HMF-Mapnet: Hybrid Memory Fusion for Accurate and Robust Vectorized Map Perception

1st Huangnan Zheng
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
zhenghuangnan@zju.edu.cn, ORCID: 0009-0008-4822-3533

2nd Ruihang Li
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
12221089@zju.edu.cn, ORCID: 0009-0008-1608-1913

3rd Heng Chen
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
loundch@163.com, ORCID: 0009-0004-6480-3395

4th Kehan Wang
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
22421179@zju.edu.cn, ORCID: 0009-0000-8960-1002

5th Wangliang Guo
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
wl_guo@zju.edu.cn, ORCID: 0009-0005-2438-1025

6th Zhijie Pan
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
zhijie_pan@zju.edu.cn, ORCID: 0000-0002-3466-7423

***Abstract*—The accurate and efficient Vectorized high-definition (HD) Map perception is a fundamental prerequisite for reliable planning and decision-making in autonomous driving. To enhance the perception of complex road geometries and temporal dynamics, dominant paradigms either leverage explicit map priors or employ sequential modeling. Nevertheless, these two methods suffer from inherent weaknesses: the former is highly sensitive to prior quality, while the latter is prone to error accumulation, both of which ultimately limit the model's accuracy and robustness. To this end, we propose HMF-MapNet, a robust framework featuring a parallel dual-path architecture that achieves an optimal trade-off between efficiency and performance. Specifically, the geometric path utilizes a historical map bank that interfaces directly with the BEV feature map to bolster its geometric reasoning. Meanwhile, the temporal path employs a lightweight recurrent mechanism that iteratively decodes static map data over time, robustly separating static and dynamic elements.HMF-MapNet not only significantly reduces memory usage but also achieves a remarkable +6.3 mAP gain over SOTA methods under extreme rotation noise ($\sigma_r = 0.1$), proving its reliability for real-world deployment. Extensive experiments on nuScenes and Argoverse2 show that our method is remarkably superior to other existing methods.**

***Index Terms*—High-Definition Map Perception, Autonomous Driving, Memory Fusion, Temporal Modeling**

I. Introduction

High-Definition (HD) maps [1]are fundamental to autonomous driving safety, supplying the precise data needed for reliable navigation [2]and decision-making [3]in complex environments. Achieving super-human driving performance depends directly on this level of accurate environmental perception.

However, achieving such perception through online map

This work was supported by the "Pioneer" and "Leading Goose" + X RD Program of Zhejiang (2025C02246(SD2)) and "Pioneer" and "Leading Goose" RD Program of Zhejiang (2024C01016) .

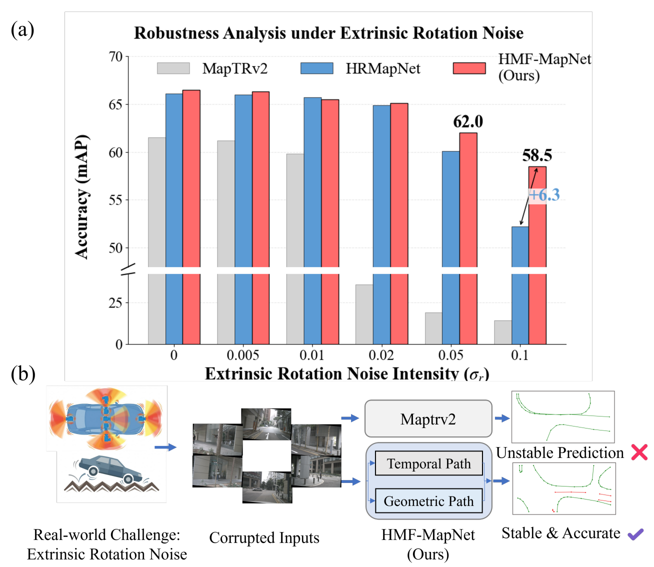

Fig. 1. Robustness Analysis. (a) Quantitative results confirm superior robustness. We outperform SOTA methods significantly, achieving a +6.3 mAP lead under extreme noise ($\sigma_r = 0.1$). (b) Sensor jitter causes geometric fractures in baselines like MapTRv2. Our HMF-MapNet leverages Temporal and Geometric Paths to ensure stable predictions.

perception models faces two formidable challenges in real-world traffic scenarios: handling complex road geometries and managing temporal dynamics. [4], [5]The former is manifested in non-standard structures such as irregular intersections, multi-lane merges, and curved ramps, which demand that the model possesses robust spatial-geometric reasoning capabilities. [6]The latter stems from the motion of other road agents, such as vehicles and pedestrians. [1] Not only do these dynamic objects cause occlusions of static road elements, but they also introduce significant noise and ambiguity into the online map construction process. [1] Addressing this fundamental challenge requires the model to effectively separate static infrastructure from dynamic objects within continuous temporal sequences.

To address these challenges, dominant research paradigms have primarily evolved along two technical pathways. The first approach is to leverage explicit map priors. These methods typically incorporate a pre-built, coarse map [7], [8] or topological structure [9], [10] as prior knowledge to enhance the model's comprehension and representation of complex geometries. The second approach employs sequential modeling. [6], [11] These methods utilize architectures like Recurrent Neural Networks (RNNs) [12] or temporal Transformers [11], [13] to process continuous sensor data streams, aiming to filter out interference from dynamic objects and infer occluded static elements through temporal modeling.

