross-reference data from heterogeneous sources. Strict token and embedding length alignment is maintained to ensure that the multi-modal content remains consistent and interpretable within LLMs.

Next, LlamaIndex stores these vectorized representations in

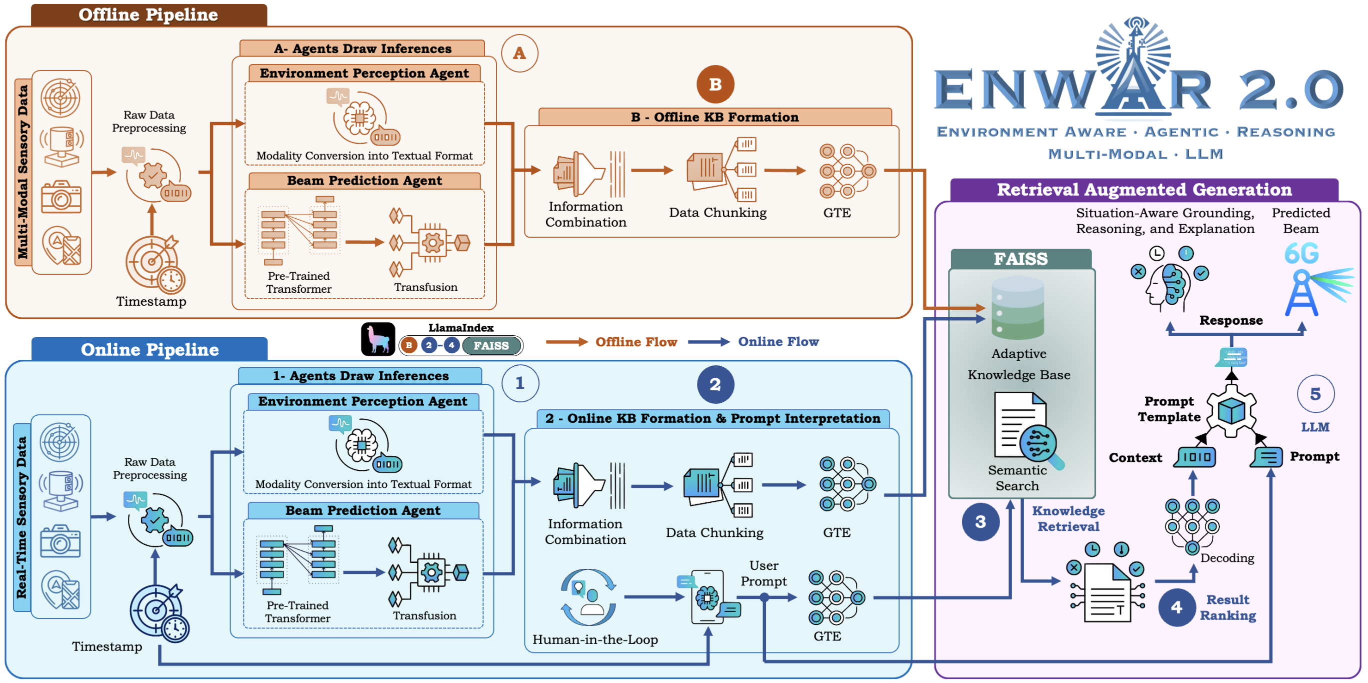


Fig. 2. Two primary pipelines of ENWAR 2.0 workflow: i) offline pipeline is responsible for KB generation involving steps Ⓐ-Ⓑ, and ii) online pipeline incorporates real-time process of response generation, comprising of steps ①-⑤. ENWAR 2.0 utilizes LlamaIndex for steps 🅑 , ❷ , ❸ , and ❹ .

a domain-specific KB by employing similarity search frameworks such as Facebook AI Similarity Search (FAISS) [32], which facilitate efficient indexing and retrieval of embeddings, even in large-scale or high-dimensional datasets.

B. Online Pipeline: Retrieval Augmented Generation

To retrieve contextually relevant information, update KB, and guide LLM responses, the *Online Pipeline* leverages the RAG in real-time following the below five steps, of which steps ❷ -❹ are implemented by using LlamaIndex.
① *Agentic inference drawing* mirrors the procedures described in step-Ⓐ of the *offline pipeline* but exploits the real-time multi-modal sensory data.
❷ *Online KB formation and prompt interpretation* follows the footprints of step-🅑 to update KB with the current embeddings. Unlike the offline pipeline, this step involves prompt interpretation. In addition to automated beam tracking request coming from ZSM, ENWAR 2.0 also allows human-in-the-loop interventions, which may occasionally demand explanations based on a flagging mechanism. As exemplified in App. C, the time-stamped prompt is vectorized so that it can be matched against the KB, which is explained next.
❸ *Semantic search* exploits the vectorized prompt to query the FAISS-indexed repository, generating a set of relevant text chunks that provide historical or contextual information. This retrieval process is powered by semantic similarity computations that identify potentially valuable correlations among environmental factors, past beam performance, and user-specific objectives. By leveraging the embeddings available in the repository, ENWAR 2.0 rapidly locates the most suitable knowledge segments for the ongoing scenario.
❹ *Result Ranking* filters and prioritizes the top-p percentile (e.g., $p = 95$) of retrieved chunks to ensure only the most contextually relevant information is used. This improves response coherence, reduces unnecessary token consumption, and preserves high-priority data even when nearing the LLM's context window limit. The selected vectorized results are then decoded into text to form a context, as illustrated in App. B.
⑤ *Response generation* begins with constructing a final prompt template [c.f. App. D], which includes a predefined task description, retrieved context from step-❹ [c.f.App. B], and the user query [c.f. App. C]. The LLM then generates responses by incorporating environmental observations, beam properties, and historical insights. Using top-p sampling or similar techniques, it balances diversity and precision to ensure responses are contextually rich and non-repetitive.

At this stage, ENWAR 2.0 selects the most suitable beams and explains its reasoning. For instance, it can justify decisions based on angular alignment, beam power levels, or real-time sensor updates affecting mobility. By fusing multi-modal data with RAG, ENWAR 2.0 enhances situational awareness and supports informed decision-making. Further details on the RAG framework are provided in Sec. VII.

IV. Environment Perception Agent

This section presents the design and processes involved in ENWAR 2.0's environment perception agent. By leveraging real-world sensor data from the DeepSense6G dataset [33], [34], the agent constructs a holistic understanding of the environment to facilitate robust beam prediction and situation-aware reasoning. Specifically, we utilize Scenarios 31–34, which provide synchronized data streams, including GPS coordinates for a vehicle and a RSU in communication, camera frames, LiDAR point clouds, and radar scans. In the rest of this section, we first delineate the data preprocessing steps, then explain how environmental perceptions are inferred from various sensor modalities.

A. Data Preprocessing

DeepSense6G offers diverse sensors that capture the environment from multiple perspectives. Ensuring the integrity and consistency of these data streams is vital for accurate beam prediction and reliable system inference. To that end, each modality is subjected to domain-specific preprocessing routines. Each data source undergoes careful timestamping and preprocessing procedures before being incorporated into ENWAR 2.0 to maintain consistency across modalities and mitigate errors introduced by noisy measurements.

1) Image Preprocessing: All camera frames are resized to a resolution of 256×256, and pixel values are normalized to stabilize subsequent feature extraction.

2) GPS Preprocessing: GPS data is calibrated to account for vertical, horizontal, and position dilution of precision (DoP) measurements, correcting biases that may arise from satellite geometry or signal noise. After calibration, motion-related features such as displacement, speed, directional angle, and acceleration are extracted. Additionally, angular velocity and curvature information is derived to capture both turning and linear motion patterns.

3) LiDAR Preprocessing: To handle the high density of LiDAR point clouds, a voxel-grid downsampling procedure is applied to reduce the volume of points while preserving spatial structure [35]. A random sample consensus (RANSAC) plane-fitting algorithm then identifies and removes ground-plane points, leaving only points associated with obstacles or objects of interest [35]. Outliers are removed by statistical analysis, as described in Sec. IV-D. The remaining 3D points are normalized by centering them around the origin and scaling them to fit within a unit cube.

4) Radar Preprocessing: Radar scans typically include clutter and spurious reflections. To address this, a Gaussian-based noise filter enhances valid targets by removing low-intensity points. Outliers are eliminated via statistical analysis, as detailed in Sec. IV-E, to isolate reliable reflections from moving and static objects. Finally, the cleaned radar data is normalized through max-min normalization.

B. Camera Perception

1) Image-Based Object Detection: To extract high-level semantics from images, ENWAR 2.0 employs YOLO, a real-time object detection system that produces bounding boxes, confidence scores, and class labels (e.g., vehicles or pedestrians). The confidence threshold is set to 0.5, discarding bounding boxes with lower scores to reduce false detections. Each bounding box is represented by coordinates, (x, y, w, h), where x and y are the bounding box's center coordinates, and w and h are the bounding box's height and width. This representation is then assigned a class label. These outputs are converted into textual descriptions that are integrated into the RAG pipeline. This transformation ensures ENWAR 2.0 provides comprehensive and visually context-rich responses.

2) Image-to-Text for Enhanced Explainability: Beyond bounding boxes, ENWAR 2.0 augments scene understanding with descriptive text generated from images. Specifically, the first image in each five-image sequence is passed to a multi-modal LLM optimized for vision-language tasks. Using the first image ensures that the description reflects the environment at the start of the sampling period, which is generally stable within the sample intervals. An instructional prompt guides the vision model to provide contextually relevant scene descriptions (e.g., *"a white sedan waiting at a crosswalk"*), thereby improving the interpretability of the sensed environment. Incorporating these descriptive captions into the system's KB furnishes valuable contextual cues that can clarify beam decisions and environment-aware reasoning.

C. GPS Perception

Position data from the vehicle (Unit 2) and the RSU (Unit 1) is used to calculate their relative distance and bearing angle. Latitude and longitude measurements are calibrated before being processed by a long short-term memory (LSTM)-based GPS network (detailed in Section V). This network predicts future trajectories, capturing projected distance and bearing over time. By converting current and forecasted GPS data into textual descriptions (e.g., *"vehicle at 35.1236, -80.9421 at a bearing of 43.31° with speed 10km/h"*), ENWAR 2.0 easily incorporates positional information into the RAG framework, and facilitates alignment with other sensor modalities and supports high-level situational inference.

D. LiDAR Perception

Clustering in LiDAR point clouds is done via density-based spatial clustering of applications with noise (DBSCAN), which groups 3D points into meaningful objects and filters out noise. Algorithmic parameters, such as $\epsilon_{\text{lidar}} = 0.75$ and $\texttt{min\_samples}_{\text{lidar}} = 5$, determine the maximum separation between points within a cluster and the minimum cluster size, respectively. These parameters were selected by plotting a k-distance graph, where the distance to the k-th nearest neighbor for each data point is calculated with $k = \texttt{min\_samples}_{\text{lidar}}$, and the point of maximum curvature in this plot represents a good value for ϵ_{lidar}. Each identified cluster is translated into a bounding box with estimated dimensions and location. These bounding box coordinates, along with cluster metadata (e.g., point density), are compiled into textual summaries (for instance, *"large object spanning 2.3m in length at 45° to the LiDAR"*). In addition to providing environmental context, this structured textual data helps ENWAR 2.0 refine its understanding of the physical layout surrounding the RSU, supporting more accurate beam direction decisions.

