2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

RelMap: Enhancing Online Map Construction with Class-Aware Spatial Relation and Semantic Priors

Tianhui Cai, Yun Zhang, Zewei Zhou, Zhiyu Huang*, Jiaqi Ma
University of California, Los Angeles

***Abstract*— Online high-definition (HD) map construction is crucial for scaling autonomous driving systems. While Transformer-based methods have become prevalent in online HD map construction, most existing approaches overlook the inherent spatial dependencies and semantic relationships between map elements, which constrains their accuracy and generalization capabilities. To address this, we propose *RelMap*, an end-to-end framework that explicitly models both spatial relations and semantic priors to enhance online HD map construction. Specifically, we introduce a *Class-aware Spatial Relation Prior*, which explicitly encodes relative positional dependencies between map elements using a learnable class-aware relation encoder. Additionally, we design a Mixture-of-Experts-based *Semantic Prior*, which routes features to class-specific experts based on predicted class probabilities, refining instance feature decoding. *RelMap* is compatible with both single-frame and temporal perception backbones, achieving state-of-the-art performance on the nuScenes and Argoverse 2 datasets.**

I. INTRODUCTION

High-definition (HD) maps encode rich geometric and semantic information about road infrastructure and play a pivotal role in autonomous driving systems, enabling localization [1], prediction [2], [3], and planning [4], [5]. However, the maintenance and frequent updating of HD maps impose substantial costs. Furthermore, the reliance on pre-constructed HD maps significantly limits the scalability and adaptability of autonomous vehicles, particularly in dynamically changing environments. To mitigate these limitations, recent works have explored online HD map construction using onboard sensory data [6], [7], [8], [9], [10], [11], [12]. These methods generate HD maps using perception inputs, thus obviating the need for pre-annotated maps and enhancing scalability in real-world scenarios.

Most existing methods adopt a bird's-eye-view (BEV) perception backbone [13], [14] that lifts 2D image features into BEV space, followed by DETR-style [15] decoders to predict vectorized map instances [7], [16], [17], [8], [18]. While effective, these methods typically treat all map elements independently and uniformly, neglecting the spatial dependencies and semantic regularities inherent in map topology. Such relationships are crucial for accurate map construction: for example, lane dividers often run parallel to road boundaries, and crosswalks are typically positioned near intersections.

To address this limitation, we propose **RelMap** that explicitly incorporates spatial and semantic priors into

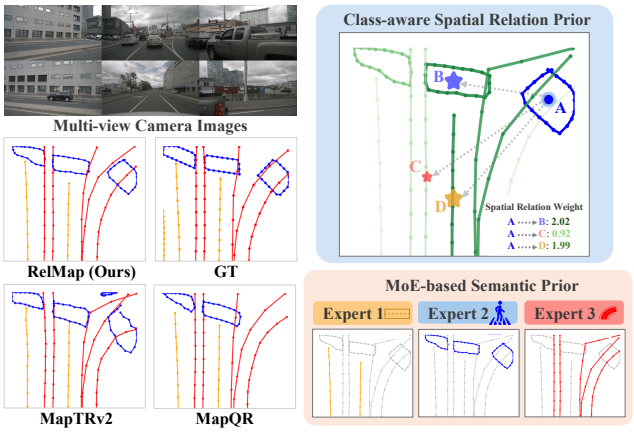

Fig. 1. Two major improvements proposed in **RelMap** for online vectorized map construction. The **Class-aware Spatial Relation Prior** explicitly encodes spatial dependencies between map elements and decodes important weights, while the **MoE-based Semantic Prior** routes features to class-specific experts based on predicted class probabilities for refined decoding.