Nevertheless, both technical pathways suffer from inherent limitations. Methods based on map priors are highly sensitive to the quality of the prior information [8]; an inaccurate or outdated prior map can mislead the model, leading to erroneous perception results.Furthermore, sequential modeling methods are prone to error accumulation [11]. In this process, minor prediction errors from previous frames are propagated and amplified over time, ultimately causing performance degradation in long sequences.Ultimately, both deficiencies severely constrain the model's accuracy and robustness in real-world, complex scenarios.

Crucially, these weaknesses are further amplified in real-world deployment, where stable sensors cannot be assumed. Mechanical vibrations, uneven terrain (e.g., speed bumps), and calibration drift often introduce extrinsic rotation noise. Under such perturbations, map-prior methods fail because the projected prior becomes misaligned with the visual evidence. Sequential models are even more fragile, as minor feature misalignments caused by frame-wise jitter can accumulate and be amplified through recurrent updates, eventually leading to global geometric fractures. As illustrated in Fig. 1 (b), existing paradigms lack sufficient robustness against transient disturbances, suffering catastrophic degradation with accuracy drops of over 40% under strong interference.

To address these limitations, we propose HMF-MapNet, a novel framework based on decoupled perception. Inspired by human navigation, which combines long-term structural memory with short-term observation, HMF-MapNet explicitly separates geometric retention from temporal smoothing. Unlike existing methods that entangle static and dynamic features in a single stream, our framework adopts a geometric-temporal dual-path design. By decoupling invariant structural geometry from transient feature fluctuations, it mitigates the impact of instantaneous sensor noise and provides a robust basis for perception in complex real-world environments.

Specifically, the geometric path builds a probabilistic Historical Map Bank as a stable structural anchor. Instead of conventional binary masking, we employ a log-odds occupancy mechanism to accumulate historical confidence, yielding what we term temporal inertia. This mechanism enables high-confidence structural memory to absorb transient sensor jitter and suppress geometric outliers. As a result, the model preserves consistent structural reasoning even under extrinsic misalignment, providing reliable and noise-resistant priors for subsequent decoding.

Our main contributions can be summarized as follows:

- We identify fundamental limitations in dominant online mapping paradigms: map-prior methods are vulnerable to inaccurate priors, while sequential models suffer from error accumulation.
- We propose HMF-MapNet, a novel framework with a parallel geometric-temporal dual-path architecture that synergistically enhances both geometric reasoning and temporal modeling.Its probabilistic historical bank effectively filters sensor jitter.
- Extensive experiments validate our superiority. HMF-MapNet achieves the optimal accuracy-efficiency trade-off and demonstrates remarkable robustness, outperforming SOTA methods by +6.3 mAP under extreme rotation noise ($\sigma_r = 0.1$).

II. Related Work

A. BEV Representation and Temporal Perception

Recent vision-based autonomous driving increasingly adopts Bird's-Eye-View (BEV) representations for unified multi-camera perception and geometry-aware scene understanding. BEVFormer [14] introduces spatial attention for BEV feature construction, while StreamMapNet [6] improves temporal consistency with historical BEV alignment. Depth-aware multi-camera perception further strengthens spatial reasoning, as shown by M2Depth [15] and CFD [16]. In parallel, occupancy and semantic scene understanding methods [17]–[19] highlight the value of geometry-aware and temporally coherent representations for online map perception.

B. Online Vectorized Map Construction

Online mapping aims to directly predict vectorized map elements from sensor inputs. VectorMapNet [20] pioneers polyline prediction, and MapTR [1] improves it with permutation-equivalent point-set modeling. Subsequent methods enhance geometric accuracy and structural consistency through differentiable rasterization [21], mask-guided learning [22], interaction modeling [23], and stream query denoising [11]. However, these methods still struggle under occlusion, dynamic interference, and sensor noise.

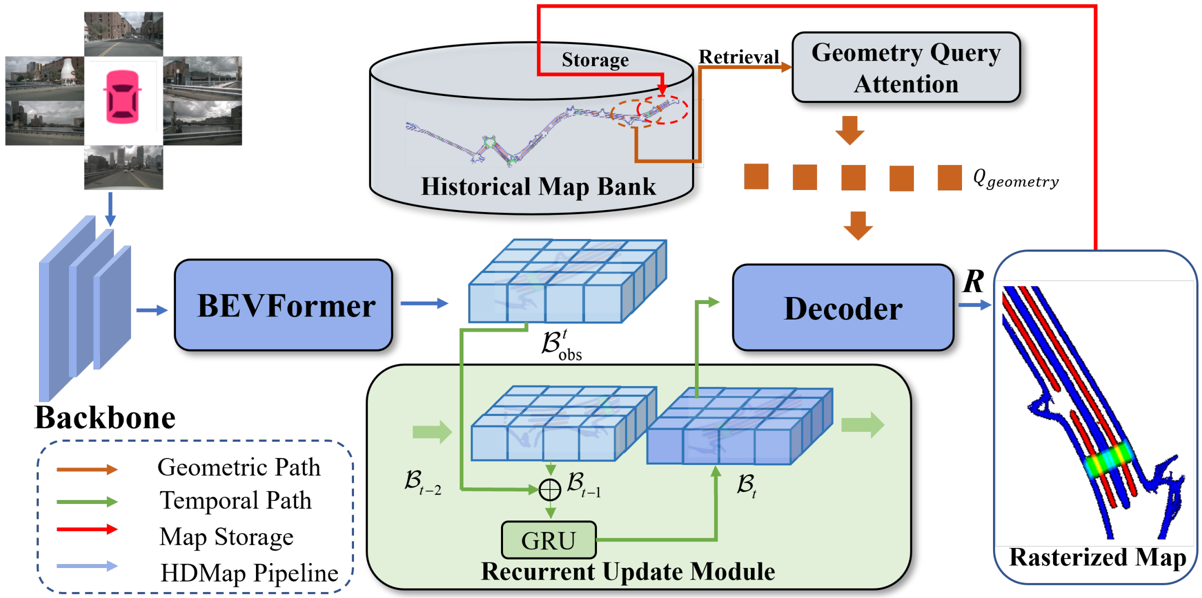

Fig. 2. Proposed HMF-Mapnet architecture.Our framework employs a dual-path architecture that concurrently refines BEV features with a recurrent (GRU) module for temporal smoothing, while injecting geometric priors from a historical map bank via attention to guide the decoder. ***R*** denotes rasterization.