E. Radar Perception

Radar perception begins by filtering raw scans with an intensity threshold to eliminate weak reflections. Detected points are then grouped into clusters using DBSCAN, configured through $\epsilon_{\text{radar}} = 2.5$ and $\texttt{min\_samples}_{\text{radar}} = 2$. These parameters were selected similarly to the parameter selection

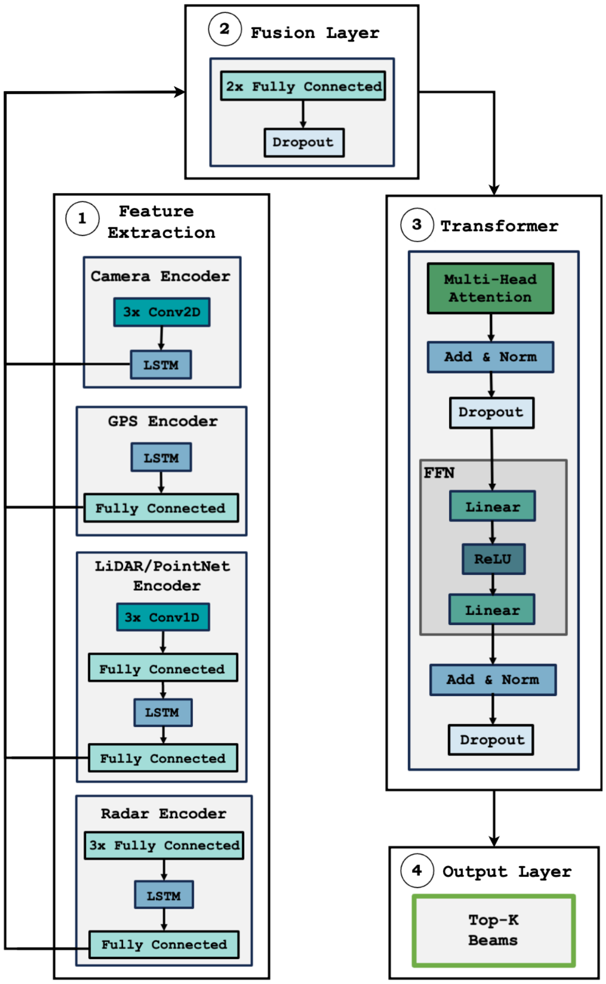


Fig. 3. TransFusion beam prediction model architecture.

in LiDAR Perception. Each cluster's average range, velocity, and angular span are calculated to characterize the detected object in radar coordinates. Transforming these radar clusters into text (e.g., *"cluster at range 15m, velocity 2m/s"*) enhances ENWAR 2.0's capacity to integrate spatiotemporal data across modalities. Correlating radar measurements with LiDAR or camera detections further enriches multi-modal awareness and can reveal inconsistencies, such as an undetected object in other modalities but present in radar scans.

V. BEAM PREDICTION AGENT

ENWAR 2.0 employs the TransFusion architecture to predict optimal beam selections by integrating multi-modal sensory inputs. This model enhances accuracy and adaptability in dynamic wireless environments by processing diverse data sources through a structured pipeline. As illustrated in Fig. 3, the agent follows these steps: ① specialized encoders extract high-dimensional features from preprocessed multi-modal data; ② extracted features are then fused into a unified representation, enabling cross-modality interaction; ③ the fused data passes through a transformer block, leveraging multi-head self-attention to capture complex dependencies; and ④ the output layer assigns scores to potential beam directions. This pipeline ensures robust and reliable beam predictions. To analyze the contribution of each modality, these steps are repeated for different modality combinations. The following sections provide a detailed explanation of each stage.

① **Feature Extraction:** The first step in the model is to extract features from the multi-modal inputs. The extracted features represent high-dimensional, modality-specific insights essential for understanding the environment and predicting optimal beams.

Image Features: The `Camera Encoder` processes image sequences through convolutional layers and an LSTM network. The preprocessed images are transformed into a tensor and reshaped, then passed through three consecutive convolutional layers with a ReLU activation function following each layer. After the final convolutional layer, the output tensor is reshaped into a flattened vector sequence. These flattened features are then fed into an LSTM network layer, which processes the temporal dependencies across the sequence. The LSTM generates 128 hidden states for each time step, and the hidden state from the last time step is used as the final encoded representation. This representation captures the spatial features from the convolutional layers and the temporal dependencies modeled by the LSTM network.

GPS Features: The `GPS Encoder` takes preprocessed GPS data and encodes temporal dependencies using an LSTM network. Preprocessed and derived GPS features are then concatenated and normalized using a feature scaler trained on the entire dataset to ensure consistency and minimize variability. The resulting normalized feature vector is then passed into a two-layer LSTM network, which processes the input sequence to capture temporal dependencies. At each time step, the LSTM generates 128 hidden states, with the final hidden state representing the learned temporal features of the sequence. This hidden state is passed through a fully connected layer to produce the final encoded GPS representation.

LiDAR Features: The `PointNet Encoder` extracts spatial and temporal features from point cloud data, leveraging both convolutional layers and recurrent processing. At each time step t in the sequence, the corresponding point cloud frame is processed through three 1D convolutional layers with kernel size 1 and ReLU activations. Max-pooling is then applied across all points in the cloud to aggregate features into a fixed-dimensional representation. The sequence of spatially encoded frames is passed through an LSTM network with 128 hidden states to model temporal dependencies across consecutive LiDAR frames. The last hidden state of the LSTM captures both spatial and motion-aware features, forming the final encoded representation of the LiDAR sequence. This integration of CNN-based feature extraction with LSTM-based temporal modeling enables the network to learn meaningful spatiotemporal patterns from the evolving point cloud data.

Radar Features: The `Radar Encoder` processes radar data by combining fully connected layers for spatial feature extraction with an LSTM network for capturing temporal dependencies similar to the PointNet architecture [36]. At each time step t in the sequence, the radar input tensor is flattened and passed through three fully connected layers with ReLU activations to extract frame-wise features. the sequence of extracted radar features is passed through an LSTM network with 128 hidden states, which models temporal dependencies across consecutive radar frames. The last hidden state of the LSTM serves as the final encoded representation, capturing both the spatial structure of individual frames and the temporal evolution of radar reflections over time.

② **Early Fusion:** An important design choice in the TransFusion architecture is to decouple the initial feature extraction from the cross-modal reasoning by performing feature fusion prior to transformer processing. Each sensor modality produces features with distinct statistical properties, dimensionalities, and temporal characteristics. Directly feeding these heterogeneous outputs into a transformer would require the model to simultaneously handle feature extraction and cross-modal alignment, increasing the learning burden and training instability. Instead, TransFusion adopts a staged strategy: modality-specific encoders first transform raw sensor inputs into structured, modality-aligned feature embeddings. These embeddings are then fused into a single, unified representation before being processed by the transformer.

By presenting the transformer with a fused, high-dimensional feature vector, the architecture allows the transformer to focus solely on modeling complex dependencies and interactions across modalities without being overwhelmed by low-level modality-specific noise or structural mismatches [1]. This pre-fusion strategy reduces computational complexity, improves model convergence, and ensures that cross-modal relationships are fully captured for more accurate and stable beam prediction performance.

The Fusion Layer concatenates the outputs from all modality-specific encoders into a unified feature vector. This fused vector is then processed through two fully connected layers with dropout regularization to promote generalization. The first fully connected layer applies a linear projection followed by a ReLU activation function, capturing higher-level feature interactions. Dropout is applied to prevent overfitting, and the output is further refined by a second fully connected layer with another ReLU activation. The resulting fused representation effectively integrates spatial, semantic, temporal, and motion information from all modalities, providing a compact and rich input to the downstream Transformer Block for sequence-level reasoning and final beam prediction.

③ **Transformer Block:** The Transformer Block uses fused feature representation to model cross-modal interactions and capture long-range dependencies across the input sequence. The fused input first passes through a multi-head self-attention mechanism, which allows the model to dynamically weigh the importance of different features across modalities and time steps. The transformer comprises four attention heads, each with 512 hidden dimensions, supported by two-layer normalization stages and two dropout layers (dropout rate = 0.1) to enhance generalization and stabilize training. The outputs of the attention heads are aggregated and combined with the original input via a residual connection, followed by layer normalization and dropout to maintain training stability and prevent overfitting.

Following attention, the output is processed through a position-wise feed-forward network consisting of two fully connected layers separated by a ReLU activation. The feed-forward output is then combined with the attention output via residual connection, followed by another round of normalization and dropout. This hierarchical processing allows the Transformer Block to capture complex cross-modal dependencies, align modality-specific features, and generate a transfused representation that encodes joint spatial, temporal, and semantic relationships for accurate beam prediction.

④ **Output Layer:** The Output Layer services as the final component of the model, responsible for predicting the beam scores across all possible beam directions. The input to the output layer is the processed transfused feature representation obtained from the Transformer Block, passed through a fully connected layer that maps the input vector to a score for each of the Q beam indices. The output assigns a score for each beam, where higher scores indicate higher beam suitability. This scoring mechanism enables the system to identify and prioritize the optimal beam direction based on the fused multi-modal sensory data and the modeled interactions within the transformer block.

VI. OFFLINE PIPELINE: KNOWLEDGE-BASE GENERATION

With data preprocessing and agents delineated in the previous two sections, this section focuses on the remaining components of the offline pipeline and explains how KBs are generated.

A. Information Combination

A core innovation of ENWAR 2.0 lies in its ability to seamlessly integrate processed multi-modal information from both the environment and beam prediction agent into a unified textual format, enabling effective storage, retrieval, and reasoning within the KB. A detailed example of a full KB entry, i.e. information combination, is shown in App. B. After pre-processing explained in Sec. IV, we populate the KB by meticulously identifying a total of 150 scenes/samples, each representing a comprehensive multi-modal snapshot of the environment with scenarios showing unobstructed communications (e.g., scenes in Fig. 4), a busy environment (e.g., scenes in Fig. 5), and obstructed communications (e.g., scenes in Fig. 6). These samples are composed of the following elements:

1) Timestamp: Each sample is assigned a unique identifier and a sequential index corresponding to its timestamp/tag, which ensures precise organization within the KB, preventing contextual collisions and ensuring accurate data retrieval.

2) Network Environment Description: A high-level textual summary of the communication devices of interest within the environment and the multi-modal input sampling period is appended at the beginning of each KB entry.

3) Camera Detections: Each sample consists of five consecutive images, processed using YOLO-based object detection to detect and classify objects such as vehicles, pedestrians, and obstacles. Detected bounding boxes, confidence scores, and class labels are converted into textual descriptions, capturing each detected object's spatial extent and semantic meaning.

4) Image-to-Text Descriptions: Scene-level descriptions generated by the LLaMA3.2-Vision model for the first image of each sequence. These natural language descriptions capture high-level visual insights, such as "a pedestrian waiting to cross the road" or "a vehicle merging into traffic," providing a semantic overview of the scene.

5) GPS Coordinates: Five consecutive GPS samples provide precise positional data for the receiver vehicle. Each timestamp is transformed into textual descriptions that include the vehicle's latitude, longitude, distance, and bearing relative to the RSU. These details serve as a fixed spatial reference for other modalities.

6) LiDAR Detections: Five consecutive point clouds are processed using DBSCAN to extract object clusters, which are then converted into textual data. These descriptions detail detected clusters' dimensions, spatial locations, and relationships, facilitating robust 3D spatial reasoning.

7) Radar Detections: Five consecutive radar scans are analyzed to extract object clusters, including their ranges, velocities, and angular properties. The processed radar data is then converted into textual descriptions, capturing temporal and spatial dynamics.

8) Beam Properties: Textual representations of the Top-k predicted beams, including attributes such as beam center angle, angular width, estimated beamwidth (calculated based on vehicle trajectory), and real-time received beam power. These descriptions enable detailed interpretability of beam predictions regarding the dynamic environment.

