

V2X-UniPool: Unifying Multimodal Perception and Knowledge Reasoning for Autonomous Driving

Xuewen Luo¹, Fengze Yang¹, Ding Fan², Xiangbo Gao³, Bo Yu¹, Zihao Li³
Zhengzhong Tu³, Yang Zhou³, Chenxi Liu¹

¹University of Utah, ²Monash University, ³Texas A&M University

xuewen.luo@utah.edu

Abstract

Autonomous driving (AD) has achieved significant progress, yet single-vehicle perception remains constrained by sensing range and occlusions. Vehicle-to-Everything (V2X) communication addresses these limits by enabling collaboration across vehicles and infrastructure, but it also faces heterogeneity, synchronization, and latency constraints. Language models offer strong knowledge-driven reasoning and decision-making capabilities, but they are not inherently designed to process raw sensor streams and are prone to hallucination. We propose V2X-UniPool, the first framework that unifies V2X perception with language-based reasoning for knowledge-driven AD. It transforms multimodal V2X data into structured, language-based knowledge, organizes it in a time-indexed knowledge pool for temporally consistent reasoning, and employs Retrieval-Augmented Generation (RAG) to ground decisions in real-time context. Experiments on the real-world DAIR-V2X dataset show that V2X-UniPool achieves state-of-the-art planning accuracy and safety while reducing communication cost by more than 80%, achieving the lowest overhead among evaluated methods. These results highlight the promise of bridging V2X perception and language reasoning to advance scalable and trustworthy driving. Our code is available at: <https://anonymous.4open.science/r/V2X-UniPool-7326>

1. Introduction

Autonomous driving (AD) has achieved remarkable progress, however single-vehicle intelligence remains fundamentally constrained by limited sensing range and occlusion. To overcome these challenges, **Vehicle-to-Everything (V2X) communication** has emerged as a promising paradigm, enabling collaborative perception across vehicles and infrastructure [17, 40, 46]. By integrating distributed sensing resources, V2X extends perception

range and improves safety in complex scenarios [16]. In parallel, the rapid advancement of language models including **large language models (LLMs)** and **vision-language models (VLMs)** is catalyzing a shift from purely data-driven to knowledge-driven pipelines [32]. Equipped with common-sense knowledge and reasoning ability, these models offer the potential to support human-like decision-making and more intelligent driving policies in dynamic traffic environments [13, 38].

Despite their promise, both approaches face fundamental limitations. V2X fusion introduces cross-modal heterogeneity, temporal misalignment, and non-negligible communication overhead [35, 39], hindering the establishment of a unified and temporally consistent representation. Meanwhile, although LLMs and VLMs excel at processing natural language and static images, they cannot directly consume raw multimodal sensor data essential to AD (e.g., LiDAR point clouds, high-frequency trajectories), and are prone to hallucination when deprived of real-time grounding [4, 15, 28]. **This exposes a critical gap: existing V2X frameworks remain perception-centric, while language models remain knowledge-centric; no solution unifies the two into a language-based communication paradigm that enables temporally consistent and knowledge-grounded reasoning.**

We propose **V2X-UniPool**, the first framework that bridges V2X perception with language-based reasoning for AD. The system comprises three modules: (i) a translation layer that converts raw multimodal inputs into a unified language-based representation, (ii) a time-indexed knowledge pool that organizes static and dynamic scene elements for synchronized reasoning, and (iii) a Retrieval-Augmented Generation (RAG) mechanism that grounds decisions in structured, real-time environmental knowledge. Together, these components directly address format fragmentation, temporal desynchronization, and weak grounding, providing a scalable bridge between collaborative sensing and knowledge-driven planning. Our contributions are

threefold:

- We introduce **V2X-UniPool**, the first **knowledge-driven language-based V2X framework** for AD systems in scenario-aware, smarter decision making. The framework contains: (i) a language translation module, (ii) a time-indexed knowledge pool, and (iii) contextualized RAG grounding.
- To address heterogeneity and synchronization issues, we construct a **time-indexed knowledge pool** that integrates infrastructure-side sensor records into a unified, language-based representation. Accompanied by temporal static and dynamic data, the pool enables efficient storage, real-time retrieval, and temporally consistent reasoning.
- V2X-UniPool incorporates a **RAG mechanism** that allows models to retrieve this time-indexed knowledge pool for relevant scene context. This real-time grounding effectively bridges the modality gap and reduces hallucinations.
- Experiments on the real-world DAIR-V2X dataset demonstrate that V2X-UniPool achieves **state-of-the-art planning accuracy and safety**, while reducing communication cost by more than 80% and attaining the **lowest overhead**, even under zero-shot vehicle-side model.

2. Related Work

2.1. V2X Data Fusion in Cooperative Perception

Vehicle-independent perception, which relies solely on on-board sensors and intra-vehicle fusion, often suffers from limited field of view and occlusions, making it difficult to ensure reliable situation awareness in dense or unstructured environments [23]. In contrast, cooperative perception, enabled by Vehicle-to-Everything (V2X) communication, facilitates the sharing of sensory information among vehicles and infrastructure, leading to a more comprehensive understanding of the traffic environment [14, 19, 37]. This approach has proven particularly valuable in high-occlusion scenarios, where external perspectives can compensate for blind spots and improve safety-critical decision making [20, 25, 33].