C. Prior-Augmented and Memory-Based Mapping

To improve robustness, recent methods incorporate structural priors or historical memory into online mapping. HRMapNet [9] enhances temporal stability through recurrent HD map fusion. Related studies in 3D perception show that prior modeling also improves geometric and semantic consistency [17], [18]. Cross-domain studies also indicate that semantic guidance, geometric consistency, and structured representation learning can improve model robustness and feature alignment [24], [25]. More generally, structured representation learning methods [26], [27] suggest that well-designed priors can improve stability and expressivity. Different from prior-based or single-path memory methods, our HMF-MapNet uses a dual-path hybrid memory to capture long-term consistency and short-term details without external priors.

III. Method

A. Architecture Overview

The overall architecture of HMF-MapNet is illustrated in Fig. 2. We build our framework upon the foundational principles of advanced end-to-end vectorized map learning paradigms [1], inheriting their strengths in permutation-equivalent modeling and hierarchical query mechanisms while introducing critical architectural innovations for robustness. **Feature Extraction and Transformation.** The pipeline commences with a shared visual backbone (ResNet-50 [28] initialized with ImageNet weights) equipped with a Feature Pyramid Network (FPN) [29] to extract multi-scale 2D image features. These features are then lifted into a unified 3D Bird's-Eye-View (BEV) representation using a view transformer based on deformable attention mechanisms [14]. This results in a comprehensive feature map $\mathcal{B}_{obs}$ (typically 100×50) covering a perception range of $\pm$30m by $\pm$15m, which serves as the fundamental input for subsequent processing.

Design Philosophy: Decoupling Geometry and Dynamics. While existing single-stream baselines achieve impressive performance under ideal conditions, they inherently couple geometric structural information with temporal dynamic features. We identify this coupling as a significant vulnerability: when sensors suffer from extrinsic noise or jitter, the entangled features degrade globally, leading to geometric fractures. To address this, HMF-MapNet introduces a novel **Parallel Dual-Path Architecture** designed to strategically decouple these two aspects:

- The **Geometric Path** (Top branch) acts as a static anchor. It maintains a probabilistic Historical Map Bank to provide long-term, noise-resistant geometric priors, isolating structural reasoning from instantaneous sensor fluctuations.
- The **Temporal Path** (Bottom branch) acts as a dynamic smoother. It employs a recurrent update mechanism to model motion and consistency across frames, filtering out high-frequency disturbances.

Synergistic Decoding. Crucially, these two paths do not operate in isolation; they synergistically guide the decoding stage. The geometric path injects context-rich priors into the initial map queries via cross-attention, while the temporal path provides temporally refined BEV features. Finally, a 6-layer Transformer decoder [30] processes these enhanced inputs to predict N map instances. Following standard vectorized output formulations, each instance is modeled as a point set sequence (e.g., 20 points), enabling end-to-end regression without post-processing.

B. Geometric Path

The Geometric Path serves as the **Static Anchor** of our dual-path system. Our design is driven by a core insight: while visual observations are transient and susceptible to noise, the physical road structure is immutable and persistent. Therefore, rather than reconstructing the map "from scratch" at every frame, we aim to maintain a probabilistic memory of the environment—similar to how human drivers rely on road familiarity. This philosophy necessitates a representation that is both **cumulative** (to build confidence over time) and **inertial** (to resist sudden perturbations). To materialize this concept, we introduce two key components: the Historical Map Bank for robust storage and Geometric Query Attention for efficient retrieval.

1) Historical Map Bank Construction: To build this persistent memory, we employ a global semantic grid, $\mathcal{M}_{\text{global}}$. Instead of storing raw probabilities, each cell g stores an occupancy *log-odds* value $\mathcal{L}(g)$. We adopt the log-odds representation because it transforms complex Bayesian probability multiplication into efficient additive updates, which is computationally ideal for online inference.

The Additive Update Rule. When a new BEV observation aligns with the grid, we update the belief state of each cell. This process allows the map to incrementally accumulate confidence. The update rule is formulated as:

$$\mathcal{L}(g) \leftarrow \text{clamp}\left(\mathcal{L}(g) + m \cdot \lambda^{+} + (1-m)\cdot \lambda^{-}\right) \quad (1)$$

This formula can be decomposed into three logical components:

- **Prior State $\mathcal{L}(g)$:** The accumulated belief from all previous frames, representing the historical memory.
- **Hit Reward $(m \cdot \lambda^{+})$:** If the current sensor detects an element ($m = 1$), we add a positive score λ^{+} to boost confidence.
- **Miss Penalty $((1-m)\cdot\lambda^{-})$:** If the sensor sees empty space ($m = 0$), we add a negative penalty λ^{-} to gradually reduce confidence.