By transforming this rich multi-modal data into a unified textual format, ENWAR 2.0 provides a robust foundation for its RAG framework. The KB is the cornerstone for efficient retrieval, enabling contextual reasoning and dynamic decision-making in complex, wireless network environments.

B. Data Chunking

To efficiently handle the multi-modal data within the constraints of embedding models and LLMs, ENWAR 2.0 segments the transformed textual data into equal-sized and manageable chunks. This chunking process ensures that the data is consistently embedded into the GTE model. Consider a preprocessed and transformed textual dataset of size L. This dataset is divided into smaller chunks of size C. To maintain continuity and prevent loss of meaning between chunks, each chunk overlaps with the next by C_{overlap} characters. This overlapping ensures that the segments are connected, preserving the context and meaning across the chunks. For ENWAR 2.0's chunking process, the following parameters were selected: $C = 800$ and $C_{\text{overlap}} = \text{int}(C/10) = 80$. For example, if the original dataset is divided into chunks of 800 characters with an overlap of 80 characters, the first chunk would contain characters 1 to 800, the second chunk would contain characters 721 to 1600, and repeats for all chunks. This method ensures that important information is not lost between chunks, maintaining high fidelity and contextual accuracy across all modalities. This configuration balances preserving contextual continuity and keeping retrieval efficient, as smaller chunks risk fragmentation while larger chunks increase redundancy and slow down indexing.

C. Embedding

Once chunked, the textual data is processed using a GTE model, which converts each chunk into dense, high-dimensional vector embeddings, encoding the semantic meaning of the text in a machine-readable format. The embedding model first tokenizes the input text into subword units, which are then mapped to numerical representations. These token representations are passed through a deep neural network, typically a transformer-based architecture, which captures contextual relationships and semantic meaning. The final output is a fixed-size vector embedding that encapsulates the most salient information from the text.

Alignment across modalities is achieved during the tokenization and embedding stages, where padding ensures uniform tokenized chunk lengths and consistent semantic representation. For simplicity, final embedded vector with padding is matched to the GTE model's maximum token size. This uniformity allows LLMs to handle redundancy, conflicts, and synergies effectively. These embeddings are then stored in the KB facilitated by LLaMaIndex for retrieval. This integrated approach facilitates robust, conflict-free sensing and a unified understanding of the environment. By analyzing transformed and integrated data, the LLM resolves discrepancies, identifies shared patterns, and utilizes the strengths of each modality

VII. ONLINE PIPELINE: ADAPTIVE KNOWLEDGE-BASE FORMATION AND RETRIEVAL AUGMENTED GENERATION

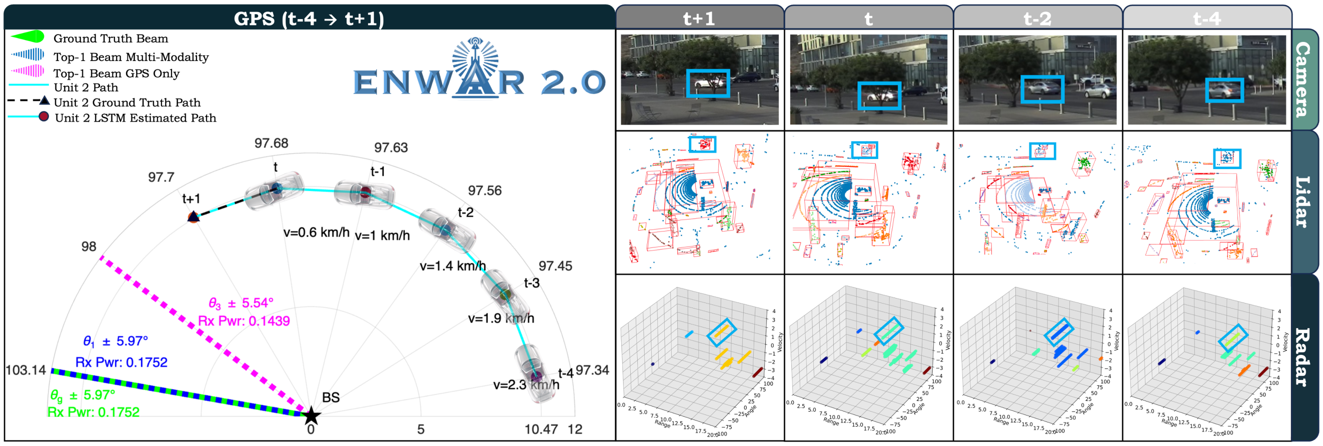
This section outlines the adaptive KB formation and RAG processes introduced in Sec. III-B, providing relevant examples to illustrate their implementation.

A. Adaptive Knowledge-Base Formation

In order to update the KB, ENWAR 2.0's online pipeline leverages LlamaIndex that follows the same offline KB generation steps on the current multi-modal sensory data and corresponding beam predictions. Notice that KB is updated independent of user prompts, ensuring that ENWAR 2.0 continuously refines its understanding of the environment. Additionally, the adaptive KB provides an inherent historical record, capturing network behavior over time and enabling the LLM to refine its responses dynamically. This historical awareness, either through explicit KB entries or adaptive prompting, allows ENWAR 2.0 to maintain response consistency and adjust its level of numerical detail based on prior interactions, improving both interpretability and reasoning.

B. Prompt Processing

When a user prompt is received, it is vectorized and embedded, then used in the semantic search process, as explained in the next section, to compare against entries in the KB. While user prompts can reference any stored KB entry, our deployment, by default, assumes that the user is inquiring about the most recent entry, applying a recency tag to streamline retrieval. Since the KB updates dynamically with each new multi-modal input, user prompts inherently align with the



Prompt 1

Based on the Beam Prediction and Multi-Modal Sensor Visualization, select and evaluate the best beam's alignment with Unit 2's trajectory, focusing on GPS location, speed, bearing, and movement. Compare top-k beams for differences in received power and coverage. Determine suitability based on environmental factors (e.g., camera, lidar, radar) and their influence. Assess if the top-k beams are within ground truth bounds and align well with the vehicle's trajectory and bearing.

ENWAR 2.0

Environmental Context Analysis:

- **LiDAR Data Analysis:** The detected objects clusters may create reflective surfaces that affect beam propagation. High-rise buildings and narrow bounding boxes suggest a potentially cluttered urban environment, which may result in multipath propagation and shadowing.
- **Radar Data Analysis:** Clusters show variations in range, velocity, and density, with moving and static objects that can reflect or scatter RF signals, affecting the beam's power and alignment with Unit 2.
- **Camera Data Analysis:** The camera detected multiple vehicles at varying positions. Objects like trees and streetlights may block or diffract signals, impacting beamwidth and coverage.
- **GPS Data Analysis:** The transmitter vehicle's consistent movement in a specific bearing suggests the beam needs to align with this trajectory to ensure strong signal propagation and minimal interference. UNIT 2's movement is gradually slowing down (from 2.307 km/h to 0.6 km/h), indicating it is approaching an intersection, stop sign/traffic light, as confirmed by the scene description.

Unit 2's trajectory and estimated next position indicate a movement consistent with a bearing of ~97.68°. Both the Top-1 and Top-2 Beams fall within this trajectory and align closely with the vehicle's estimated location. The Top-3 Beam, however, with its center angle at 98°, is marginally offset but still within a tolerable range.

Predicted Beam Properties and Analysis:

- **Top-1 Beam:** Index: 21, Center Angle: 103.14°, Width: 1.41862 meters, Coverage: ±5.97°, Power: 0.1753
- **Top-2 Beam:** Index: 22, Center Angle: 103.14°, Width: 1.31729 meters, Coverage: ±5.55°, Power: 0.1722
- **Top-3 Beam:** Index: 26, Center Angle: 98.00°, Width: 1.31729 meters, Coverage: ±5.55°, Power: 0.1439
- Both Top-1 and Top-2 beams have identical center angles, aligning well with UNIT 2's bearing and the estimated trajectory. Their beam width and coverage ensure adequate reception, the captured power is highest for these beams, suggesting optimal signal reception. For optimal beam steering, selecting the Top-1 Beam, since it has the higher received power, is most suitable for this scenario.
- The network can rely on a consistent beam since UNIT 2 is not making rapid direction changes, and the detections from the multi-modal data indicate clear Line-of-Sight Communications
- The Top-3 Beam deviates from the ground truth center angle by about 5° and has reduced power. This deviation could make it less effective for maintaining a reliable connection.
- ***Ground Truth Beam Properties:*** *Index: 21, Center Angle: 103.14°, Width: 1.41862 meters, Coverage: ±5.97°, Power: 0.1753*
- *The Top-1 Beam is perfectly aligned with the ground truth and receiver trajectory, making it the most suitable choice.*

Prompt 2

Create a table summarizing key features for this sample while including the following: object types and positions from LiDAR, Camera, and Radar (e.g., bounding boxes, clusters), environmental bounds (e.g., point cloud dimensions), trajectory data (e.g., GPS locations, bearings, speed), and beam properties (e.g., index, center angle, power)

ENWAR 2.0

Detection Properties	**Key Features**
Camera	Detected 9 cars at various positions with bounding box widths and heights ranging from (14x12) to (136x50).
LiDAR	Detected 6 objects with bounding box centers (e.g., (0.84, -0.13, 0.01)) and extents (e.g., (2.73, 2.00, 0.69)).
Point Cloud	Min bounds (-79.43, -94.78, -2.04), Max bounds (88.63, 80.17, 13.88), Dimensions (168.06 x 174.95 x 15.92).
Radar	Detected clusters with ranges (5.07m to 24.0m), angles (-75° to 72.66°), velocities (0 to 2.59 km/h), and counts (2 to 122).
GPS	UNIT 2 locations: [(33.42412937, -111.9349708) to (33.4241327, -111.93497087)], speeds: [2.307 km/h to 0.6 km/h], bearings: [97.34° to 97.68°].
Beam	Top-1 Beam: Index 21, Center Angle 103.14°, Width 1.41862m, Power 0.1753.
	Top-2 Beam: Index 22, Center Angle 103.14°, Width 1.31729m, Power 0.1722.
	Top-3 Beam: Index 26, Center Angle 98.00°, Width 1.31729m, Power 0.1439.
	Ground Truth Beam: Index 21, Center Angle 103.14°, Width 1.41862m, Power 0.1753.

Fig. 4. Visualization of ENWAR 2.0's best case scenario response with perfect top-1 beam alignment, situation-aware reasoning and explanation, and a tabular summary of situation-aware environment sensing, perception, reasoning, and prediction.

latest contextual information unless explicitly specified otherwise. A timestamp associated with each KB entry ensures that ENWAR 2.0 maintains relevance in its responses, accurately referencing stored scenarios while allowing for flexibility in historical data retrieval as inferred from the prompt itself (an example is seen in App. C).

C. Semantic Search

Once the user prompt is vectorized and embedded, ENWAR 2.0 conducts a semantic search to retrieve the most relevant information from its domain-specific KB within the RAG framework. This search targets contexts linked to the prompt's timestamp stored in FAISS. By computing semantic similarity between the prompt and stored embeddings, ENWAR 2.0 efficiently identifies matching entries. FAISS accelerates vector similarity searches through hierarchical indexing and clustering, maintaining low latency even as data scales. As detailed in the following subsection, the top-ranked, contextually aligned results are then selected for further processing.