Transformer-based decoders for online HD map construction. Our design is motivated by two key insights. First, **structured spatial dependencies**: map elements exhibit consistent geometric patterns (e.g., lane dividers parallel to road boundaries). To leverage such spatial regularities, we introduce a learnable *Class-aware Spatial Relation Prior* that models inter-instance geometric dependencies. Specifically, we develop a *Position Relation Encoder* inspired by relation encoding in vision Transformers [19], [20], [21], which computes relative positions between instances and propagates this information across Transformer layers. Additionally, to incorporate class-specific cues into spatial modeling, we propose a *Class Relation Modeling Prior* implemented as a learnable relation matrix that encodes semantic affinities among map classes. This matrix modulates the spatial relation embeddings with class-aware signals, allowing the model to focus on semantically relevant spatial contexts. Second, **class-specific semantic decoding**: map elements from different classes exhibit distinctive appearance and structural patterns (e.g., crosswalks are rectangular, while lane dividers are elongated and linear). To capture these variations, we introduce a *Class-conditioned Mixture-of-Experts (MoE)-based Semantic Prior*. Rather than decoding all map instances within a shared feature space, our approach dynamically routes each instance to a specialized expert based on its predicted class probability. Each expert learns to model the distinct characteristics of a specific class, thereby promoting more accurate and semantically consistent decoding. This

*Corresponding author. zhiyuh@ucla.edu

MoE mechanism, inspired by prior works in dynamic routing [22], enables scalable learning of class-specific decoding functions while preserving model efficiency.

To validate the effectiveness of the proposed priors, we build two variants of RelMap: a single-frame model (**RelMap-SF**) and a temporal model with historical memory integration (**RelMap-TF**). These two variants demonstrate the applicability of our priors across different temporal settings and architectures, consistently improving vectorized map prediction quality. The key improvements of our online vectorized HD map construction model are illustrated in Fig. 1. The main contributions are summarized as follows:

1) We propose **RelMap**, an online HD map construction framework that introduces two learnable priors: a *Class-aware Spatial Relation Prior* for modeling of geometric dependencies and a *Class-conditioned MoE Semantic Prior* for class-specific decoding.
2) We demonstrate that these priors are compatible with both single-frame and temporal perception backbones, highlighting their broad applicability.
3) Our approach achieves state-of-the-art performance on the nuScenes and Argoverse 2 benchmarks in both single-frame and temporal settings, demonstrating the effectiveness and generalization capability.

II. RELATED WORK

Vectorized Map Construction. HD map construction aims to generate structured maps from sensor data. MapTR [7] and MapTRv2 [23] enable end-to-end vectorized prediction via hierarchical queries, inspiring a series of follow-up works [24], [18], [17], [25], [9]. Recent works such as StreamMapNet [9] and MapTracker [10] further enhance performance by leveraging historical context through memory aggregation, temporal query modeling, or feature propagation [9], [26], [10], [27], [28], [29], [30]. Several prior-based methods have been proposed to improve HD map construction. Some leverage external data like SD maps [31], [32], [30], while others learn priors to guide training [6], [25]. For example, MGMap [6] introduces instance- and position-guided masks to refine feature localization, and PriorMapNet [25] integrates structure priors from clustered map elements into query initialization. However, these works neglect spatial relationship priors across instances. We address this by introducing a learnable, class-aware spatial relation prior that promotes more structured and context-aware predictions.

Relation Modeling. Relation modeling has demonstrated effectiveness across various computer vision tasks, including image recognition [33], object detection [34], and generative modeling [19], [35]. Category-level modeling captures class co-occurrence patterns [33], [36], while instance-level methods focus on spatial or semantic relations between objects [21], [20]. Despite the effectiveness of relation modeling in vision tasks, its application in HD map construction remains underexplored [37], [38], [39], [40], particularly explicit spatial relation modeling. GeMap [24] separates attention into shape attention for intra-instance refinement and relation attention for inter-instance interactions, but lacks explicit relation modeling conditioned on classes. We introduce an instance-level spatial relation modeling approach that incorporates class-aware adjustments, enabling the model to capture spatial dependencies more effectively across different map categories.

Mixture-of-Experts. MoE [22] enhances model capacity expressiveness by dynamically routing inputs to specialized sub-networks ("experts"). Instead of processing all inputs through a shared network, MoE assigns each input to selected experts, typically FFNs, via a gating function. Sparse gating selects only a subset of experts per input, often using a top-K strategy [41], [42], whereas dense gating activates all experts with different weights [43], [44]. MapExpert [45] employs a sparse MoE [22] with a top-K gating mechanism, where each input is assigned to a subset of experts with the highest gating scores. Like traditional MoE approaches, it relies on a router network for expert selection and an auxiliary expert balance loss to prevent imbalanced expert utilization. In contrast, our method eliminates the need for a separate router and expert balance loss by directly leveraging the classification predictions from the previous decoder layer. This simplifies the training process, reduces additional parameters, and achieves better performance with fewer experts.