Finally, the value is clamped to a range $[\mathcal{L}_{\min}, \mathcal{L}_{\max}]$ to prevent confidence from exploding to infinity, ensuring the map retains the plasticity to adapt to environmental changes.

For downstream perception, we convert this continuous belief back into a binary decision to form the local map prior $\mathcal{M}_{\text{local}}$:

$$\mathcal{M}_{\text{local}}(g) = \begin{cases} 1 & \text{if } \mathcal{L}(g) > \theta_{\text{occ}} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

2) Robustness Analysis: The Inertia Mechanism: The core innovation of our Historical Map Bank is the physical property of "Temporal Inertia" provided by the log-odds mechanism. Unlike standard metrics, log-odds acts as a **Geometric Stabilizer**, creating a "confidence buffer" that resists sudden changes.

Physical Interpretation: The Confidence Buffer. Consider a scenario where severe vehicle vibration causes extrinsic rotation noise, leading to a lane boundary "drifting" by 2 meters.

- *In conventional networks:* This corrupted frame directly overwrites the output, causing a geometric fracture.
- *In HMF-MapNet:* A cell g at the true lane position likely possesses a saturated value (e.g., $\mathcal{L}(g) \approx \mathcal{L}_{\max} = 5.0$) from history. The transient drift results in a single "miss" penalty ($\lambda^{-} = -0.5$). The updated value becomes $5.0 - 0.5 = 4.5$. Since 4.5 is still far above the threshold θ_{occ} (typically 0), the cell remains occupied.

Essentially, the model "absorbs" the penalty using its accumulated confidence buffer, ignoring the transient noise.

Theoretical Stability Bound. We formally derive how much noise the system can withstand. Let N_{fail} be the number of consecutive erroneous frames required to deplete the confidence buffer and flip a stable state (from Occupied to Free).

The state flips only when the accumulated penalties overwhelm the current confidence $\mathcal{L}_{\text{curr}}$:

$$\mathcal{L}_{\text{curr}} + N_{\text{fail}} \cdot \lambda^{-} \leq \theta_{\text{occ}} \quad (3)$$

Rearranging this inequality gives us the stability lower bound:

$$N_{\text{fail}} \geq \left\lceil \frac{\theta_{\text{occ}} - \mathcal{L}_{\text{curr}}}{\lambda^{-}} \right\rceil \quad (4)$$

With typical settings ($\mathcal{L}_{\max} = 5.0, \theta_{\text{occ}} = 0, \lambda^{-} = -0.5$), the system requires $N_{\text{fail}} \geq 10$ consecutive error frames to erase a map element. Since sensor jitter is high-frequency and typically lasts only 1-2 frames, it falls far below this threshold, mathematically guaranteeing geometric robustness.

3) Geometric Query Attention with Sparsity: Geometric Query Attention exploits the retrieved map priors to initialize the decoder. Beyond improving accuracy, this module is critical for the system's **Efficiency**.

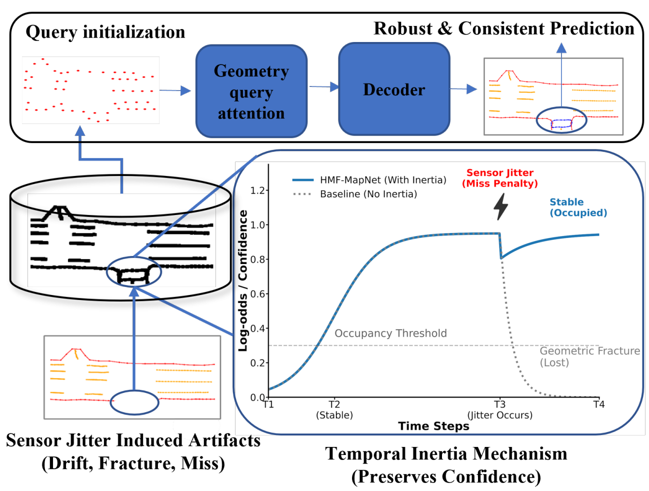

Fig. 3. Illustration of the Robustness Mechanism. (a) Problem: Sensor jitter causes geometric fractures in baseline models. (b) Mechanism: Our log-odds inertia (blue line) acts as a confidence buffer to resist transient jitter, whereas baselines (grey dashed) collapse. (c) Solution: Stable priors initialize sparse queries, ensuring robust predictions under instability.