D. Result Ranking

To ensure the relevance and quality of retrieved data, ENWAR 2.0 employs a refined result-ranking mechanism that prioritizes results based on their section headers and semantic alignment with the user prompt. This approach guarantees that only the most contextually relevant portions of the KB are used in generating responses, improving their clarity and precision. The ranking process employs top-p percentile relevance filtering, where $p = 95$ is consistently applied. In the context of RAG, this means retaining only the top 95% most relevant results based on their semantic similarity scores with the user prompt. FAISS's hierarchical indexing optimizes the ranking process by sorting and scoring entries based on semantic similarity. These similarity scores are calculated using cosine similarity between the embedded vector representation of the prompt and the entries stored in the KB. Entries in the bottom 5% of similarity scores are excluded, which helps minimize noise and ensures a focused and high-quality response.

This top-p ranking mechanism provides several advantages. By excluding less relevant entries, the system avoids diluting the response with marginally related or irrelevant content, thereby preserving the contextual alignment between the user query and the retrieved data. This approach also scales efficiently for large KBs containing millions of embeddings, narrowing the search space to only the most relevant subset. As a result, the system can maintain low computational overhead without sacrificing precision. Furthermore, focusing on high-relevance data enhances interpretability, ensuring the final response remains clear, concise, and actionable, facilitating informed decision-making and better user understanding.

E. Response Generation

In ENWAR 2.0, once the top-ranked results (i.e., embeddings) from the semantic search are retrieved, they provide the critical contextual foundation for the LLM to generate accurate and contextually relevant responses. These retrieved results, aligned with the user's prompt, enable the LLM to construct coherent and informed outputs tailored to specific scenarios.

At this point, the next step is crafting a final prompt template, as illustrated in step-⑤ of Fig. 2. To this aim, retrieved embeddings are first transformed into textual format using a decoder, yielding the context [c.f. App. B] necessary for the prompt template. The prompt template also includes a predefined task description and instruction with the user prompt [cf. App. C]. An example of the final version of a template is shown in App. D]. This template guides the LLM to accurately analyze the wireless environment and generate network insights and reasoning based on the relevant extracted information.

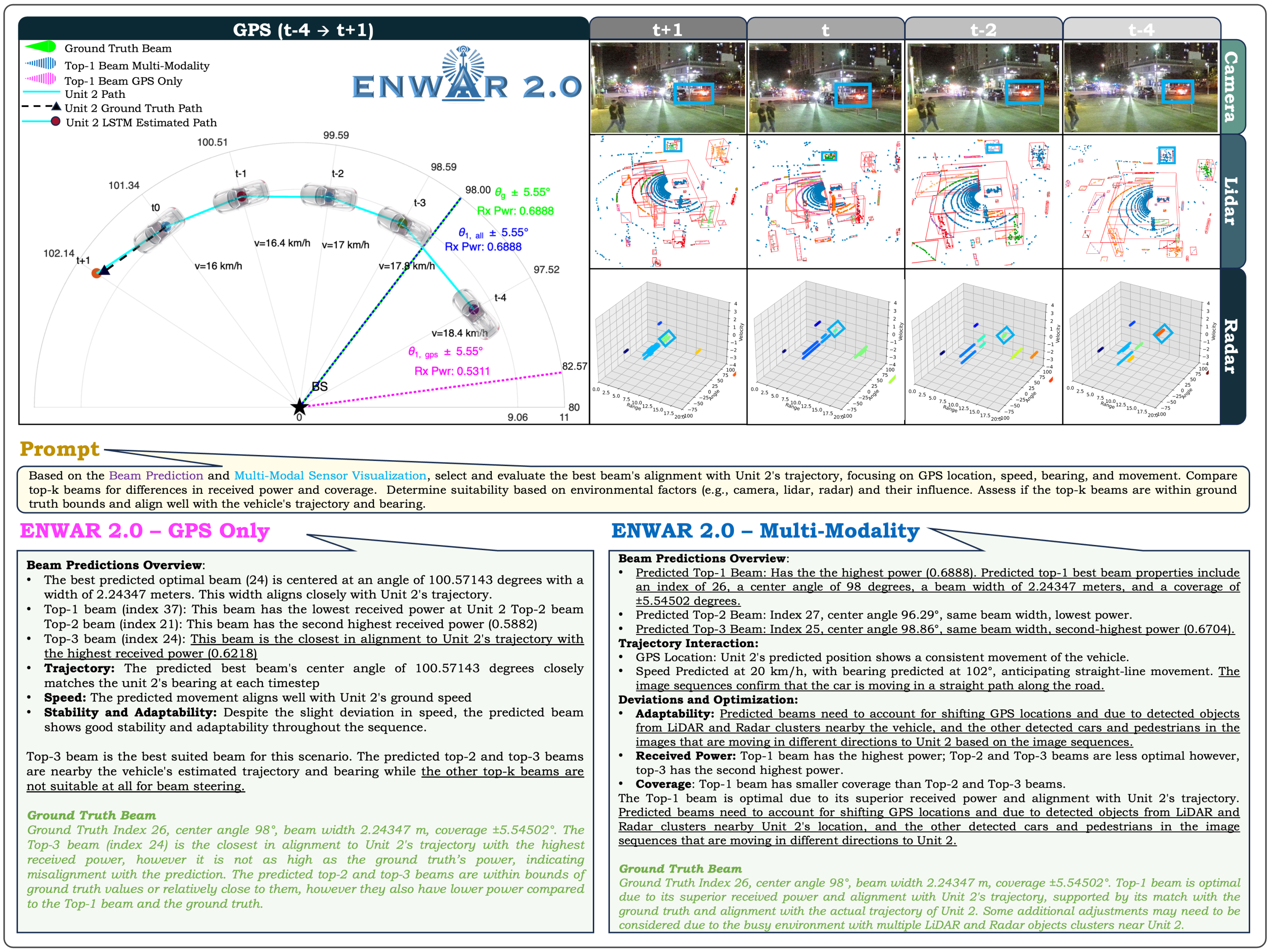
To ensure the LLM effectively utilizes retrieved KB entries, ENWAR 2.0 uses a structured approach that ties response generation to predefined instructional tasks. The retrieval mechanism supplies the most relevant context. In parallel, the LLM follows explicit reasoning patterns and a response template that maps retrieved data to explain beam selection, trajectory alignment, and environmental conditions. The user prompt, retrieved entries, and task instructions collectively shape responses, ensuring interpretability and relevancy. This predefined task design was refined iteratively to deliver consistent, reliable outputs without depending on explicit ground truths, as demonstrated by our evaluation metrics and response examples.

The LLM processes a prompt template that includes the user prompt and the most relevant retrieved KB contexts. Chunking the combined prompt at this stage is unnecessary, given the LLM's larger context window compared to the GTE model. The input flows through the LLM's internal architecture, where its generative abilities synthesize coherent, contextually relevant responses. Leveraging deep language modeling and pattern recognition, the LLM draws logical connections between retrieved data and the user prompt, enabling it to infer meaningful insights beyond direct retrieval. This capability allows the LLM to generate detailed environmental representations, identifying entities like vehicles and pedestrians, localizing them, and explaining their interactions. Response generation is guided by task-specific reasoning and predefined instructions, ensuring structured, explainable, and contextually relevant outputs that support interpretability and decision-making. Further details on our response generation taxonomy are provided in App. E.

Figs. 4-6 visually demonstrate ENWAR 2.0's response generation across different levels of performance, showcasing its adaptability, situation-aware reasoning, grounding, and human-interpretable explainability capabilities[6]. For visualization purposes, the ground truth contexts are also included in these figures; however, in real-time deployment, ENWAR 2.0 operates without access to ground truth validation data, making its ability to infer and justify beam selections solely based on multi-modal inputs a key strength of the system.

Fig. 4 illustrates ENWAR 2.0's best-case scenario, where perfect top-1 beam alignment is achieved alongside precise situation-aware reasoning and detailed tabular summaries of perception and beam prediction outcomes. A detailed version of the best-case scenario response is found in App. F. Beyond providing environmental descriptions, ENWAR 2.0 excels in reasoning and inference. For instance, it can predict potential vehicle interactions or assess how environmental conditions, such as high traffic density or signal interference, influence network performance. This capability is demonstrated in Fig. 5, which compares single-modality and full-modality responses. Fig. 5 highlights how ENWAR 2.0 leverages multi-modal data integration to improve perception and beam prediction performance significantly. Specifically, ENWAR 2.0 reasons that the Top-3 beam is the best choice given the criteria within the scenario and justifies its reasoning, and in multi-modality, it perceives and explains the environment while noting perfect beam alignment in ENWAR 2.0's analysis. A detailed version of this scenario with modality comparisons and its response is given in App. G.

[6] **Note on Visualization Scaling:** For clarity, the polar graph tick intervals are kept small (e.g., 97.63, 97.68, 97.8). Although the real beamwidth is about 6 degrees, angular distances are overscaled to better show vehicle trajectories and beam alignment, while the beamwidth is visually minimized to highlight alignment precision.



Prompt

Based on the Beam Prediction and Multi-Modal Sensor Visualization, select and evaluate the best beam's alignment with Unit 2's trajectory, focusing on GPS location, speed, bearing, and movement. Compare top-k beams for differences in received power and coverage. Determine suitability based on environmental factors (e.g., camera, lidar, radar) and their influence. Assess if the top-k beams are within ground truth bounds and align well with the vehicle's trajectory and bearing.

ENWAR 2.0 – GPS Only

Beam Predictions Overview:

- The best predicted optimal beam (24) is centered at an angle of 100.57143 degrees with a width of 2.24347 meters. This width aligns closely with Unit 2's trajectory.
- Top-1 beam (index 37): This beam has the lowest received power at Unit 2 Top-2 beam Top-2 beam (index 21): This beam has the second highest received power (0.5882)
- Top-3 beam (index 24): This beam is the closest in alignment to Unit 2's trajectory with the highest received power (0.6218)
- **Trajectory:** The predicted best beam's center angle of 100.57143 degrees closely matches the unit 2's bearing at each timestep
- **Speed:** The predicted movement aligns well with Unit 2's ground speed
- **Stability and Adaptability:** Despite the slight deviation in speed, the predicted beam shows good stability and adaptability throughout the sequence.

Top-3 beam is the best suited beam for this scenario. The predicted top-2 and top-3 beams are nearby the vehicle's estimated trajectory and bearing while the other top-k beams are not suitable at all for beam steering.

Ground Truth Beam

Ground Truth Index 26, center angle 98°, beam width 2.24347 m, coverage ±5.54502°. The Top-3 beam (index 24) is the closest in alignment to Unit 2's trajectory with the highest received power, however it is not as high as the ground truth's power, indicating misalignment with the prediction. The predicted top-2 and top-3 beams are within bounds of ground truth values or relatively close to them, however they also have lower power compared to the Top-1 beam and the ground truth.

ENWAR 2.0 – Multi-Modality

Beam Predictions Overview:

- Predicted Top-1 Beam: Has the the highest power (0.6888). Predicted top-1 best beam properties include an index of 26, a center angle of 98 degrees, a beam width of 2.24347 meters, and a coverage of ±5.54502 degrees.
- Predicted Top-2 Beam: Index 27, center angle 96.29°, same beam width, lowest power.
- Predicted Top-3 Beam: Index 25, center angle 98.86°, same beam width, second-highest power (0.6704).

Trajectory Interaction:

- GPS Location: Unit 2's predicted position shows a consistent movement of the vehicle.
- Speed Predicted at 20 km/h, with bearing predicted at 102°, anticipating straight-line movement. The image sequences confirm that the car is moving in a straight path along the road.

Deviations and Optimization:

- **Adaptability:** Predicted beams need to account for shifting GPS locations and due to detected objects from LiDAR and Radar clusters nearby the vehicle, and the other detected cars and pedestrians in the images that are moving in different directions to Unit 2 based on the image sequences.
- **Received Power:** Top-1 beam has the highest power; Top-2 and Top-3 beams are less optimal however, top-3 has the second highest power.
- **Coverage**: Top-1 beam has smaller coverage than Top-2 and Top-3 beams.