Data fusion plays a vital role in this approach [14, 27]. Its methods are commonly divided into three categories: early fusion, intermediate fusion, and late fusion. Early fusion aggregates raw sensor data across agents before any local processing, offering fine-grained cooperation but incurring high bandwidth and strict temporal synchronization requirements [2, 5]. Intermediate fusion shares encoded features among agents to achieve a balance between communication cost and accuracy, and has therefore become a widely adopted paradigm [19, 29, 30, 35]. However, it faces major challenges in heterogeneous environments [11, 16]. In contrast, late fusion, especially when implemented through

LLM-based reasoning, can better accommodate agent heterogeneity by deferring integration to a high-level semantic space [6, 10].

In this work, we treat shared outputs as language-level abstractions rather than transmitting raw data or intermediate features as in prior data-fusion methods. This design not only significantly reduces communication cost but also supports efficient, interpretable reasoning under realistic bandwidth constraints.

2.2. Knowledge-Empowered Autonomous Driving

The rapid advancement of language models is reshaping the field of AD, enabling a shift from traditional, situation-aware pipelines to more holistic and knowledge-driven paradigms. By leveraging their ability to understand complex driving scenes, they are increasingly regarded as foundational components in modern AD architectures [21, 47]. Their powerful contextual reasoning capabilities allow AD systems to interpret intricate traffic scenarios and support more informed decision-making. Recent studies have demonstrated that language models not only enhance perception but also contribute to downstream tasks such as trajectory prediction and motion planning, thereby improving the overall safety, adaptability, and intelligence of AD systems [12, 38].

Building on this foundation, recent work has shown the potential of language models to support end-to-end systems from initial sensing and scene understanding to planning and control [7]. However, their reasoning capacity is often constrained by static pre-training, which restricts real-time adaptability and environmental grounding [18]. To address this, the RAG paradigm has emerged as a promising enhancement, allowing models to proactively query external databases for up-to-date and context-specific knowledge [9]. Systems such as RAG-Driver [45] and RAG-Guided [44] demonstrate how integrating RAG can reduce hallucinations, improve decision consistency, and expand the effective knowledge scope of LLMs in autonomous driving contexts [22].

In contrast to prior knowledge-driven AD that suffers from hallucinations and lacks environmental grounding, our framework anchors reasoning in a structured knowledge pool, providing temporally aligned, interpretable, and query-efficient context. Furthermore, V2X-UniPool enables precise and proactive retrieval of spatiotemporal knowledge tailored to the ego vehicle’s state, ensuring both consistency and real-time responsiveness.

3. Methodology

3.1. Framework Overview

To achieve extended situational awareness and accurate reasoning, we introduce V2X-UniPool, a unified multimodal

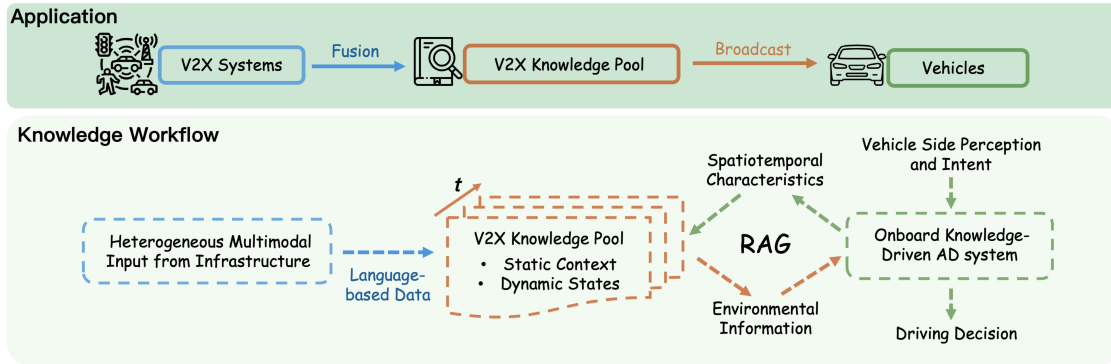


Figure 1. Framework of V2X-UniPool

framework as shown in Fig 1. This framework bridges the gap between raw multimodal sensor data from real-world infrastructure and structured, language-based representations. It continuously aggregates heterogeneous multimodal inputs into a temporally organized V2X Knowledge Pool, while the onboard knowledge-driven AD system interacts with this pool through a RAG mechanism. At each reasoning step, the infrastructure encodes the latest knowledge into a compact language-based snapshot, which is broadcast; the vehicle then obtains environmental information conditioned on its spatiotemporal characteristics. This fused representation supports anticipatory planning, thereby enabling real-time, grounded, and adaptive reasoning in complex traffic scenarios.

The core of this framework is a time-indexed **V2X Knowledge Pool**, which serves as a comprehensive repository for environmental scene understanding. Environmental scene understanding refers to the perception system’s ability to extract, recognize, and interpret features of all surrounding elements relevant to driving [24]. In this paper, the V2X Knowledge Pool formalizes such scene understanding as an integration of language-based representations of the surrounding world, encompassing both static elements (e.g., road geometry, traffic signs, map features) and dynamic elements (e.g., vehicles, traffic lights). Accordingly, the V2X Knowledge Pool is organized into a Static Pool and a Dynamic Pool.

RAG reasoning mechanism is integrated with the V2X Knowledge Pool. At each iteration, the pool broadcasts a language-based snapshot from this pool, and the ego vehicle retrieves relevant environmental information based on timestamp and location. This retrieved knowledge that covers static context and dynamically updated traffic states is then fused into a unified representation together with the local sensory input of the vehicle. The fused context and local perception are subsequently processed by the vehicle-side model to generate the final driving decisions.