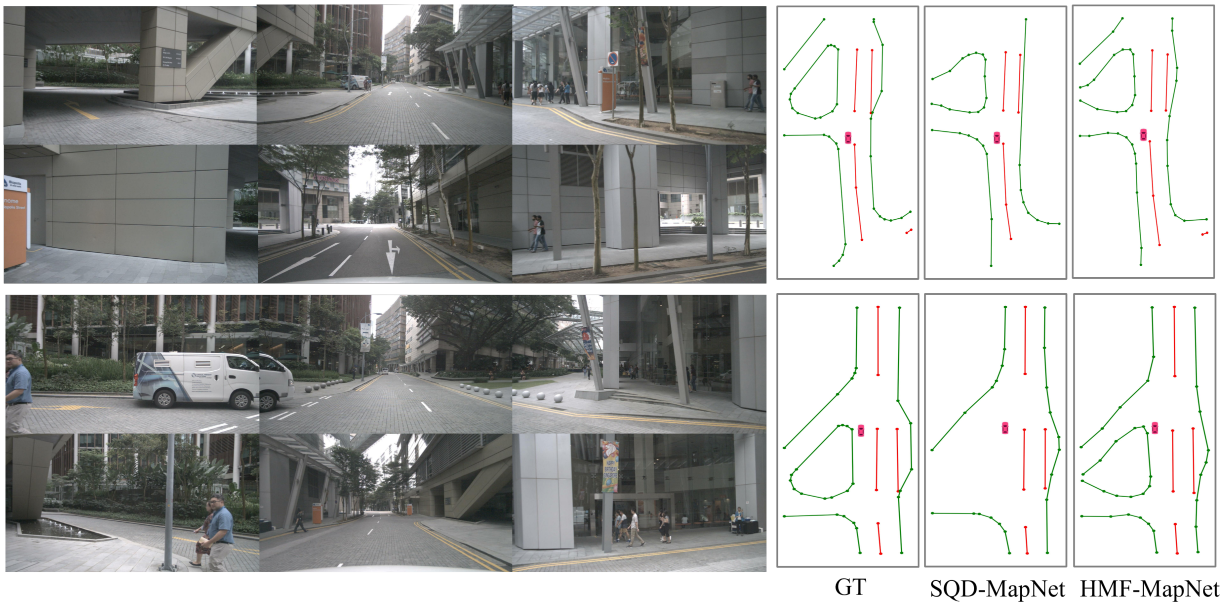

Fig. 4. The visual results of SQD-MapNet, our proposed HMF-MapNet approach and the corresponding ground truth.

Sparse Set Selection. First, we strictly define the sparse set of active geometric priors, denoted as $\mathcal{P}$, by filtering the local map prior $\mathcal{M}_{\text{local}}$ obtained in Eq. (3):

$$\mathcal{P} = \{p \mid \mathcal{M}_{\text{local}}(p) = 1\} \tag{5}$$

This filtering step enforces computational sparsity by discarding all background cells, reducing the search space from $H \times W$ to a significantly smaller subset $|\mathcal{P}| \ll H \times W$.

Hybrid Query Initialization. To satisfy the decoder's requirement for a fixed number of queries N_q [31], we construct the base queries Q_{base} using a hybrid strategy. For each retrieved location $p \in \mathcal{P}$, we generate its prior embedding:

$$\mathbf{E}_{prior}(p) = \text{PE}(p) + \mathbf{E}_{occ}. \tag{6}$$

However, the size of the sparse set $|\mathcal{P}|$ may vary. To ensure tensor consistency:

- **Padding:** If $|\mathcal{P}| < N_q$ (e.g., during initialization or in sparse areas), we supplement the insufficient slots with learnable queries initialized at random positions.
- **Truncation:** If $|\mathcal{P}| > N_q$, we select the top-N_q elements based on their log-odds confidence.

This strategy ensures that the model remains robust even in "cold start" scenarios where historical priors are scarce.

The decoder then utilizes these hybridized embeddings as the initial query set Q_{base} to attend to the sparse priors:

$$Q_{geometry} = \text{CrossAttn}(Q_{base}, \mathbf{E}_{prior}(p), \mathbf{E}_{prior}(p)) \tag{7}$$

This sparse interaction design is the theoretical foundation for the low memory consumption and high inference speed (FPS) demonstrated in our experiments, achieving an optimal trade-off between performance and efficiency.

C. Temporal Path

The temporal path introduces a Recurrent Update Module that interacts with the BEV features. Beyond simply aggregating features, the core purpose of this module is to function as a **temporal low-pass filter**, modeling dynamic changes while actively suppressing high-frequency sensor noise to maintain long-term prediction consistency.

At each timestep t, the update process begins by aligning the BEV feature map from the previous step, $\mathcal{B}_{t-1}$, to the current frame's coordinate system. This is achieved via a warping function $\mathcal{W}$ that utilizes the ego-motion transformation $\mathbf{T}_{t\leftarrow t-1}$. The resulting aligned features, denoted as $\mathcal{B}'_{t-1}$, serve as the history prior. These are then fused with the newly extracted BEV features from the current observation, $\mathcal{B}^t_{\text{obs}}$, using a Gated Recurrent Unit (GRU) [32]. The final output is normalized by a LayerNorm layer to produce the refined BEV map $\mathcal{B}_t$.

This entire recurrent update is concisely formulated as:

$$\mathcal{B}_t = \text{LayerNorm}\left(\text{GRU}\left(\mathcal{B}^t_{\text{obs}}, \mathcal{W}\left(\mathcal{B}_{t-1}, \mathbf{T}_{t\leftarrow t-1}\right)\right)\right) \tag{8}$$

This inherent robustness stems from the GRU's gating mechanism, which functions as a learnable low-pass filter. Mathematically, the update rule corresponds to:

$$\mathcal{B}_t = (1 - z_t) \odot \mathcal{B}'_{t-1} + z_t \odot \tilde{\mathcal{B}}_t \tag{9}$$

where $z_t \in [0, 1]$ is the update gate activation and $\tilde{\mathcal{B}}_t$ is the candidate activation derived from the current observation $\mathcal{B}^t_{\text{obs}}$. This formulation is mathematically equivalent to an **Exponential Moving Average (EMA)**, which acts as a temporal low-pass filter.

Under ideal conditions, the update gate z_t adapts to capture dynamic object movements. However, when extrinsic rotation

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON NUSCENES.