III. METHOD

We formulate online vectorized HD map construction as the task of generating structured map instances from surrounding multi-view perspective-view (PV) images. Each map instance is represented as a vectorized polyline defined by an ordered set of 2D points, denoted as $(x_i, y_i)_{i=1}^{N_p}$, where N_p is the number of points composing the instance.

A. Main Framework

To enhance spatial and semantic reasoning in vectorized map construction, as illustrated in Fig. 2, we introduce two learnable priors: 1) a Learnable Class-aware Spatial Relation Prior, which enhances spatial dependency modeling between map instances, and 2) a MoE-based Semantic Prior, which enables class-specific feature refinement.

We instantiate these priors in two model variants: **RelMap-SF**, a single-frame map construction model built on MapQR [16], and **RelMap-TF**, a temporal map construction model built on MapTracker [10]. Although these models differ in specific architectural components, they share a common foundation inherited from the MapTR series [7], [23], including: 1) a ResNet-50 [46] vision backbone for multi-view image feature extraction, 2) a BEV encoder to transform image features into a BEV representation, and 3) a deformable Transformer decoder that outputs vectorized map instances. MapQR enhances efficiency via a GKT-h BEV encoder and scatter-and-gather instance queries for one-to-one decoding in the single-frame regime. MapTracker extends the MapTRv2 decoder with a Strided Memory Fusion module, aggregating temporal features for stable predictions. Despite these differences, the core query-based deformable Transformer decoding structure remains consistent, allowing seamless integration of our proposed priors.