3.2. V2X Knowledge Pool Construction

V2X Raw Data Processing To construct a temporally aligned and semantically structured knowledge representation, the raw sensor data are first transformed into language-based data. These raw data are categorized into two modalities: **Unstructured Data**, including camera images and LiDAR point clouds; and **Structured Data**, comprising HD maps, traffic signal phases, average flow speed, vehicle counts, density metrics, and other infrastructure sensors. This focus leverages the global field-of-view and fixed positioning of infrastructure sensors, offering more comprehensive and stable environmental coverage than the ego vehicle’s limited and dynamic perception. All collected raw data are temporally synchronized and spatially calibrated within a unified coordinate frame to ensure cross-modality alignment. The processing pipeline is divided according to data modality and preprocessing requirements:

For Unstructured Data, camera and LiDAR inputs from roadside units are converted into interpretable semantic representations. High-resolution bird’s-eye-view (BEV) images are denoised, standardized, and normalized before being processed by a state-of-the-art vision-language model (GPT-4o 2024). Task-specific prompts guide the model to extract scene-level semantics, outputting structured descriptions that include a `reason` field (natural-language explanation) for interpretability and a `prediction` field to infer short-term state transitions. This reasoning-aware design improves both transparency and utility [4, 26, 31]. In parallel, LiDAR point clouds are geometrically filtered to remove noise and normalized for consistent density and frame alignment. Combined with image semantics, these 3D spatial cues enhance the detection and localization of road users and infrastructure elements, yielding a consistent representation of the traffic scene. For Structured Data, inputs are cleaned, temporally synchronized, and spatially aligned. Relevant fields are parsed, missing values are interpolated or replaced with defaults chosen based on infrastructure priors or domain-specific heuristics, and numeri-

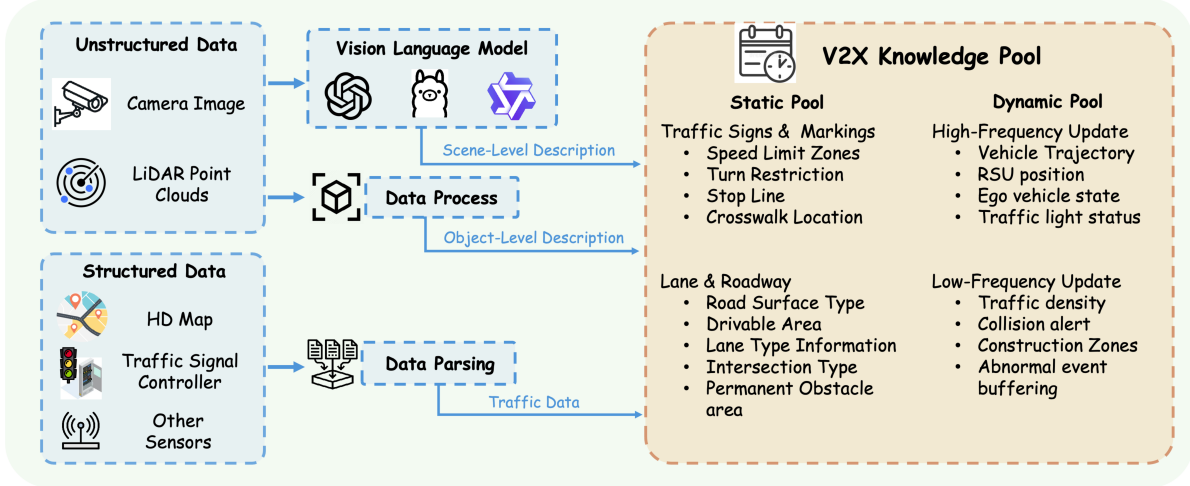


Figure 2. Overview of the V2X Knowledge Pool Construction

cal fields are standardized. These records are then matched with unstructured data by timestamp and location, enabling unified integration.

The resulting dataset comprises temporally aligned, semantically normalized heterogeneous data, where both structured and unstructured modalities are transformed into a unified language-based representation. This forms the basis of the V2X Knowledge Pool, supporting interpretable and efficient reasoning across complex traffic scenarios.

Static Pool The Static Pool of the V2X Knowledge Pool is defined as the collection of infrastructure-level environmental semantics whose temporal dynamics are negligible within operational planning horizons [16]. These elements are effectively time-invariant in practice, as their variations remain below a stability threshold ϵ during short time intervals, thereby providing a reliable basis for reasoning and decision-making across time-aligned traffic scenarios. Formally, a data element $\mathcal{D}_{\text{static}}$ is included in the Static Pool if its feature representation satisfies Eq 1:

$$\left| \frac{\partial f(\mathcal{D}_{\text{static}})}{\partial t} \right| < \epsilon, \quad \forall t \in [t_0, t_0 + T] \quad (1)$$

where $f(\cdot)$ denotes the element’s numerical representation, $\epsilon \ll 1$ is a predefined stability threshold, and T is the planning horizon.

The Static Pool is typically constructed offline or during system initialization for a given region and updated only when long-term infrastructure changes occur (e.g., lane reconfiguration, construction updates, or map revisions). Each entry is represented in a language-based format, enabling interpretable and semantically grounded descriptions of the environment, including traffic signs, lane markings, and roadway geometry.

By maintaining a stable, high-level abstraction of the physical scene, the Static Pool functions as the long-term memory of the reasoning system. It ensures consistent grounding for dynamic perception, supports scenario understanding, and reduces the computational burden of repeatedly processing static semantics in real time.