The Top-1 beam is optimal due to its superior received power and alignment with Unit 2's trajectory. Predicted beams need to account for shifting GPS locations and due to detected objects from LiDAR and Radar clusters nearby Unit 2's location, and the other detected cars and pedestrians in the image sequences that are moving in different directions to Unit 2.

Ground Truth Beam

Ground Truth Index 26, center angle 98°, beam width 2.24347 m, coverage ±5.54502°. Top-1 beam is optimal due to its superior received power and alignment with Unit 2's trajectory, supported by its match with the ground truth and alignment with the actual trajectory of Unit 2. Some additional adjustments may need to be considered due to the busy environment with multiple LiDAR and Radar objects clusters near Unit 2.

Fig. 5. Visualization and comparison of ENWAR 2.0's performance with GPS-only sensing and beam prediction to multi-modality situation-aware environment sensing, reasoning, and prediction.

ENWAR 2.0 demonstrates robust inference capabilities even in challenging scenarios where beam alignment is affected by environmental obstructions and noisy sensor data. This challenge is evident in Fig. 6, which illustrates a highly dynamic environment with multiple obstacles disrupting the beam selection process. Despite these challenges, ENWAR 2.0 effectively interprets the scene, identifying limitations in beam alignment and offering alternative explanations for network conditions. By analyzing the contextual information from the sensor inputs, ENWAR 2.0 provides insight into why certain beam selections may be unreliable while suggesting viable adjustments to maintain network connectivity. A detailed version of this scenario and its response is shown in App. H.

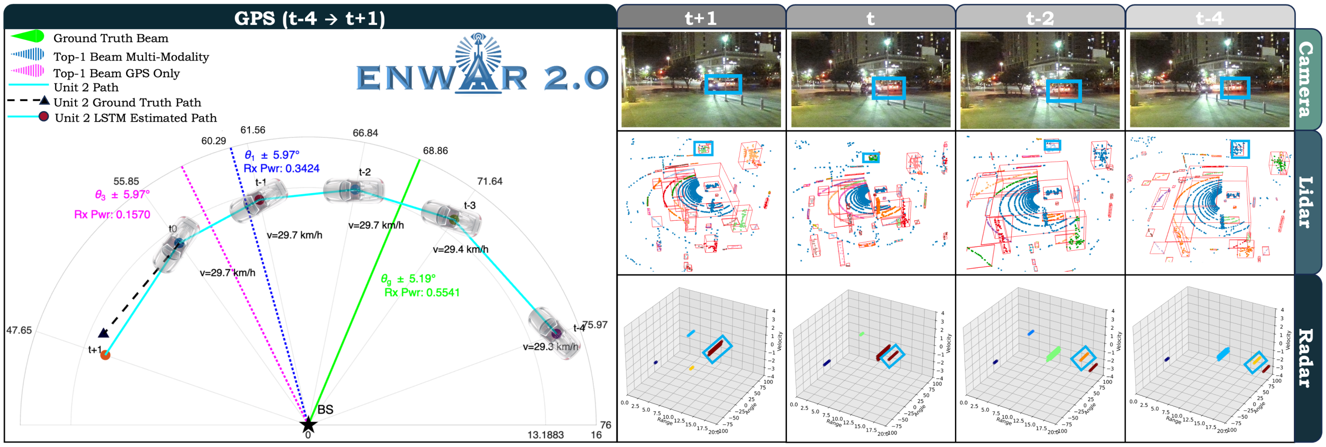
Moreover, the generated responses align closely with user-defined tasks or objectives. In ENWAR 2.0, this means providing actionable insights, such as explaining the alignment of beam properties with the vehicle trajectory or justifying beam selections in the context of environmental constraints. For instance, in Fig. 4, the system provides a clear and detailed explanation of why a particular beam was selected, based on its center angle, width, and power in relation to the predicted vehicle trajectory. Additionally, it perceives the environment and the vehicle's actions fully with respect to the dynamic occurrences of the environment, along with a tabular summary of all the situation-aware environment perception information. Prompts and responses showing these capabilities and CoT reasoning are provided in App. I.

VIII. Evaluation of ENWAR 2.0

This section evaluates ENWAR 2.0, focusing on its beam prediction accuracy and capability for situation-aware responses. The beam prediction agent is tested across multiple multi-modal sensor configurations, while ENWAR 2.0's interpretability is assessed through its justifications for beam selections based on environmental perception. The following subsections describe the evaluation methods, results, and insights into how ENWAR 2.0 integrates beam prediction and explainability for AI-driven network management.

A. Evaluation of Beam Prediction Agent

In order to comprehensively assess the agent's ability to predict optimal beams effectively, we consider three KPIs:



Prompt

Based on the Beam Prediction and Multi-Modal Sensor Visualization, select and evaluate the best beam's alignment with Unit 2's trajectory, focusing on GPS location, speed, bearing, and movement. Compare top-k beams for differences in received power and coverage. Determine suitability based on environmental factors (e.g., camera, lidar, radar) and their influence. Assess if the top-k beams are within ground truth bounds and align well with the vehicle's trajectory and bearing.

ENWAR 2.0

Predicted Top-1 Best Beam Index: 51, Top-2 Best Beam Index: 62, Top-3 Best Beam Index: 55. The predicted top-k beams are not suitable due to their lack of coverage of Unit 2's trajectory, their much lower received power and narrower coverage compared to the ground truth.

Trajectory Interaction

The predicted top-1 beam (51) is not aligned well with Unit 2's trajectory. At Timestep 3, the unit has a bearing of 66.84 degrees from the Base Station (UNIT 1), but the predicted top-1 beam index suggests an angle of 60.28571 degrees and Unit 2's final expected trajectory is at a bearing of 47.65 degrees.The predicted top-1 beam has an angle of 60.28571 degrees, which does not match Unit 2's bearing direction at all times, especially considering its speed changes and the degree difference of around ~12.5 degrees indicates the beam and Unit 2 are not near each other at all.

Optimality Justification

- Considering the multi-modal data and beam metrics, the detected LiDAR and Radar object clusters are overlapping each other in a small vicinity, and they are close to Unit 2's relative location. The image detections shows cars going in different directions which may indicated obstructions and blockages, affecting and potentially breaking line-of-sight between Unit 1 and Unit 2.
- At Timestep 4, the unit is at a steady speed (29.703 kmph). A wider angle and longer coverage might be beneficial in this situation as the vehicle seems to be turning based on the GPS data. However, for other timestamps where Unit 2 has a ground speed of approximately 29.748 km/h, a narrower beam with more precise direction might be necessary since the movement seems to stabilize.
- The predicted top-k beams are not suitable due to their lack of coverage of Unit 2's trajectory, have much lower received power and narrower coverage

Revised Beam Steering Approach

- A revised predicted optimal beam index considers a balance in adjusting the top-1 beam index. This would involve adjusting the angle, width, and power of the predicted beam while maintaining adaptability to Unit 2's trajectory changes. For instance, using the predicted top-1 beam but reducing its power and increasing its coverage slightly, and adding beam angle and width adjustments based on multi-modal detections.

Ground Truth Beam Index: 46

Given the properties of the ground truth beam index, it also does not align well with Unit 2's trajectory and movement. This observation confirms that this environment has obstructions which are affecting clear communication between Unit and Unit 2 as shown by the multi-modal detections.

Fig. 6. Worst case scenario of ENWAR 2.0's performance where the ground truth and predicted beams do not align with the vehicle due to poor conditions; however, ENWAR 2.0 shows impressive response on alternate beam steering schemes.

- Top-k Accuracy measures the proportion of cases where the ground truth beam appears within the Top-k predicted beams across the evaluation dataset.
- APL quantifies the power loss associated with selecting a non-optimal beam and is defined as:

$$PL_{[dB]} = 10 \log_{10} \left(\frac{p'}{p} \right), \tag{3}$$

  where p' is the highest-power beam among the Top-k predictions, and p is the ground truth beam power.
- Distance-based accuracy (DBA) score evaluates the directional accuracy of beam predictions and is defined as:

$$\text{DBA} = \frac{1}{K} \left(Y_1 + Y_2 + \cdots + Y_K \right), \tag{4}$$

  where

$$Y_K = 1 - \frac{1}{N} \sum_{n=1}^{N} \min_{1 \leq k \leq K} \min \left(\frac{|\hat{y}_{n,k} - y_n|}{\Delta}, 1 \right),$$

  $\hat{y}_{n,k}$ is the k-th most likely beam for input sequence n, and y_n is the ground truth beam for sequence n.

To illustrate ENWAR 2.0's target-in-the-loop beam tracking, Fig. 7 depicts the time evolution of APL throughout a session, which is defined as the period during which a vehicle remains within the BS's coverage area. Initially, as the vehicle enters the RSU's view, APL is high due to suboptimal beam alignment. As the vehicle moves within range, ENWAR 2.0 accurately predicts the optimal beams, minimizing APL to near-zero levels. However, as the vehicle exits the coverage area, APL rises again, reflecting the system's natural limitations in tracking beyond the RSU's predictive range.

ENWAR 2.0's evaluation demonstrates the significant impact of multi-modal sensor fusion on beam prediction accuracy, where carefully selected modality combinations yield substantial performance gains. A latency-aware lookahead strategy ensures real-time beam adaptation based on inference time. Given an input sampling period of 100ms and a sequence length of 500ms (five samples), configurations exceeding 200ms of inference time predict t+3, those between 100ms and 200ms predict $t+2$, and configurations below 100ms predict t+1. The accuracy values presented in this section correspond to the $t+3$ prediction accuracy. Although $t+1$ predictions yield slightly higher accuracy across all configurations, their longer inference latency would exceed the input capture window, making them impractical for real-time deployment. This measure ensures that even the configuration with the longest inference time (using all modalities) completes predictions

TABLE I
BEAM PREDICTION AGENT PERFORMANCE SUMMARY [$M = 16$, $Q = 64$] FOR PREDICTING THE BEAM AT $t + 3$

Modality Combination	Inference Time (ms)	Top-1			Top-3			Top-5		
		Accuracy	APL	DBA	Accuracy	APL	DBA	Accuracy	APL	DBA
camera_gps_lidar	200	77.8%	-0.02981	84.7%	90.0%	-0.009552	92.8%	93.4%	-0.006611	95.2%
camera_radar_lidar	212.8	77.4%	-0.03133	84.1%	89.8%	-0.009970	92.3%	93.0%	-0.006810	94.9%
camera_radar_gps_lidar	220	76.2%	-0.04025	82.3%	88.7%	-0.01016	90.2%	91.4%	-0.007003	92.7%
camera_lidar	197.4	73.6%	-0.07158	80.9%	82.4%	-0.01670	86.3%	89.1%	-0.007715	90.0%
camera_radar_gps	203.1	70.1%	-0.09937	78.7%	80.6%	-0.02080	81.2%	86.3%	-0.009831	89.1%
camera_radar	172.7	64.2%	-0.1725	74.2%	79.0%	-0.02793	80.7%	84.1%	-0.01092	88.4%
camera_gps	149.3	63.7%	-0.1790	74.0%	76.3%	-0.02847	80.5%	82.4%	-0.01641	87.3%
radar_gps_lidar	188	57.3%	-0.2073	72.8%	74.8%	-0.03625	78.1%	80.6%	-0.01901	84.9%
camera_only	122.6	53.9%	-0.2313	71.5%	72.1%	-0.04001	77.4%	77.3%	-0.02773	82.2%
radar_gps	145.3	49.4%	-0.3310	70.3%	67.9%	-0.1281	74.0%	74.8%	-0.07443	79.3%
gps_lidar	153.2	43.7%	-0.3748	67.1%	67.8%	-0.1355	73.6%	74.3%	-0.07910	79.2%
radar_lidar	170.9	41.7%	-0.3900	65.1%	65.2%	-0.1591	73.0%	70.1%	-0.09052	77.6%
radar_only	133.1	37.3%	-0.4111	63.4%	61.0%	-0.1630	72.1%	67.8%	-0.09819	75.6%
lidar_only	145	35.2%	-0.7201	61.6%	59.4%	-0.1809	68.5%	66.2%	-0.1006	69.9%
gps_only	87.4	31.4%	-0.7641	61.3%	58.3%	-0.2377	67.7%	65.0%	-0.1054	68.3%