Method	Backbone	Epoch	Accuracy Metric				Efficiency Metric	
			AP_{ped}	AP_{div}	AP_{bou}	mAP	FPS (↑)	Mem. (MB) (↓)
MapTR [1]	R50	24	41.2	49.5	51.1	47.3	**15.7**	**465.0**
VectorMapNet [20]	R50	110	38.3	36.1	39.2	37.9	2.1	693.0
StreamMapNet [6]	R50	24	60.4	61.9	58.9	60.4	14.5	681.6
MapVR [21]	R50	24	47.7	54.4	51.0	51.0	16.1	465.1
GlobalMapNet [10]	R50	24	64.8	58.6	57.5	60.3	14.6	669.9
SQD-MapNet [11]	R50	24	63.0	65.5	63.3	63.9	12.2	685.0
MGMap [22]	R50	24	61.8	65.0	**67.5**	64.8	13.9	1033.6
Maptrv2 [30]	R50	24	59.8	62.4	62.4	61.5	15.2	674.4
HRMapNet [9]	R50	24	66.0	66.2	66.1	66.1	10.3	1044.4
InteractionMap* [23]	R50	24	67.5	**67.0**	65.4	**66.7**	11.6	1236.7
HMF-MapNet (Ours)	R50	24	**67.9**	66.8	64.9	66.5	14.6	793.8

TABLE II
COMPARISON ON THE **ARGOVERSE2** DATASET.

Method	AP_{ped}	AP_{div}	AP_{bou}	mAP
HDMapNet	5.7	13.1	37.6	18.8
MapTR	50.6	60.7	61.2	57.4
VectorMapNet	37.2	50.4	40.6	42.7
StreamMapNet	62.0	59.5	63.0	61.5
SQD-MapNet	**64.9**	60.2	64.9	63.3
HMF-MapNet (Ours)	64.1	**63.7**	**66.8**	**64.8**

noise (sensor jitter) occurs, it manifests as **high-frequency, transient fluctuations** in the feature space ($\mathcal{B}^t_{\text{obs}}$). In such cases, the gate z_t learns to assign a lower weight to the current noisy observation ($z_t \rightarrow 0$), thereby forcing the model to retain the stable, aligned historical state $\mathcal{B}'_{t-1}$. This gating mechanism effectively provides *temporal damping*, shielding the map representation from instantaneous geometric fractures caused by sensor instability while preserving the continuity of static elements.

The geometric path is designed to leverage short-term accumulated experience, providing the model with a stable and reliable Historical Map Bank. By maintaining a persistent probabilistic representation of the environment, this path decouples geometric reasoning from instantaneous sensor fluctuations, serving as a robust anchor for the dual-path system.

D. Implementation Details

To enhance temporal consistency and accelerate convergence, we incorporate the Stream Query Denoising (SQD) strategy [11]. Specifically, a subset of queries is initialized with noise-perturbed ground truth elements from the preceding frame, supervising the model to reconstruct the precise geometry of the current frame. The overall objective function L_{total} is formulated as:

$$L_{\text{total}} = L_{\text{map}} + \lambda_{\text{sqd}} L_{\text{denoise}} \tag{10}$$

where L_{denoise} shares the same formulation as the primary matching loss L_{map}, computed on the denoising query group. The map construction loss L_{map} is a linear combination of classification and geometric regression terms:

$$L_{\text{map}} = \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{pts}} L_{\text{pts}} + \lambda_{\text{dir}} L_{\text{dir}} \tag{11}$$

Here, L_{cls} denotes the Focal Loss for element classification, L_{pts} represents the L1 loss for point-wise coordinate regression, and L_{dir} is the edge direction loss used to regularize the geometric connectivity of polylines.

IV. EXPERIMENTS

A. Experimental Settings

Datasets. We evaluate HMF-MapNet on nuScenes [33] (6 cameras) and Argoverse2 [34] (7 ring cameras). For consistency, we unify the frame rate to 2Hz, downsampling Argoverse2 from its original 10Hz.

Evaluation Metrics. We focus on 3 static map categories: lane-divider, ped-crossing, and road-boundary. The thresholds for calculating AP are set to {0.5m, 1.0m, 1.5m}. Additionally, we measure inference speed (FPS) and GPU memory consumption to evaluate efficiency.

B. Comparison with State-of-the-Art

Results on nuScenes. Table I presents a comprehensive quantitative comparison on the nuScenes dataset. In terms of perception accuracy, HMF-MapNet achieves a competitive **mAP of 66.5**, establishing its position among top-tier paradigms.

- **Significant Improvement over Baselines:** Our method demonstrates a substantial performance leap compared to standard baselines, surpassing MapTR by **+19.2 mAP** and StreamMapNet by **+6.1 mAP**. This confirms that our dual-path architecture effectively resolves the geometric fractures common in baseline methods.
- **Competitive with SOTA:** HMF-MapNet outperforms the strong competitor HRMapNet (66.1 mAP) by **+0.4 mAP**. We also compare with the recent InteractionMap [23]. Since its code is unavailable, we report results based on our reproduction (InteractionMap*). While InteractionMap* attains a marginally higher accuracy (66.7 mAP),

our method remains highly competitive, with a negligible gap of only 0.2 mAP.

Results on Argoverse 2. Table II reports the results on the Argoverse 2 dataset. Our model records a **mAP of 64.8**, consistently surpassing previous temporal methods like StreamMapNet (61.5 mAP) and SQD-MapNet (63.3 mAP). This demonstrates the robust generalization capability of our dual-path architecture across different sensor layouts and environments.For a fair comparison, the backbone (R50) and training schedule (24 epochs) remain consistent with the settings used in our nuScenes experiments.