Dynamic Pool The Dynamic Pool of the V2X Knowledge Pool captures temporally evolving traffic states. Unlike the Static Pool that includes invariant infrastructure, the Dynamic Pool is continuously updated to reflect changes in traffic participants, signal states, and environmental conditions. Formally, a data element $\mathcal{D}_{\text{dynamic}}$ is classified as part of the Dynamic Pool if it violates the constraint of temporal invariance and exhibits perceptible change within the planning horizon as Eq 2:

$$\left| \frac{\partial f(\mathcal{D}_{\text{dynamic}})}{\partial t} \right| \geq \epsilon, \quad \exists t \in [t_0, t_0 + T] \quad (2)$$

where $f(\cdot)$ denotes the numerical feature representation of the element, and ϵ is the minimum change rate required to be considered temporally dynamic. The threshold is task-defined and reflects the system’s sensitivity to environmental variation. This definition ensures that only non-stationary elements requiring timely updates are included in the Dynamic Pool, enabling responsive and adaptive planning.

The Dynamic Pool is constructed online and continuously updated to capture temporally evolving states, and to balance responsiveness and semantic granularity, we divide it into two sub-pools: a **High-Frequency Update Sub-Pool** and a **Low-Frequency Update Sub-Pool**.

High-Frequency Update Sub-Pool contains data streams that change rapidly over short time intervals and are critical for short-term prediction and motion planning.

This includes object trajectories, velocities, types, and traffic light states updated at rates of at least 10 Hz. Formally, a dynamic data instance $\mathcal{D}_{\text{dynamic}}$ is assigned to the High-Frequency Update Sub-Pool if it satisfies Eq 3:

$$\left| \frac{\partial f(\mathcal{D}_{\text{dynamic}})}{\partial t} \right| \geq \epsilon_{\text{high}}, \quad f_s(\mathcal{D}_{\text{dynamic}}) \geq 10 \text{ Hz} \quad (3)$$

where ϵ_{high} denotes the minimum rate of change for highly dynamic content, and $f_s(\cdot)$ represents the sampling frequency. This condition ensures that only fast-evolving, high-resolution data are routed to this sub-pool for real-time planning.

Low-Frequency Update Sub-Pool stores moderately dynamic yet semantically rich information that evolves over slower time scales. This includes aggregated traffic density, collision alerts, construction updates, and abnormal events, typically updated at rates around 1 Hz. Formally, a dynamic data instance $\mathcal{D}_{\text{dynamic}}$ is assigned to the Low-Frequency Update Sub-Pool if it satisfies Eq 4:

$$\epsilon_{\text{low}} \leq \left| \frac{\partial f(\mathcal{D}_{\text{dynamic}})}{\partial t} \right| < \epsilon_{\text{high}}, \quad f_s(\mathcal{D}_{\text{dynamic}}) < 10 \text{ Hz} \quad (4)$$

where ϵ_{low} and ϵ_{high} define the acceptable range of perceptual change for moderate dynamics. Elements with variation below ϵ_{low} are instead assigned to the Static Pool, ensuring a clear separation between static and dynamic semantics. This guarantees that mid-rate signals are preserved for semantic fusion and planning over longer horizons.

Together, these two sub-pools provide a comprehensive multi-resolution temporal representation of the current traffic environment, enabling knowledge-driven AD to anticipate, reason, and respond adaptively under dynamic conditions.

3.3. RAG-based Reasoning

We design a RAG-based collaborative reasoning mechanism that enables the ego vehicle to continuously access infrastructure-side knowledge while preserving temporal and spatial relevance.

Language-based Knowledge Broadcasting The infrastructure encodes the aggregated contents of the Static Pool, the High-Frequency Update Sub-Pool, and the Low-Frequency Update Sub-Pool into a compact language-based snapshot S_t as Eq 5, which is then broadcast to nearby vehicles rather than transmitting the entire pool history:

$$S_t = \text{Encode}\left(P^{\text{Static}} \oplus P^{\text{HF}} \oplus P^{\text{LF}}\right), \quad (5)$$

where P^{Static} , P^{HF} , and P^{LF} denote concatenation-based aggregation from the respective sub-pools, and \oplus represents modality-wise concatenation.

Environmental Information Retrieval Upon receiving the broadcast stream S_t , the ego vehicle filters the knowledge based on its spatiotemporal state. Specifically, it further extracts the information most relevant to its current timestamp t and location l as Eq 6:

$$E_{t,l} = \text{Retrieve}(t, l; S_t), \quad (6)$$

where $\text{Retrieve}(\cdot)$ returns spatiotemporally filtered knowledge such as nearby objects, traffic lights, road conditions, aggregated traffic density, and abnormal events. This design ensures temporal consistency while filtering the knowledge to what is most relevant for the vehicle.

Joint Perception and Reasoning The retrieved environmental information $E_{t,l}$ is then combined with the vehicle’s local sensory input V . In practice, the fusion is implemented through structured prompt construction. The vehicle-side language model then performs reasoning over this fused representation to generate the final driving decision \hat{D} as Eq 7:

$$\hat{D} = \text{LM}(\text{Fuse}(V, E_{t,l})). \quad (7)$$

This RAG-based design ensures that each vehicle always reasons with the most spatiotemporally relevant infrastructure knowledge, while the language-based snapshot representation minimizes communication overhead and provides interpretable semantics for complex urban scenarios.