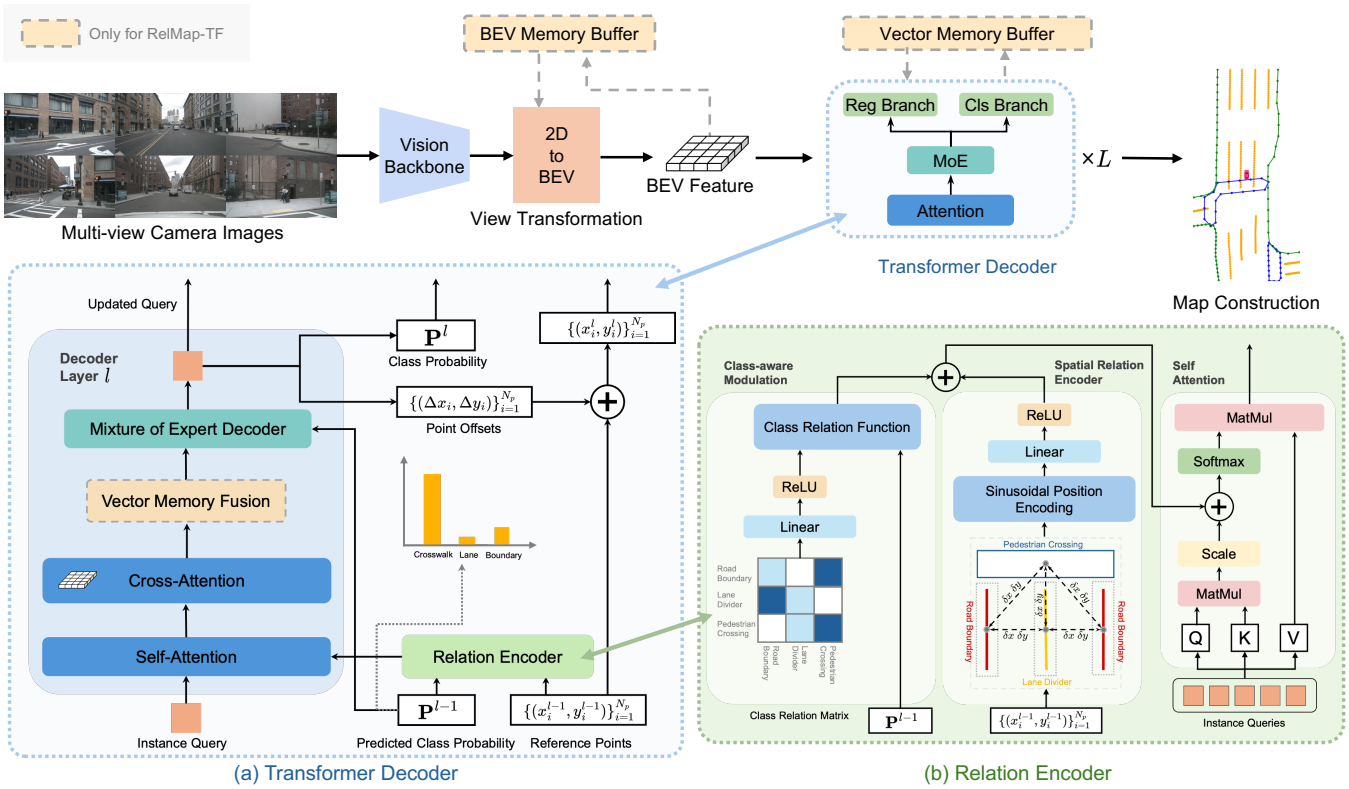

Fig. 2. Overview of **RelMap**, our proposed online vectorized map construction framework. Given multi-view camera images, the model first extracts image features using a vision backbone and transforms them into the BEV space. Subsequently, an enhanced Transformer decoder predicts vectorized map elements. In the temporal variant, RelMap-TF, memory buffers and a fusion module integrate historical context. To improve spatial and semantic reasoning, we refine the decoder's self-attention mechanism using a class-aware relation encoder and introduce a MoE decoder, where class probabilities guide the routing of map instances to specialized expert networks.

B. Class-aware Spatial Relation Prior

To better capture structured spatial dependencies between map instances, we introduce a Learnable Class-aware Spatial Relation Prior, which modulates self-attention by incorporating relational cues from geometric and semantic structures. As illustrated in Fig. 2(b), this prior is computed via a Relation Encoder comprising two components: 1) a Spatial Relation Encoder, which encodes relative positional relationships between instances, and 2) a Class-aware Modulation, which adjusts spatial dependencies based on instance class relationships. As shown in Fig. 2(a), this prior is integrated into the self-attention mechanism of the Transformer decoder to enhance relation modeling.

Spatial Relation Encoder. Given a set of predicted map instances $\{(x_i^{l-1}, y_i^{l-1})\}_{i=1}^{N_p}$ from the Transformer decoder layer $l-1$, we compute their relative spatial relationships using bounding box-based positions. To obtain the bounding box for each instance, we use the smallest axis-aligned bounding box that encloses all the N_p points in the instance. Each instance is then represented by its bounding box $\mathbf{b}_i = (cx_i, cy_i, w_i, h_i)$, where (cx_i, cy_i) denotes the center and (w_i, h_i) represents the width and height. This provides a compact representation of the instance's spatial extent without requiring additional model layers. To encode spatial dependencies, we construct a position relation embedding following [21] by computing the relative displacement and scale differences between instances. For each pair of map instances i and j, we compute their relations:

$$\delta x = \log\left(\frac{|cx_i - cx_j|}{w_i + \epsilon} + 1\right), \tag{1}$$

$$\delta y = \log\left(\frac{|cy_i - cy_j|}{h_i + \epsilon} + 1\right), \tag{2}$$

$$\delta w = \log\left(\frac{w_i}{w_j + \epsilon}\right), \ \delta h = \log\left(\frac{h_i}{h_j + \epsilon}\right), \tag{3}$$

where ϵ is a small constant to prevent numerical instability.