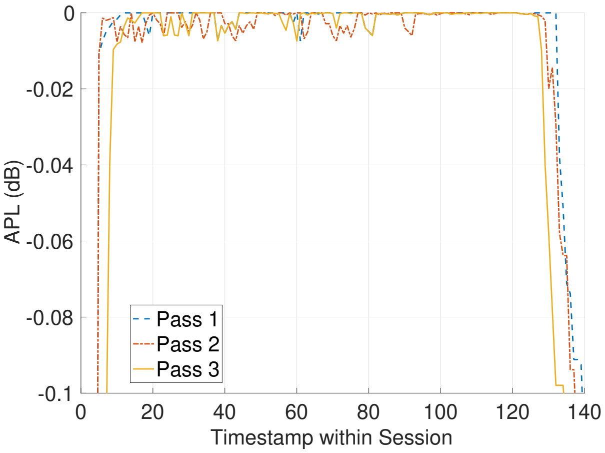


Fig. 7. Time-series analysis of beam APL per time instance within a session (pass) for three sessions. The figure illustrates the characteristic pattern of APL: high values when the vehicle enters the base station's view, a drop to near-zero levels as optimal beams are predicted within range, and a subsequent rise in APL as the vehicle exits the coverage area.

before the next sequence is available.

As shown in Table I, the highest Top-3 accuracy of 90.0% is achieved by the *camera_gps_lidar* configuration, emphasizing the critical role of combining spatial, depth, and semantic perception for precise beam alignment. Similarly,configurations incorporating radar, such as *camera_radar_lidar* (89.8%) and *camera_radar_gps* (80.6%), highlight the value of motion tracking in enhancing beam selection under dynamic conditions. Camera data consistently emerges as the most influential modality, while LiDAR and radar provide valuable complementary depth and motion information.

Notably, despite incorporating all modalities, the *camera_radar_gps_lidar* configuration achieves a slightly lower Top-3 accuracy of 89.8% compared to *camera_gps_lidar*. This reduction in accuracy can be attributed to the increased model complexity introduced by redundant or partially overlapping features across modalities, which may lead to optimization challenges during training and introduce additional noise when certain features conflict or saturate the model capacity. Such behavior is consistent with prior observations in multi-modal learning, where adding excessive modalities can lead to feature redundancy and representational conflicts that hinder generalization. This observation highlights that while adding more modalities generally enhances beam selection, excessive fusion may not always yield proportional gains, emphasizing the importance of selective, task-aware modality integration to balance prediction performance and computational efficiency.

Ultimately, camera data remains the top contributor to accurate beam prediction, with LiDAR and radar serving as important enhancers in constructing a robust, real-world AI-driven beam prediction system.

B. Evaluation of ENWAR 2.0 *Response Generation*

1) Enwar 2.0 Setup: To evaluate the performance of ENWAR 2.0, we utilized models from the LLaMa3 family: Llama3.2-3B/LLaMa3.1-8B/LLaMa3.3-70B for text-based tasks and LLaMa3.2-Vision-11B for image-to-text processing. The models were deployed on an A100 GPU equipped with 40GB of VRAM. LLaMa3.2-Vision-11B fully utilized 8GB of VRAM, LLaMa3.2-3B required 3.4GB, LLaMa3.1-8B utilized 16GB, and LLaMA3.3-70B required 32GB of VRAM. LLaMa3.2 is a lightweight version of LLaMa3.1, and LLaMa3.3-70B is the upgraded version LLaMa3.1-70B. All models support a maximum context length of 128k tokens. For deployment considerations and system limitations of ENWAR 2.0, we refer the readers to App. K.

Pretrained LLMs can be deployed with configurable hyperparameters influencing their generation capabilities and interpretive performance. For ENWAR 2.0, we selected hyperparameters tailored to support multi-modal perception tasks, balancing interpretability, creativity, conciseness, and response relevancy. Specifically, we selected `max_new_tokens` = 4096 to allow sufficiently long responses for detailed reasoning; `temperature`=0.5 to moderate output diversity while maintaining factual precision; and `repetition_penalty`=1.15 to reduce redundant phrases and encourage concise, varied responses. These settings were chosen after preliminary ablation tests and align with best practices reported in prior LLM-driven reasoning frameworks, providing stable performance across diverse scenarios while avoiding excessive verbosity or hallucinations.

To improve retrieval granularity, LlamaIndex was integrated with FAISS, providing more fine-grained control over data selection and ensuring high relevancy during inference. The GTE model, `stella_en_400M_v5` [37] with a maximum token size of 8192, was used to generate embeddings. To maintain semantic coherence across segments, text chunking was applied using a size of 800 characters and an 80 character overlap for minimal information loss across chunk boundaries.

On average, the extracted multi-modal data, including all sensor inputs and associated beam properties, amounted to $\approx$ 20k tokens. With RAG, this number was reduced to $\approx$13k tokens per user prompt by retrieving only the most relevant context before integrating it into the final prompt template.

For benchmarking purposes, off-the-shelf pretrained versions of each textual model (denoted as Vanilla LLaMa) were used as a baseline without RAG capabilities. In this baseline, the entire 20k tokens were directly injected into the prompt template without using RAG capabilities, relying solely on learning from real-time data. Our second baseline (static, non-adaptive RAG) involved preemptively appending all data samples into the KB to evaluate whether having full, static knowledge alters performance. This comparison allowed for a clear assessment of the improvements gained from the relevance-based retrieval process in ENWAR 2.0.

2) Evaluation Dataset Creation: A comprehensive evaluation of ENWAR 2.0 was conducted using a diverse, multi-modal dataset to assess both its predictive performance and interpretability. To analyze the system's ability to effectively integrate and leverage multi-modal data, 15 distinct modality permutations were designed. A total of 150 carefully curated samples were selected to construct a dedicated KB for each permutation, resulting in a total of 15 specialized KBs (2250 total samples). These samples encompassed a wide range of real-world scenarios, ensuring diverse contextual representations across different sensor configurations.

Each sample consisted of multi-modal input streams where combinations of GPS, LiDAR, radar, and camera data, were varied to assess their contribution to contextual understanding and predictive performance. The embedding storage and indexing processes enabled efficient real-time retrieval, ensuring that the KBs remained optimized for speed and accuracy. The dataset was then preprocessed and transformed into a textual format suitable for RAG-based processing, following the methodology outlined in Section VII-A.

To assess the interpretability of ENWAR 2.0, a set of 50 manually crafted validation Q&A pairs was developed for each of the 15 modality combinations, resulting in a total of 750 evaluation pairs. Half of the questions were designed to follow the interpretation category, and the other half follows the perception category, which will be explained in the next subsection. Each question, $\approx$ 200 tokens in length, was designed to probe ENWAR 2.0's reasoning behind beam selection decisions. The questions examined factors such as: "Why a particular beam was selected among the Top-k predictions" or "How multi-modal inputs, such as object detection, trajectory alignment, and spatial awareness, contributed to the decision." The corresponding answers served as benchmarks for evaluating ENWAR 2.0's explanatory capabilities, providing a structured basis for measuring the system's ability to generate accurate, coherent, and contextually grounded explanations. Examples of these questions can be found in Apps. F-J.

3) Performance Criteria and KPIs: The evaluation questions were presented to ENWAR 2.0, which generated explanations for situation-aware environment reasoning and beam predictions. The evaluation focused on two primary criteria:

- *Beam Interpretation* assesses ENWAR 2.0's ability to justify its beam selection based on beam properties and the dynamic scene conditions. It measures how effectively the model links its predictions to factors such as trajectory alignment, signal strength, and spatial coverage.
- *Environmental Perception* evaluates ENWAR 2.0's ability to interpret the surrounding environment, focusing on the analysis of object detections, overlaps, traffic density, and their influence on network performance. Perception assesses how well the model infers the reasons for a beam's performance under specific environmental conditions.

The performance of ENWAR 2.0 was assessed using both general benchmarks and domain-specific evaluation frameworks. While standard benchmarks such as General Language Understanding Evaluation (GLUE) and Massive Multitask Language Understanding (MMLU) provide a general measure of LLM capabilities, including answer relevancy, factual correctness, and hallucination detection, these are insufficient for domain-specific tasks involving RAG-based frameworks.

To ensure a thorough evaluation of both interpretability and perception, we defined the following KPIs: 1) *faithfulness* evaluates the consistency of responses with the retrieved context, 2) *correctness* assesses the factual correctness and semantic similarity of generated responses, and 3) *answer relevancy* measures the semantic alignment of ENWAR 2.0's responses with the user's prompt and context to ensure the ability to generate contextually relevant situation-aware environment reasoning and beam prediction responses. We refer readers to ENWAR 1.0 and the LlamaIndex documentation[7] for their mathematical formulation [9].

4) Evaluation of Interpretation and Perception: The performance of ENWAR 2.0 was evaluated against baselines using KPIs that measure interpretation, perception, faithfulness, and relevancy. Results show that adaptive RAG significantly improves correctness and contextual accuracy by leveraging the most recent multi-modal inputs while maintaining retrieval efficiency.

Understanding how modality combinations affect beam prediction and environmental perception is critical for robust reasoning in dynamic scenarios. As shown in Fig. 8, visual sensor-based configurations consistently enhance scores across all KPIs, highlighting their importance for capturing semantic and spatial features essential for accurate beam alignment and justification.