4. Experiment

4.1. Experiment Settings

We conduct comprehensive evaluations of V2X-UniPool on the DAIR-V2X-Seq dataset [42], a large-scale real-world V2X benchmark featuring both vehicle-side and infrastructure-side sensors. The dataset comprises sequential perception and trajectory forecasting subsets. The sequential perception dataset includes over 15,000 frames across 95 scenarios, collected at 10 Hz from vehicle and roadside sensors. The trajectory forecasting dataset is significantly larger, with 210,000 scenarios from 28 intersections, including 50,000 cooperative-view, 80,000 ego-view, and 80,000 infrastructure-view cases. Each scenario is supplemented with 10-second trajectories, high-definition vector maps, and real-time traffic light signals. This comprehensive setup enables fine-grained exploration of cooperative perception and planning, allowing assessment of infrastructure-enhanced reasoning in diverse urban traffic conditions.

All models follow the OpenEMMA-style planning formulation [34], where each input consists of front-view images and a history of ego vehicle speeds and curvatures, and the output is a prediction of future speed and curvature vectors. This design reflects real-world driving requirements

Table 1. Planning Evaluation Results. V2X-UniPool integrates external knowledge via structured RAG reasoning. The vehicle-side model is based on Qwen-2.5-vl-7B (lower is better).

Method	L2 Error (m) ↓				Collision Rate (%) ↓				Transmission Cost ↓
	2.5s	3.5s	4.5s	Avg.	2.5s	3.5s	4.5s	Avg.	
V2VNet [30]	2.31	3.29	4.31	3.30	0.00	1.03	1.47	0.83	8.19×10^7
CooperNaut [8]	3.83	5.26	6.69	5.26	0.59	1.92	1.63	1.38	8.19×10^7
UniV2X - Vanilla [43]	2.21	3.31	4.46	3.33	0.15	0.89	2.67	1.24	8.19×10^7
UniV2X [43]	2.60	3.44	4.36	3.00	0.00	0.74	0.49	0.41	8.09×10^5
V2X-VLM [41]	1.09	1.12	1.42	1.21	0.02	0.03	0.03	0.03	1.24×10^7
V2X-UniPool	0.95	1.44	2.02	1.47	0.00	0.02	0.05	0.04	1.51×10^5

and supports precise evaluation of planning performance. For the experiment, we construct over 10,000 ego-view scenarios from DAIR-V2X-Seq, each spanning 10 seconds and formatted as structured planning prompts with historical motion states and aligned future trajectory labels. During inference, all vehicle-side models receive the same V2X-UniPool RAG retrieval results, ensuring that observed performance differences are solely attributed to the models’ reasoning ability over external semantic context.

Beyond the dataset setup and input–output design, our experiments first benchmark V2X-UniPool against representative planning baselines, jointly evaluating planning accuracy, safety, and communication efficiency together with a detailed latency analysis of the infrastructure-to-vehicle pipeline. We then assess generality through ablations across diverse vehicle-side models and examine scaling behavior within the Qwen-2.5-VL family. Together, these results provide a comprehensive evaluation of accuracy, safety, efficiency, and scalability for V2X-UniPool.

4.2. Experiment Results

Table 1 reports the overall planning and communication results. **V2X-UniPool** achieves the best short-horizon planning accuracy (**0.95 m at 2.5s**) and sustains competitive average error (1.47 m), comparable to the state-of-the-art V2X-VLM (1.21 m). In terms of safety, it consistently maintains a near-zero collision rate (**0.00%** at 2.5s, **0.02%** at 3.5s, average 0.04%), matching the reliability of V2X-VLM (0.03%). These results highlight that V2X-UniPool is capable of delivering both accurate and safe planning, even under challenging horizons.

Beyond accuracy and safety, V2X-UniPool establishes a new benchmark in efficiency. It reduces transmission cost to only **1.51×10^5** bytes, which is over two orders of magnitude smaller than V2X-VLM (1.24×10^7) and three orders lower than traditional cooperative perception approaches (8.19×10^7). This drastic reduction makes V2X-UniPool significantly more deployable under realistic bandwidth constraints.

To further quantify the practical efficiency of this de-

sign, we conduct a detailed latency breakdown analysis. This evaluation decomposes the infrastructure-to-vehicle pipeline into its constituent stages, allowing us to measure the contribution of raw data preprocessing, temporal indexing, communication, and vehicle-side retrieval. These results demonstrate that by transmitting compact language-based knowledge snapshots, V2X-UniPool not only achieves the lowest communication cost, but also realizes a lightweight end-to-end pipeline, incurring negligible overhead beyond data preprocessing.

Table 2 reports the latency breakdown of V2X-UniPool. The total processing takes approximately **302 ms**, where the dominant component is raw V2X data preprocessing (**250 ms, 82.8%**). Notably, this latency reflects the maximum among parallel sensor streams, since preprocessing for different modalities is executed concurrently. Knowledge pool operations, including temporal indexing and updates, add only **40 ms (13.2%)**, while vehicle-side retrieval and prompt construction incur negligible overhead (<0.1%). In terms of communication, with 5G NR-V2X links (typical throughput approaching 100 Mbps), V2X-UniPool requires only **12 ms per broadcast (4.0%)**. On the vehicle side, the latency for reasoning depends on the deployed model size. Nonetheless, V2X-UniPool consistently ensures that the *infrastructure-to-vehicle* operations remain lightweight, leaving the computation–latency trade-off mainly determined by the chosen vehicle-side model.

4.3. Ablation Study

To systematically evaluate the effectiveness of the proposed V2X-UniPool, two complementary ablation studies are designed with the goal of assessing its benefits across both *different model architectures* and *different model scales*. (i) *Cross-model evaluation*: We compare the planning performance of representative vehicle-side language models with and without V2X-UniPool integration. The evaluation spans both proprietary and open-source paradigms, including **GPT-4o (2024)**, **GPT-4.1 Mini**, **Gemini-2.0**, and **Llama-3.2**. (ii) *Scaling behavior*: To examine how V2X-UniPool interacts with model capacity, we further conduct

Table 2. Latency breakdown of V2X-UniPool. The total latency corresponds to the infrastructure-to-vehicle processing time for one pipeline, including communication (estimated with 100 Mbps link).