These relative features are then transformed using sinusoidal position encoding and processed through an MLP to obtain a spatial relation embedding:

$$R_{i,j}^{spatial} = f\left(\mathrm{PE}\left(\delta x, \delta y, \delta w, \delta h\right)\right), \tag{4}$$

where $f = \mathrm{ReLU} \circ \mathrm{Linear}$, and $R^{spatial} \in \mathbb{R}^{N_{\text{ins}} \times N_{\text{ins}} \times N_{\text{head}}}$ represents the spatial relationships between all N_{ins} map instances, with N_{head} as the number of attention heads.

Class-aware Modulation. To further refine spatial relationships based on semantic structure, we introduce a *learnable class relation matrix* $\mathbf{R}_{cls} \in \mathbb{R}^{N_c \times N_c \times D}$, which models interactions between different classes. Here, N_c denotes the number of classes, and D is the embedding dimension. This matrix encodes the semantic dependencies between different

classes, allowing the model to modulate spatial relationships based on class-specific contextual cues.

Previous works such as [36] rely on one-hot assignments to query the class relation matrices, making them highly sensitive to classification errors. Since our model updates relational reasoning based on the classification results from previous layers, inaccuracies in classification predictions in early layers can lead to incorrect relational dependencies, which may propagate through subsequent layers and affect final predictions. To mitigate this, we compute a **soft class relation prior** using a weighted sum of the learnable class relation matrix, where the predicted class probabilities determine the weights. Specifically, we obtain the predicted class probabilities for each map instance from the previous decoder layer $\mathbf{P}^{l-1} \in \mathbb{R}^{N_{\text{ins}} \times N_c}$, where P^{l-1}_{i,c_k} denotes the probability of instance i belonging to class c_k, and C is the total number of classes. Given class probabilities P^{l-1}_i and P^{l-1}_j for map instances i and j, the class-aware relation embedding is computed using the following function:

$$R^{cls}_{i,j} = \sum_{c_k=1}^{N_C} \sum_{c_h=1}^{N_C} P^{l-1}_{i,c_k} g(\mathbf{R}_{cls}(c_i, c_j)) P^{l-1}_{j,c_h}, \tag{5}$$

where $g = \text{ReLU} \circ \text{Linear}$, and $R^{\text{cls}} \in \mathbb{R}^{N_{\text{ins}} \times N_{\text{ins}} \times N_{\text{head}}}$ is class relation embedding, which encodes the pairwise class-aware relational dependencies between instances, capturing how different map elements interact based on their predicted class probabilities.

Integration into Self-attention. The Learnable Class-aware Spatial Relation Prior is obtained by combining the learned spatial relation embedding R^{spatial} and the class-aware modulation R^{cls}:

$$R^{\text{rel}} = R^{\text{spatial}} + R^{\text{cls}}. \tag{6}$$

We incorporate the relation prior into self-attention, where relationships between queries are learned. As shown in Fig. 2 (b), the relation prior is used to modulate the attention weights by adding R^{rel} to the attention logits before applying the softmax function. The modified attention is given as:

$$\text{Attn}_{\text{self}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + R^{\text{rel}}\right)\mathbf{V}. \tag{7}$$

By integrating R^{rel}, the self-attention computation is enhanced beyond feature similarities, incorporating structured spatial and semantic dependencies.

C. MoE-based Semantic Prior

Different types of map elements, such as pedestrian crossings, lane dividers, and road boundaries, exhibit distinct semantic characteristics that influence their contextual roles in HD maps. However, standard Transformer-based models process all instances within a shared feature space, limiting their capacity to capture class-specific nuances. To address this limitation, we introduce a MoE-based Semantic Prior, allowing the model to adaptively refine instance features based on their predicted class probabilities.