C. Efficiency Analysis

To demonstrate the real-world deployability of HMF-MapNet, we measured the inference speed (FPS) and GPU memory usage of all models on a single RTX 3090 GPU. The rightmost columns of Table I highlight the optimal trade-off achieved by our method:

- **Vs. HRMapNet:** Under a fair comparison using official reproduction, our method is **41.7% faster** (14.6 FPS vs. 10.3 FPS) and reduces memory usage by ∼250MB compared to HRMapNet, while achieving higher accuracy (+0.4 mAP).
- **Vs. InteractionMap:** Although InteractionMap achieves higher accuracy, it is computationally heavy. HMF-MapNet boosts FPS by **25.9%** and reduces memory usage by **35.8%** (∼443MB) compared to InteractionMap, making it far more suitable for on-board deployment where resources are constrained.

This efficiency stems from our architectural philosophy: unlike methods utilizing heavy dense autoregressive mechanisms to store full-label features, HMF-MapNet decouples geometry from semantics. By maintaining only lightweight binary geometric priors (Log-odds), we eliminate the need to maintain heavy dense feature maps, directly driving the significant memory reduction and speed boost.

D. Robustness Analysis

Real-world autonomous driving systems inevitably face sensor instability caused by uneven terrain, speed bumps, or mechanical vibrations. To rigorously evaluate the reliability of HMF-MapNet under such harsh conditions, we conducted a stress test by simulating sensor noise. Specifically, we introduced random rotation noise to the camera extrinsics, with the standard deviation σ_r ranging from 0 to 0.1 radians. This setup mimics the pitch and roll jitter often encountered in daily driving, which typically disrupts the spatial alignment of geometric features.

Performance under Extreme Noise. As quantified in Table III, our framework demonstrates exceptional stability compared to state-of-the-art methods.

- **Baseline Collapse:** The standard end-to-end method, MapTRv2, exhibits a *catastrophic failure* as noise increases. Its accuracy plummets from 61.5 mAP to a negligible 14.2 mAP at $\sigma_r = 0.1$. This confirms that single-frame, prior-free architectures are extremely brittle when spatial alignment is compromised.
- **SOTA Comparison:** While HRMapNet maintains reasonable performance at low noise levels, it begins to degrade significantly under strong interference ($\sigma_r \geq 0.05$). At the most extreme noise level ($\sigma_r = 0.1$), HRMapNet drops to 52.2 mAP.
- **Our Superiority:** In sharp contrast, HMF-MapNet maintains high integrity. At $\sigma_r = 0.1$, it achieves a remarkable **58.5 mAP**, outperforming HRMapNet by a substantial margin of **+6.3 mAP** and surpassing MapTRv2 by over **+44 mAP**.

Mechanism Analysis. This decisive advantage is attributed to the **Temporal Inertia** provided by our Log-odds Map Bank. As illustrated in Fig. 3, the probabilistic grid acts as a low-pass temporal filter. High-frequency sensor jitter (transient noise) may cause instantaneous misalignments, but these are insufficient to overturn the accumulated high-confidence occupancy states in the history bank. Consequently, HMF-MapNet effectively "filters out" the geometric outliers caused by sensor instability, ensuring consistent planning-ready maps.

TABLE III
ROBUSTNESS ANALYSIS UNDER EXTRINSIC ROTATION NOISE (MAP).

Model	σ_r=0	0.005	0.01	0.02	0.05	0.1
HRMapNet	66.1	66.0	**65.7**	64.9	60.1	52.2
Maptrv2	61.5	61.2	59.8	35.6	18.8	14.2
HMF-MapNet	**66.5**	**66.3**	65.5	**65.1**	**62.0**	**58.5**

E. Ablation Study

To validate the effectiveness of HMF-MapNet's dual-path architecture, we constructed two variants: *w/o G* (removing the Geometric Path/Historical Map Bank) and *w/o T* (removing the Temporal Path/Recurrent Update Module).

As shown in Table IV, both variants exhibited significant performance degradation compared to the full model (66.5 mAP). Specifically, removing the Geometric Path caused a drop of 4.4 mAP, confirming that the historical map bank provides crucial long-term static priors. Removing the Temporal Path resulted in a 2.8 mAP drop, highlighting the importance of the recurrent module in smoothing dynamic features.

TABLE IV
ABLATION STUDY ON NUSCENES. 'W/O G': WITHOUT GEOMETRIC PATH; 'W/O T': WITHOUT TEMPORAL PATH.

Method	AP_{ped}	AP_{div}	AP_{bou}	mAP
HMF-MapNet w/o G	61.2	61.9	63.3	62.1
HMF-MapNet w/o T	63.1	65.2	63.8	63.7
HMF-MapNet	**67.9**	**66.8**	**64.9**	**66.5**

V. CONCLUSION

We proposed HMF-MapNet, a dual-path framework that resolves key limitations in vectorized HD map perception.

By synergistically modeling geometry with a historical map bank and time with a recurrent network, it mitigates prior sensitivity and error accumulation, achieving high performance on nuScenes and Argoverse2. Future work will explore multi-modal sensor fusion to further enhance autonomous driving safety.

References

[1] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang, "Maptr: Structured modeling and learning for online vectorized hd map construction," *arXiv preprint arXiv:2208.14437*, 2022.

[2] Fang Da and Yu Zhang, "Path-aware graph attention for hd maps in motion prediction," in *2022 International conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 6430–6436.

[3] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al., "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17853–17862.

[4] Wenjie Ding, Limeng Qiao, Xi Qiu, and Chi Zhang, "Pivotnet: Vectorized pivot learning for end-to-end hd map construction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3672–3682.