Process	Description	Latency (ms)	Proportion (%)
Preprocessing	V2X Raw Data Processing (parallel, max sensor)	~250	82.8
Knowledge Pool Ops	Temporal indexing + pool update	~40	13.2
Communication (100 Mbps)	Transmission delay per broadcast	~12	4.0
Vehicle Extraction	Query relevant pool entries (retrieval)	≈0.0	<0.1
Prompt Construction	Build structured input for vehicle-side model	≈0.0	<0.1
Total	–	~302	100.0

experiments across multiple scales of the same architecture. Specifically, we adopt the **Qwen-2.5-VL** family at four parameter sizes (**3B**, **8B**, **32B**, and **72B**) to analyze scaling trends under both baseline and V2X-UniPool enhanced settings. This two-fold design enables us to disentangle the effects of model scaling from those of structured external knowledge, and to demonstrate the generality and scalability of V2X-UniPool.

While collision rate is a critical safety indicator, for most strong models it already remains near zero, leaving limited room for further reduction. Therefore, we additionally introduce a **Comfort Score** to better capture improvements in trajectory smoothness and ride quality. Beyond safety, passenger experience is a critical factor for autonomous vehicles, as the primary stakeholders are the passengers themselves. To capture the smoothness of the trajectory and the comfort of the ride, we adopt a differentiable metric based on longitudinal acceleration and jerk, two key factors widely recognized in studies of transportation safety and comfort [1, 3, 36]. Here, \ddot{v} denotes longitudinal acceleration and $\ddot{\ddot{v}}$ is jerk (the rate of change of acceleration). The averages are computed over n predicted future frames. Following prior literature, we set $\alpha = 1.0$ and $\beta = 1.0$ to balance the contributions of acceleration and jerk. The resulting score lies within $[0, 1]$, where higher values indicate smoother and more comfortable planned trajectories. The score function is defined as Eq 8:

$$\text{ComfortScore} = 1 - \tanh(\alpha \cdot |\ddot{v}| + \beta \cdot |\ddot{\ddot{v}}|) \quad (8)$$

From Table 3, it is clear that **V2X-UniPool consistently enhances all evaluated models across the three key metrics**. On average, the *L2 Error* decreases by about **1–8%** across horizons (mean: 4–5%), confirming more accurate trajectory generation. The *Comfort Score* improves steadily for most models, with average gains of **0.035**, reflecting smoother driving behavior. The most pronounced effect, however, is observed in the *Collision Rate*: while GPT-4o and Gemini-2.0 achieve modest reductions, Llama-3.2 shows the most striking improvement, with its collision rate dropping from 0.11 to 0.03 at the 4.5s horizon equivalent

to a **73% reduction**. This indicates that V2X-UniPool not only improves planning accuracy, but also provides substantial safety benefits. Overall, the ablation results highlight that the integration of V2X-UniPool leads to more precise trajectories, significantly lower collision risks, and improved comfort, with safety improvement being the most prominent gain.

Table 4. Overall Average ADE comparison of Qwen models with and without V2X-UniPool (lower is better).

Model	Base	+V2X-UniPool ↓
Qwen-3B	1.6482	1.5851
Qwen-7B	1.3699	1.3456
Qwen-32B	1.3803	1.2671
Qwen-72B	1.3241	1.2998

We evaluate the effect of V2X-UniPool on model scaling using the Overall Average ADE across same 100 driving scenarios. As shown in Table 4, incorporating V2X-UniPool consistently reduces planning error across all model sizes. For example, Qwen-3B with UniPool improves from 1.65 to 1.59, effectively narrowing the gap to the 7B base model (1.37). Similarly, Qwen-32B with UniPool achieves 1.27, surpassing the performance of the 72B base model (1.32). These results highlight that V2X-UniPool enables smaller models to approach or even exceed the performance of much larger counterparts, thereby achieving near state-of-the-art accuracy at substantially lower computational cost.

Building on these error-based comparisons, we further examine how improvements propagate into downstream driving quality metrics, including comfort and collision. Figure 3 summarizes these effects across the Qwen-2.5-VL family. Overall, **V2X-UniPool brings consistent benefits across all model scales**, with the most notable gains achieved by **Qwen-3B** and **Qwen-32B**. Among the evaluation metrics, *Error* and *Comfort* show the most substantial improvements, indicating that external structured knowledge not only enhances prediction accuracy but also con-

Table 3. Comprehensive ablation study on different vehicle-side models with and without V2X-UniPool. Metrics include **L2 Error** (m, lower is better), **Collision Rate** (% , lower is better), and **Comfort Score** (higher is better), reported at 2.5s, 3.5s, and 4.5s horizons.