Adaptive RAG further amplifies these gains, consistently outperforming static retrieval across all metrics. Dynamically updating the KB aligns retrieval with evolving environments,

[7] https://docs.llamaindex.ai/en/stable/

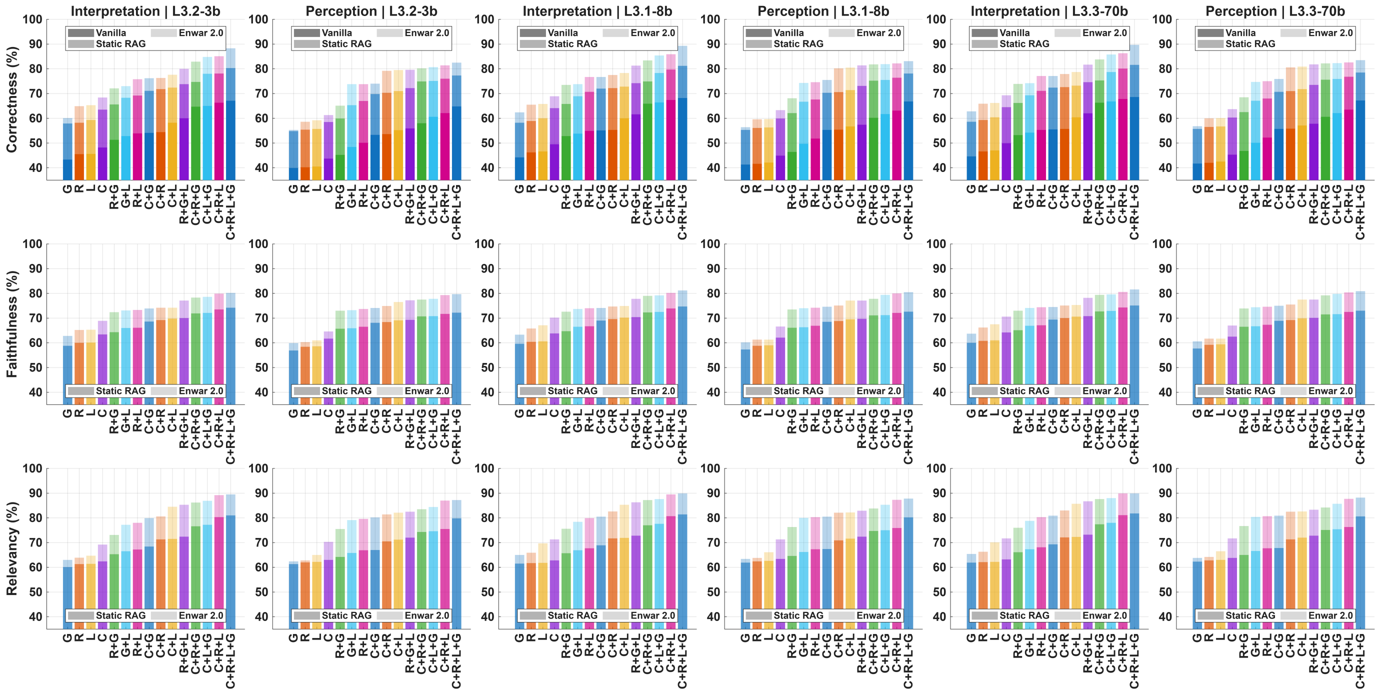


Fig. 8. KPI comparison of ENWAR 2.0 across various modality combinations (C: camera, G: GPS, L: LiDAR, R: radar) using LLaMa3.2-3B, 3.1-8B, and 3.3-70B under different RAG strategies. Bars indicate evaluation performance: the lightest bars show ENWAR 2.0, mid-transparency bars represent static and non-adaptive RAG (static RAG), and the darkest bars depict vanilla LLaMa without RAG, which does not report relevancy or faithfulness metrics.

reducing irrelevant context and improving precision. In contrast, the static and non-adaptive RAG suffers from lower correctness due to outdated or less relevant information. While additional modalities benefit, adaptive RAG alone delivers substantial improvements, underscoring its critical role in real-time AI-driven wireless systems.

Faithfulness and relevancy are not reported for baselines without RAG (Vanilla LLaMa). In Fig. 8, transparent bars show adaptive RAG performance, while opaque bars show static retrieval. The observed improvements validate adaptive RAG as a key mechanism for ensuring accurate and contextually relevant responses while minimizing retrieval inefficiencies.

The evaluation highlights the substantial impact of adaptive RAG in improving interpretation and perception accuracy across different modality configurations. As seen in the results, adaptive RAG consistently enhances response correctness, faithfulness, and relevancy by dynamically retrieving the most recent and contextually relevant KB entries, reducing information overload and improving response precision. This subsection discusses the results from the perspective LLaMa3.3-70B.

The camera-based configurations consistently outperform other sensor combinations, emphasizing the importance of visual perception for semantic and spatial scene understanding. The *camera_radar_lidar_gps* configuration achieves the highest interpretation correctness (89.7%) and perception correctness (83.5%) with adaptive RAG, demonstrating the effectiveness of multi-modal fusion. The removal of adaptive RAG results in a performance drop to 81.6% and 78.5%, highlighting the critical role of adaptive retrieval in maintaining real-time environmental awareness.

Lower-performing configurations, such as *radar_gps* and *gps_lidar*, experience significant drops in correctness without adaptive RAG, emphasizing the limitations of static knowledge retrieval when dealing with dynamic environments. For instance, the *gps_lidar* configuration drops from 74.2% to 69.3% in interpretation correctness and from 74.7% to 67.1% in perception correctness, illustrating how adaptive retrieval improves accuracy by ensuring that the system references only the most relevant and recent environmental data.

Beyond correctness, adaptive RAG improves faithfulness and relevancy, reducing hallucinations and enhancing response alignment with real-world conditions. Across all configurations, adaptive retrieval increases interpretation and perception faithfulness by 9.1% and 9.5%, and interpretation and perception relevancy by 13.7% and 13.2%, ensuring that beam justifications remain grounded in factual, scenario-specific data rather than outdated or redundant information.

These results validate the effectiveness of adaptive RAG in optimizing real-time network decision-making. By prioritizing recent environmental updates and filtering out extraneous information, ENWAR 2.0 achieves superior interpretability, reduced retrieval inefficiencies, and enhanced response accuracy, making it more robust in dynamic wireless environments.

5) Benchmarking with Vanilla LLaMa: The Vanilla LLaMa baselines achieved only 68.6% interpretation correctness and 67.2% perception correctness using all modality extractions on LLaMa3.3-70B; substantially lower than ENWAR 2.0. This gap underscores the benefit of RAG, which selectively retrieves only the most relevant KB contexts instead of injecting all data into the prompt. By filtering out irrelevant information, ENWAR 2.0 boosts correctness while reducing computational

TABLE II
ENWAR 2.0 KEY TIME INDICATORS

Process	Time
Off-the-Shelf Llama3.3-70B prompt with approx. 20k Tokens	2.67s
Llama3.2-Vision Image-To-Text (per frame)	2.34s
Off-the-Shelf Llama3.1-8B prompt with approx. 20k Tokens	2.44s
Off-the-Shelf Llama3.2-3B prompt with approx. 20k Tokens	2.38s
Llama3.3-70B w/RAG prompt with approx. 13k Tokens	1.26s
Llama3.1-8B w/RAG prompt with approx. 13k Tokens	1.18s
Llama3.2-3B w/RAG prompt with approx. 13k Tokens	1.12s
stella GTE vectorization (150 samples=500mb)	400ms
Beam Prediction Inference (worst case)	220ms
FAISS retrieval	50ms
LiDAR DBSCAN	16.3ms
Radar DBSCAN	12.8ms
YOLO Camera Detections (per frame)	5ms
GPS Processing	$<$ 1ms

overhead and latency, demonstrating the effectiveness of RAG-enabled retrieval for real-time network decision-making.

Based on the KPIs summarized in Table II, the total processing time for ENWAR 2.0 is 4.35 seconds, significantly outperforming the baseline Vanilla LLaMa, which requires 5.77 seconds for preprocessing and response generation. A significant portion of this processing time stems from the image-to-text generation step, which provides additional scene descriptions for the LLM. While this step enhances environmental context, the primary contribution to beam prediction accuracy is modality extractions and object detections (e.g., LiDAR clustering, radar processing, and GPS trajectory estimation). Image-to-text generation can be considered an optional step, particularly in scenarios where rapid inference is prioritized over detailed scene descriptions.

This efficiency gain holds despite the extra steps introduced by the RAG pipeline, such as preprocessing, embedding generation, retrieval, and response generation. The key driver is reduced token processing. Vanilla LLaMa handles full prompts of about 20k tokens, including multi-modal data, beam properties, and a 200-token user input, whereas RAG-enabled ENWAR 2.0 processes only 13k contextually relevant tokens plus the same 200-token prompt. This 7k token reduction lightens the computational load and accelerates inference.

6) Normalized Gains with Larger LLMs: While scaling to larger LLMs like Llama3.1-8B and Llama3.3-70B yields higher KPI metrics, the marginal benefits diminish rapidly relative to model size, as seen in Fig. 9. The average normalized gains per billion parameters drop from 24.9 for Llama3.2-3B, 9.47 for Llama3.1-8B, and just 1.09 for Llama3.3-70B. Moreover, larger models introduce significantly longer inference times; for instance, Llama3.3-70B requires 1.26 seconds per RAG prompt, compared to only 1.12 seconds for Llama3.2-3B. Thus, although larger models offer slight performance improvements, they come at a considerable computational cost, making smaller models more practical for real-time deployments in ENWAR 2.0.

7) Computational and Deployment Considerations and Limitations: ENWAR 2.0 delivers real-time multi-modal beam prediction and situation-aware reasoning but faces challenges such as inference latency, memory demands, and scalable retrieval efficiency. Although deployed at BSs with stable power, larger LLMs substantially increase computational and energy costs, motivating future work on model compression and adaptive scaling. The framework integrates a TransFusion-based beam prediction agent, an environment perception agent, and an adaptive RAG pipeline for explainable decision-making. Beam prediction operates with a latency between 100–220ms, maintaining real-time network performance. While currently evaluated in a single-user context, the system is designed for multi-user scalability, with FAISS indexing enhancing retrieval speed. Still, distributed retrieval architectures and caching will be essential for larger deployments. Limitations include diminishing returns from excessive modality fusion, reliance on high-quality sensor inputs, and the potential for increased retrieval latency as KBs expand. Additional computational and deployment details are provided in App. K.

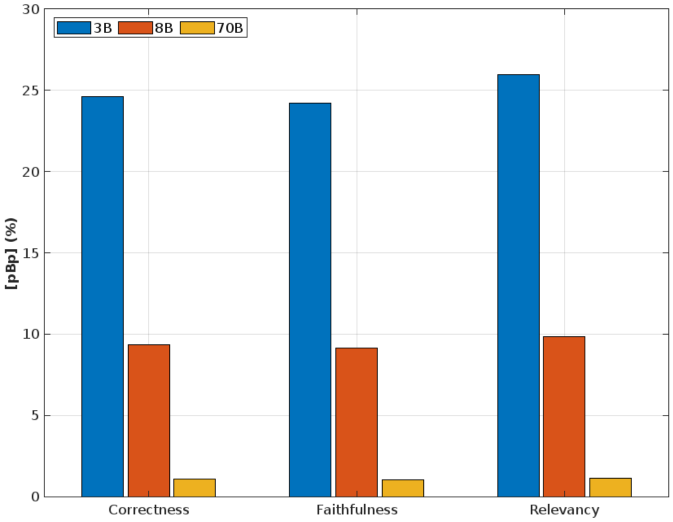


Fig. 9. KPI efficiency per billion parameters (pBp)

IX. CONCLUSION

ENWAR 2.0 is the first system to integrate LLMs, RAG, and agentic beam prediction for next-generation wireless networks, uniting situation-aware environmental perception with explainable AI-driven beam selection. Unlike conventional systems that address beam prediction or environmental perception in isolation, ENWAR 2.0 holistically fuses both, ensuring that decisions are contextually grounded and interpretable. It mitigates hallucinations and outdated information by integrating real-time multi-modal sensory data and leveraging RAG for relevant historical and contextual retrieval. Our new evaluation metrics—interpretation and perception—demonstrate the system's capability to predict optimal beams while providing meaningful justifications that enhance transparency and trust. Experimental results show ENWAR 2.0 achieves a Top-3 beam prediction accuracy of 90.0% at $t + 3$, with interpretation correctness up to 89.7% and perception correctness of 83.5%, maintaining faithfulness and relevancy up to 81.6% and 89.9%. Beyond beamforming, ENWAR 2.0 signals a move toward AI-native, proactive decision-making in 6G networks, enabling intelligent, adaptive, and self-optimizing wireless systems.

Acknowledgments

We utilized ChatGPT [38] throughout the writing process to assist with organizing content and refining the language flow of the paper, and ensuring consistency in tone and terminology.

References

[1] A. M. Nazar, A. Celik, M. Y. Selim, A. Abdallah, D. Qiao, and A. M. Eltawil, "Encoders, roll out! A multi-modal sensor transfusion for proactive I2V beam prediction," in *Asilomar Conference on Signals, Systems, and Computers*, 2025. [Online]. Available: http://hdl.handle.net/10754/703116

[2] A. Celik and A. M. Eltawil, "At the dawn of generative AI era: A tutorial-cum-survey on new frontiers in 6G wireless intelligence," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 2433–2489, 2024.