[5] Wenda Li, Tongya Zheng, Kaixuan Chen, Shunyu Liu, Haoze Jiang, Yunzhi Hao, Rui Miao, Zujie Ren, Mingli Song, Hang Shi, et al., "Physics-informed diffusion generation for geomagnetic map interpolation," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[6] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao, "Streammapnet: Streaming mapping network for vectorized online hd map construction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7356–7365.

[7] Jiaqi Li, Pingfan Jia, Jiaxing Chen, Jiaxi Liu, Lei He, and Keqiang Li, "Local map construction with sdmap: A comprehensive survey," *arXiv preprint arXiv:2409.02415*, 2024.

[8] Zhou Jiang, Zhenxin Zhu, Pengfei Li, Huan-ang Gao, Tianyuan Yuan, Yongliang Shi, Hang Zhao, and Hao Zhao, "P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors," *IEEE Robotics and Automation Letters*, 2024.

[9] Xiaoyu Zhang, Guangwei Liu, Zihao Liu, Ningyi Xu, Yunhui Liu, and Ji Zhao, "Enhancing vectorized map perception with historical rasterized maps," in *European Conference on Computer Vision*. Springer, 2024, pp. 422–439.

[10] Anqi Shi, Yuze Cai, Xiangyu Chen, Jian Pu, Zeyu Fu, and Hong Lu, "Globalmapnet: An online framework for vectorized global hd map construction," *arXiv preprint arXiv:2409.10063*, 2024.

[11] Shuo Wang, Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Zehui Chen, Tiancai Wang, Chi Zhang, Xiangyu Zhang, and Feng Zhao, "Stream query denoising for vectorized hd-map construction," in *European Conference on Computer Vision*. Springer, 2024, pp. 203–220.

[12] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[13] Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su, "Sparse4d v3: Advancing end-to-end 3d detection and tracking," *arXiv preprint arXiv:2311.11722*, 2023.

[14] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] Ruihang Li, Shanding Ye, Zhe Yin, Tao Li, ZeHua Zhang, KaiKai Xiao, and Zhijie Pan, "M2depth: A novel self-supervised multi-camera depth estimation with multi-level supervision," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[16] Ruihang Li, Tao Li, Zhe Yin, Huangnan Zheng, KaiKai Xiao, and Zhijie Pan, "Cfd: A novel cylindrical feature fusion and mapping pipeline for multi-view self-supervised depth estimation," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[17] Ruihang Li, Tao Li, Shanding Ye, Kaikai Xiao, Zheng Huangnan, Zhe Yin, and Zhijie Pan, "Enhancing generalizability via utilization of unlabeled data for occupancy perception," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 4896–4904.

[18] Ruihang Li, Huangnan Zheng, Zhe Yin, Kaikai Xiao, Hong Li, and Zhijie Pan, "Spoformer: Enhancing prior learning for multi-view 3d occupancy perception via semantic-aware attention," in *European Conference on Artificial Intelligence*, 2025.

[19] Xiyue Guo, Jiarui Hu, Junjie Hu, Hujun Bao, and Guofeng Zhang, "Sgformer: Satellite-ground fusion for 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 11929–11938.

[20] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao, "Vectormapnet: End-to-end vectorized hd map learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22352–22369.

[21] Gongjie Zhang, Jiahao Lin, Shuang Wu, Zhipeng Luo, Yang Xue, Shijian Lu, Zuoguan Wang, et al., "Online map vectorization for autonomous driving: A rasterization perspective," *Advances in Neural Information Processing Systems*, vol. 36, pp. 31865–31877, 2023.

[22] Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu, "Mgmap: Mask-guided learning for online vectorized hd map construction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14812–14821.

[23] Kuang Wu, Chuan Yang, and Zhanbin Li, "Interactionmap: Improving online vectorized hdmap construction with interaction," *arXiv preprint arXiv:2503.21659*, 2025.

[24] Jialong Xie, Fengyu Zhou, Jin Liu, and Chaoqun Wang, "Semi-supervised language-conditioned grasping with curriculum-scheduled augmentation and geometric consistency," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 4021–4028, 2025.

[25] Jialong Xie, Jin Liu, Saike Huang, Chaoqun Wang, and Fengyu Zhou, "Listen, perceive, grasp: Clip-driven attribute-aware network for language-conditioned visual segmentation and grasping," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 9729–9740, 2025.

[26] Kaixuan Chen, Jie Song, Shunyu Liu, Na Yu, Zunlei Feng, Gengshi Han, and Mingli Song, "Distribution knowledge embedding for graph pooling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 7898–7908, 2023.

[27] Kaixuan Chen, Shunyu Liu, Tongtian Zhu, Ji Qiao, Yun Su, Yingjie Tian, Tongya Zheng, Haofei Zhang, Zunlei Feng, Jingwen Ye, et al., "Improving expressivity of gnns with subgraph-specific factor embedded normalization," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 237–249.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[30] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang, "Maptrv2: An end-to-end framework for online vectorized hd map construction," *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1352–1374, 2025.

[31] Zhixin Zhang, Yiyuan Zhang, Xiaohan Ding, Fusheng Jin, and Xiangyu Yue, "Online vectorized hd map construction using geometry," in *European Conference on Computer Vision*. Springer, 2024, pp. 73–90.

[32] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[33] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631.

[34] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al., "Argoverse 2: Next generation datasets for self-driving perception and forecasting," *arXiv preprint arXiv:2301.00493*, 2023.