Model	Metric	w/o V2X-UniPool			w/ V2X-UniPool		
		2.5s	3.5s	4.5s	2.5s	3.5s	4.5s
GPT-4o (2024)	Error ↓	1.25	1.93	2.74	1.18	1.81	2.56
	Collision ↓	0.00	0.00	0.05	0.00	0.00	0.04
	Comfort ↑	0.52	0.54	0.56	0.63	0.65	0.66
GPT-4.1 Mini	Error ↓	1.45	2.21	3.11	1.41	2.18	3.10
	Collision ↓	0.00	0.00	0.10	0.00	0.00	0.01
	Comfort ↑	0.58	0.61	0.62	0.59	0.61	0.63
Gemini-2.0	Error ↓	1.37	2.18	3.17	1.29	2.01	2.92
	Collision ↓	0.00	0.05	0.05	0.00	0.04	0.04
	Comfort ↑	0.30	0.33	0.34	0.33	0.36	0.37
Llama-3.2	Error ↓	1.26	1.92	2.72	1.24	1.90	2.69
	Collision ↓	0.07	0.07	0.11	0.02	0.03	0.03
	Comfort ↑	0.53	0.56	0.58	0.51	0.56	0.59

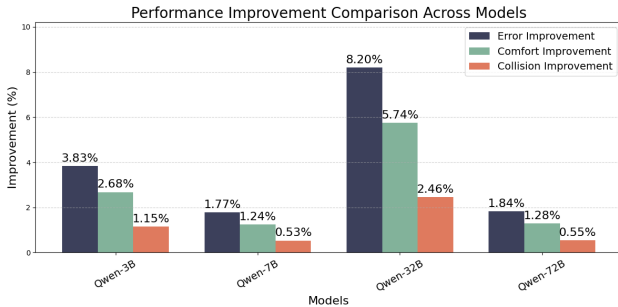


Figure 3. Relative improvements with V2X-UniPool across the Qwen-2.5-VL family. **Left:** Error reduction (%). **Middle:** Comfort improvement (%). **Right:** Collision reduction (%).

tributes to smoother and more stable driving behavior, while collision reductions are comparatively smaller due to the already low baseline rates.

5. Conclusion

This study presents **V2X-UniPool**, a unified framework for knowledge-driven AD system. By translating raw multimodal V2X signals into structured, language-based knowledge and organizing them in a temporally indexed pool, V2X-UniPool enables contextualized, RAG-based reasoning on the vehicle side. Experiments on the real-world DAIR-V2X dataset show that it improves planning accuracy and safety while reducing communication cost by more than 80%, establishing a new paradigm that integrates V2X collaboration with language-model reasoning. The results further highlight its practical potential, achieving state-of-

the-art accuracy and safety with the lowest communication overhead and small end-to-end latency, while consistently benefiting models of different architectures and scales to ensure strong scalability.

Limitations and broader impacts. V2X-UniPool is designed as an augmentation layer: it consistently improves planning quality, but its complete latency and final performance inevitably depend on the capacity of the underlying AD model. While our current evaluation is based on open-loop experiments, the next step is to advance toward closed-loop validation. Encouraged by the positive feedback from current results, we are in the process of deploying V2X-UniPool in a closed-loop setting to further assess its robustness under interactive traffic dynamics. We expect this study to advance the integration of V2X systems with knowledge-driven AD, paving the way for safer, more cost-effective, and highly scalable driving applications.

References

- [1] Mechanical vibration and shock – evaluation of human exposure to whole-body vibration – part 1: General requirements, 1997. 7
- [2] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3): 1852–1864, 2020. 2
- [3] Il Bae, Sangjae Moon, and Seung-Woo Seo. Self-driving like a human driver instead of a robocar: Personalized comfortable driving experience. *arXiv preprint arXiv:2001.03908*, 2020. 7

- [4] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1, 3
- [5] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524. IEEE, 2019. 2
- [6] Hsu-kuang Chiu, Ryo Hachiuma, Chien-Yi Wang, Stephen F Smith, Yu-Chiang Frank Wang, and Min-Hung Chen. V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. *arXiv preprint arXiv:2502.09980*, 2025. 2
- [7] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024. 2
- [8] Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, and Yuke Zhu. Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17252–17262, 2022. 6
- [9] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501, 2024. 2
- [10] Xiangbo Gao, Yuheng Wu, Rujia Wang, Chenxi Liu, Yang Zhou, and Zhengzhong Tu. Langcoop: Collaborative driving with language. *arXiv preprint arXiv:2504.13406*, 2025. 2
- [11] Xiangbo Gao, Runsheng Xu, Jiachen Li, Ziran Wang, Zhiwen Fan, and Zhengzhong Tu. Stamp: Scalable task and model-agnostic collaborative perception. *arXiv preprint arXiv:2501.18616*, 2025. 2
- [12] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. Drivemllm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *arXiv preprint arXiv:2411.13112*, 2024. 2
- [13] Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2024. 1
- [14] Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE Intelligent Transportation Systems Magazine*, 2023. 2
- [15] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 1
- [16] Tao Huang, Jianan Liu, Xi Zhou, Dinh C Nguyen, Mostafa Rahimi Azghadi, Yuxuan Xia, Qing-Long Han, and Sumei Sun. V2x cooperative perception for autonomous driving: Recent advances and challenges. *arXiv preprint arXiv:2310.03525*, 2023. 1, 2, 4
- [17] Henry Alexander Ignatious, Manzoor Khan, et al. An overview of sensors in autonomous vehicles. *Procedia Computer Science*, 198:736–741, 2022. 1
- [18] Xin Li, Yeqi Bai, Pinlong Cai, Licheng Wen, Daocheng Fu, Bo Zhang, Xuemeng Yang, Xinyu Cai, Tao Ma, Jianfei Guo, et al. Towards knowledge-driven autonomous driving. *arXiv preprint arXiv:2312.04316*, 2023. 2
- [19] Si Liu, Chen Gao, Yuan Chen, Xingyu Peng, Xianghao Kong, Kun Wang, Runsheng Xu, Wentao Jiang, Hao Xiang, Jiaqi Ma, et al. Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. *arXiv preprint arXiv:2308.16714*, 2023. 2
- [20] Jia Quan Loh, Xuewen Luo, Fan Ding, Hwa Hui Tew, Junn Yong Loo, Ze Yang Ding, Susilawati Susilawati, and Chee Pin Tan. Cross-domain transfer learning using attention latent features for multi-agent trajectory prediction, 2024. 2
- [21] Xuewen Luo, Fan Ding, Yinsheng Song, Xiaofeng Zhang, and Junnyong Loo. Pkrd-cot: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving. *arXiv preprint arXiv:2412.02025*, 2024. 2
- [22] Xuewen Luo, Chenxi Liu, Fan Ding, Fengze Yang, Yang Zhou, Junnyong Loo, and Hwa Hui Tew. Senserag: Constructing environmental knowledge bases with proactive querying for llm-based autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 989–996, 2025. 2
- [23] Lili Miao, Shang-Fu Chen, Yu-Ling Hsu, and Kai-Lung Hua. How does c-v2x help autonomous driving to avoid accidents? *Sensors*, 22(2):686, 2022. 2
- [24] Khan Muhammad, Tanveer Hussain, Hayat Ullah, Javier Del Ser, Mahdi Rezaei, Neeraj Kumar, Mohammad Hiji, Paolo Bellavista, and Victor Hugo C de Albuquerque. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22694–22715, 2022. 3
- [25] Vandana Narri, Amr Alanwar, Jonas Mårtensson, Christoffer Norén, Laura Dal Col, and Karl Henrik Johansson. Set-membership estimation in shared situational awareness for automated vehicles in occluded scenarios. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 385–392. IEEE, 2021. 2
- [26] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019. 3
- [27] Shunli Ren, Siheng Chen, and Wenjun Zhang. Collaborative perception for autonomous driving: Current status and future trend. In *Proceedings of 2021 5th Chinese Confer-*