[3] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Sys.*, 2017.

[4] C. Benzaid and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions," *IEEE Network*, vol. 34, no. 2, pp. 186–194, 2020.

[5] A. Alkhateeb, S. Jiang, and G. Charan, "Real-time digital twins: Vision and research directions for 6G and beyond," *IEEE Commun. Mag.*, 2023.

[6] E. M. Bender *et al.*, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. ACM FAccT*, 2021, p. 610–623.

[7] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, 2023.

[8] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *ArXiv*, vol. abs/2312.10997, 2023.

[9] A. M. Nazar, A. Celik, M. Y. Selim, A. Abdallah, D. Qiao, and A. M. Eltawil, "ENWAR: A RAG-empowered multi-modal LLM framework for wireless environment perception," *IEEE Commun Mag.*, 2025 [in press]. [Online]. Available: https://arxiv.org/abs/2410.18104

[10] A. Abdallah, A. Albaseer, A. Celik, M. Abdallah, and A. M. Eltawil, "NetOrchLLM: Mastering wireless network orchestration with large language models," *arXiv preprint arXiv:2412.10107*, 2024.

[11] N. Khan, S. Coleri, A. Abdallah, A. Celik, and A. M. Eltawil, "Explainable and robust artificial intelligence for trustworthy resource management in 6G networks," *IEEE Communications Magazine*, 2024.

[12] S. Xu *et al.*, "Large multi-modal models (LMMs) as universal foundation models for AI-native wireless systems," *arXiv preprint arXiv:2402.01748*, 2024.

[13] G. M. Yilma *et al.*, "TelecomRAG: Taming telecom standards with retrieval augmented generation and LLMs," *arXiv preprint arXiv:2406.07053*, 2024.

[14] A. M. Nazar, M. Y. Selim, D. Qiao, and H. Zhang, "NextG-GPT: Leveraging GenAI for advancing wireless networks and communication research," 2025. [Online]. Available: https://arxiv.org/abs/2505.19322

[15] A. Maatouk *et al.*, "TeleQnA: A benchmark dataset to assess large language models telecommunications knowledge," *arXiv preprint arXiv:2310.15051*, 2023.

[16] H. Zou *et al.*, "TelecomGPT: A framework to build telecom-specfic large language models," *arXiv preprint arXiv:2407.09424*, 2024.

[17] S. Tarkoma, R. Morabito, and J. Sauvola, "AI-native interconnect framework for integration of large language model technologies in 6G systems," *arXiv preprint arXiv:2311.05842*, 2023.

[18] Y. Shen, J. Shao, X. Zhang, Z. Lin, H. Pan, D. Li, J. Zhang, and K. B. Letaief, "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Communications Magazine*, 2024.

[19] H. Zou *et al.*, "Wireless multi-agent generative AI: From connected intelligence to collective intelligence," *arXiv preprint arXiv:2307.02757*, 2023.

[20] L. Bariah *et al.*, "Large generative AI models for telecom: The next big thing?" *IEEE Commun. Mag.*, 2024.

[21] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, D. Niyato, and O. A. Dobre, "Large language model enhanced multi-agent systems for 6G communications," *IEEE Wireless Communications*, 2024.

[22] G. Liu, Y. Liu, R. Zhang, H. Du, D. Niyato, Z. Xiong, S. Sun, and A. Jamalipour, "Wireless agentic AI with retrieval-augmented multimodal semantic perception," 2025. [Online]. Available: https://arxiv.org/abs/2505.23275

[23] B. Du, H. Du, D. Niyato, and R. Li, "Task-oriented semantic communication in large multimodal models-based vehicle networks," *IEEE Transactions on Mobile Computing*, 2025. [Online]. Available: http://dx.doi.org/10.1109/TMC.2025.3564543

[24] N. Khan, A. Abdallah, A. Celik, A. M. Eltawil, and S. Coleri, "Explainable and robust millimeter wave beam alignment for AI-native 6G networks," *arXiv preprint arXiv:2501.17883*, 2025.

[25] ——, "Explainable AI-aided feature selection and model reduction for DRL-based V2X resource allocation," *arXiv preprint arXiv:2501.13552*, 2025.

[26] S. Alikhani, G. Charan, and A. Alkhateeb, "Large wireless model (LWM): A foundation model for wireless channels," 2025. [Online]. Available: https://arxiv.org/abs/2411.08872

[27] N. Gao, Y. Liu, Q. Zhang, X. Li, and S. Jin, "Let Rff do the talking: Large language model enabled lightweight RFFI for 6G edge intelligence," *Science China Information Sciences*, 2025.

[28] K. Ding, C. Guo, Y. Yang, W. Hu, and Y. C. Eldar, "A new paradigm of user-centric wireless communication driven by large language models," 2025. [Online]. Available: https://arxiv.org/abs/2504.11696

[29] J. Shao *et al.*, "WirelessLLM: Empowering large language models towards wireless intelligence," *arXiv preprint arXiv:2405.17053*, 2024.

[30] M. Xu, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, D. I. Kim, and K. B. Letaief, "When large language model agents meet 6G networks: Perception, grounding, and alignment," pp. 63–71, 2024.

[31] Y. Feng, D. He, Y. Guan, Y. Huang, Y. Xu, and Z. Chen, "Beamwidth optimization for millimeter-wave V2V communication between neighbor vehicles in highway scenarios," *IEEE Access*, vol. 9, pp. 4335–4350, 2021.

[32] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The FAISS library," 2024.

[33] G. Charan, U. Demirhan, J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Multi-modal beam prediction challenge 2022: Towards generalization," 2022. [Online]. Available: https://arxiv.org/abs/2209.07519

[34] 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*, 2023.

[35] B. Wang, J. Lan, and J. Gao, "LiDAR filtering in 3D object detection based on improved ransac," *Remote Sensing*, vol. 14, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/9/2110

[36] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," 2017. [Online]. Available: https://arxiv.org/abs/1612.00593

[37] D. Zhang, J. Li, Z. Zeng, and F. Wang, "Jasper and stella: distillation of sota embedding models," 2025. [Online]. Available: https://arxiv.org/abs/2412.19048

[38] OpenAI. [Online]. Available: https://chat.openai.com/chat

Ahmad M. Nazar (Graduate Student Member, IEEE) is an Postdoctoral Scholar in the Department of Electrical and Computer Engineer at Iowa State University, USA. He received a Ph.D. degree in Computer Engineering from Iowa State University, USA, in 2025; and he also earned M.S. and B.S degrees in Computer Engineering in 2022 and 2020, respectively. His research interests are in the interdisciplinary research and applications of generative AI, and machine learning with a focus on LLMs and autonomous, multi-modal AI agents.

Abdulkadir Celik (Senior Member, IEEE) is an Associate Professor in the School of Electronics and Computer Science at the University of Southampton, UK, where he also serves as the Director of the Centre for Internet of Things and Pervasive Systems. He received the Ph.D. degree in co-majors of Electrical Engineering and Computer Engineering from Iowa State University, Ames, IA, USA, in 2016; wherein he also earned M.S. degrees in Electrical Engineering and Computer Engineering in 2013 and 2015, respectively. Prior to his current appointment, he was a senior research scientist from 2020 to 2025 and a post-doctoral fellow from 2016 to 2020 at King Abdullah University of Science and Technology (KAUST), Thuwal, KSA. Dr. Celik is the recipient of IEEE Communications Society's 2023 Outstanding Young Researcher Award for Europe, Middle East, and Africa (EMEA) region. He currently serves as an editor for npj Wireless Technology, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATION LETTERS, and Frontiers in Communications and Networks. His research interests are in the broad areas of next-generation wireless communication systems and networks.

Mohamed Y. Selim (Senior Member, IEEE) is an Associate Teaching Professor in the Department of Electrical and Computer Engineering at Iowa State University, USA. He received the Ph.D. and M.Sc. degrees in computer engineering from Iowa State University in 2018 and 2016, respectively, and the M.Sc. degree in electrical engineering from Port Said University, Egypt, in 2011. He has co-authored numerous papers in leading venues such as IEEE Communications Magazine, IEEE Transactions on Mobile Computing, and Computer Networks, and contributed book chapters on 5G and UAV-based RIS systems. Dr. Selim has received multiple teaching and educational impact awards at Iowa State University and has been recognized as an Exemplary Reviewer for IEEE Communications Letters (2023) and a Distinguished Reviewer for IEEE Transactions on Mobile Computing (2024). Dr. Selim's research interests span next-generation wireless communication systems, large-scale testbeds, reconfigurable intelligent surfaces, and agentic LLMs for AI-native wireless networks. He is a co-PI of the $8 million NSF-funded ARA PAWR project.

Asmaa Abdallah (Member, IEEE) received the B.S. (with High Distinction) and M.S degree in computer and communications engineering from Rafik Hariri University (RHU), Lebanon, in 2013 and 2015, respectively. In 2020, she received the Ph.D. degree in electrical and computer engineering at the American University of Beirut (AUB), Beirut, Lebanon. She was a Postdoctoral Fellow at King Abdullah University of Science and Technology (KAUST), from 2021-2024, where she is currently a Research Scientist with the Communications and Computing Systems Laboratory. From 2016 to 2020, She has been a member of the executive committee of IEEE Young Professionals Lebanon's Section. Dr. Abdallah was the recipient of the Academic Excellence Award at RHU in 2013 for ranking first on the graduating class. She also received a scholarship from the Lebanese National Counsel for Scientific Research (CNRS-L/AUB) to support her doctoral studies. In 2023, Dr. Abdallah has been selected by MIT technology review as one of the leading 15 Innovators under 35 in the MENA area. Her research interests include machine learning, communication theory, stochastic geometry, array signal processing, with emphasis on energy and spectral efficient algorithms for next-generation wireless communication systems.

Daji Qiao (Senior Member, IEEE) is a Professor in the Department of Electrical and Computer Engineering at Iowa State University. He received his Ph.D. from the University of Michigan, Ann Arbor. His research focuses on wireless networking and mobile computing, 5G/6G systems, sensor networks, and IoT. He is a Senior Member of the IEEE and a Member of the ACM.

Ahmed M. Eltawil (Senior Member, IEEE) is a Professor and Associate Dean for Research at the Computer, Electrical, and Mathematical Sciences and Engineering (CEMSE) Division at King Abdullah University of Science and Technology (KAUST). Previously, he was a Professor of Electrical Engineering and Computer Science at the University of California, Irvine (UCI) from 2005 to 2021. Professor Eltawil earned his doctorate degree from the University of California, Los Angeles, in 2003 and his Master's and Bachelor's degrees from Cairo University in 1999 and 1997, respectively. At KAUST, he established the Communication and Computing Systems Laboratory (CCSL) to conduct research on efficient architectures for computing and communications systems, with a particular focus on mobile wireless systems. His research interests encompass various application domains, such as low-power mobile systems, machine learning platforms, sensor networks, body area networks, and critical infrastructure networks. He served as a distinguished lecturer for IEEE COMSOC during the 2023/24 term. Additionally, he holds senior membership in both the IEEE and the National Academy of Inventors in the United States. He received several recognitions and awards, including the US National Science Foundation CAREER award, the 2021 "Innovator of the Year" award by the Henry Samueli School of Engineering at the University of California, Irvine, and two United States Congress certificates of merit, among other recognitions. He has served in numerous editorial roles over the years, as well as an expert reviewer for national and international funding agencies and review boards.