- ence on Swarm Intelligence and Cooperative Control, pages 682–692. Springer, 2022. 2
- [28] Shiva Sreeram, Tsun-Hsuan Wang, Alaa Maalouf, Guy Rosman, Sertac Karaman, and Daniela Rus. Probing multi-modal llms as world models for driving. *arXiv preprint arXiv:2405.05956*, 2024. 1
- [29] Rujia Wang, Xiangbo Gao, Hao Xiang, Runsheng Xu, and Zhengzhong Tu. Cocmt: Communication-efficient cross-modal transformer for collaborative perception. *arXiv preprint arXiv:2503.13504*, 2025. 2
- [30] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pages 605–621. Springer, 2020. 2, 6
- [31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3
- [32] Licheng Wen, Daoheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023. 1
- [33] Zhu Xiao, Jinmei Shu, Hongbo Jiang, Geyong Min, Hongyang Chen, and Zhu Han. Overcoming occlusions: Perception task-oriented information sharing in connected and autonomous vehicles. *IEEE Network*, 37(4):224–229, 2023. 2
- [34] Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemma: Open-source multimodal model for end-to-end autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1001–1009, 2025. 5
- [35] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *Computer Vision – ECCV 2022*, 2022. 1, 2
- [36] Fan Yan, Xiangjun Wang, Lianbo Zhang, and Yan Chen. Comfort-based trajectory and velocity planning for autonomous vehicles. *International Journal of Automotive Technology*, 22(5):1369–1380, 2021. 7
- [37] Xun Yang, Yunyang Shi, Jiping Xing, and Zhiyuan Liu. Autonomous driving under v2x environment: state-of-the-art survey and challenges. *Intelligent Transportation Infrastructure*, 1:liac020, 2022. 2
- [38] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*, 2023. 1, 2
- [39] Zhenwei Yang, Jilei Mao, Wenxian Yang, Yibo Ai, Yu Kong, Haibao Yu, and Weidong Zhang. Lidar-based end-to-end temporal perception for vehicle-infrastructure cooperation. *arXiv preprint arXiv:2411.14927*, 2024. 1
- [40] Takahito Yoshizawa, Dave Singelée, Jan Tobias Muehlberg, Stephane Delbruel, Amir Taherkordi, Danny Hughes, and Bart Preneel. A survey of security and privacy issues in v2x communication systems. *ACM Computing Surveys*, 55(9): 1–36, 2023. 1
- [41] Junwei You, Haotian Shi, Zhuoyu Jiang, Zilun Huang, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. *arXiv preprint arXiv:2408.09251*, 2024. 6
- [42] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023. 5
- [43] Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-end autonomous driving through v2x cooperation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9598–9606, 2025. 6
- [44] Jun Yu, Yunxiang Zhang, Zerui Zhang, Zhao Yang, Gongpeng Zhao, Fengzhao Sun, Fanrui Zhang, Qingsong Liu, Jianqing Sun, Jiaen Liang, et al. Rag-guided large language models for visual spatial description with adaptive hallucination corrector. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11407–11413, 2024. 2
- [45] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024. 2
- [46] Syed Adnan Yusuf, Arshad Khan, and Riad Souissi. Vehicle-to-everything (v2x) in the autonomous vehicles domain—a technical review of communication, sensor, and ai technologies for road user safety. *Transportation Research Interdisciplinary Perspectives*, 23:100980, 2024. 1
- [47] Yuxuan Zhu, Shiyi Wang, Wenqing Zhong, Nianchen Shen, Yunqi Li, Siqi Wang, Zhiheng Li, Cathy Wu, Zhengbing He, and Li Li. Will large language models be a panacea to autonomous driving? *arXiv preprint arXiv:2409.14165*, 2024. 2