As illustrated in Fig. 3, the MoE module is integrated into the feed-forward network (FFN) of each Transformer decoder layer. Instead of applying a single shared FFN to all instances uniformly, the MoE module dynamically routes instance features to specialized expert networks. The assignment is performed through a weighted combination, where the weights are derived from the predicted class probabilities of the previous decoder layer, P^{l-1}:

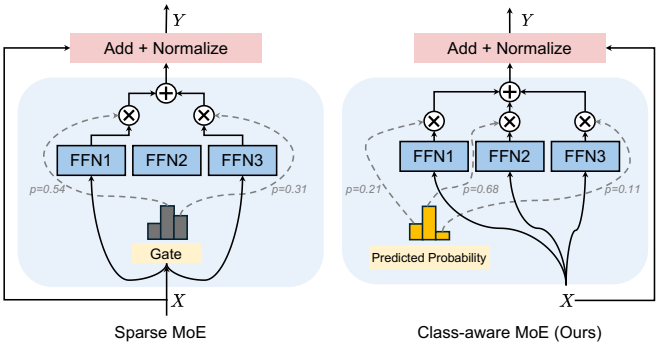

Fig. 3. Comparison between **Sparse MoE** and our **Class-aware MoE**. In Sparse MoE, a gating network selects the top-k experts and computes a weighted sum of their outputs. In contrast, our method uses predicted class probabilities to assign experts directly, eliminating the need for an additional gating network and allowing each expert to learn class-specific patterns as a semantic prior.

$$\mathbf{y}_i = \sum_{c=1}^{N_C} P^{l-1}_{i,c} E_c(\mathbf{x}_i), \tag{8}$$

where $\mathbf{x}_i$ is the feature of instance i, $P^{l-1}_{i,c}$ represents the probability of instance i belonging to class c, and $E_c(\cdot)$ is the expert network corresponding to class c.

Unlike traditional MoE, which relies on a separate routing network for expert selection, our approach directly utilizes the classification branch output from the previous decoder layer. Traditional MoE architectures require an additional trainable routing network, introducing extra parameters and increasing model complexity. Furthermore, these methods often require an auxiliary loss to balance expert utilization and prevent mode collapse. In contrast, our design eliminates the need for an explicit routing mechanism, reducing computational overhead while ensuring a more stable and interpretable expert selection process. By leveraging the instance classification predictions, our method naturally aligns expert assignment with semantic class distributions, leading to more efficient and semantically coherent feature refinement.

D. Model Training

RelMap-SF. Following MapQR [16], we supervise training with three loss terms: a one-to-one instance prediction loss $\mathcal{L}_{\text{one2one}}$, an auxiliary one-to-many loss $\mathcal{L}_{\text{one2many}}$, and a dense foreground segmentation loss $\mathcal{L}_{\text{dense}}$. The overall loss function is defined as:

$$\mathcal{L}_{\text{SF}} = \lambda_o \mathcal{L}_{\text{one2one}} + \lambda_m \mathcal{L}_{\text{one2many}} + \lambda_d \mathcal{L}_{\text{dense}}, \tag{9}$$

where λ_o, λ_m, and λ_d are loss weights.

Specifically, $\mathcal{L}_{\text{one2one}}$ employs the Hungarian algorithm with Manhattan distance for bipartite matching, incorporating classification, point-wise, and edge direction losses. $\mathcal{L}_{\text{one2many}}$

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS ON NUSCENES VALIDATION SET AT 60 M × 30 M PERCEPTION RANGE

Method	Epoch	AP_{ped}	AP_{div}	AP_{bou}	mAP	FPS
MapTRv2 [23]	24 | 110	59.8 | 68.1	62.4 |68.3	62.4 | 69.7	61.5 | 68.7	23.6
MGMap [6]	24 | 110	61.8 | 64.4	65.0 | 67.6	67.5 | 67.7	64.8 | 66.5	-
MapQR [16]	24 | 110	63.4 | 70.1	68 | 74.4	67.7 | 73.2	66.4 | 72.6	22.2
HIMap [8]	30 | 110	62.6 | 71.3	68.4 | 75.0	69.1 | 74.7	66.7 | 73.7	-
PriorMapNet [25]	24 | 110	64.0 | 71.5	69.0 | 73.2	68.2 | 73.3	67.1 | 72.7	-
MGMapNet [18]	24 | 110	64.7 | **74.3**	66.1 | 71.8	**69.4** | **74.8**	66.8 | 73.6	-
RelMap-SF (Ours)	24 | 110	**66.3** | 72.2	**70.5** | **75.1**	68.4 | 74.7	**68.4** | **74.0**	20.8
StreamMapNet [9]	110	70.0	72.9	68.3	70.4	21.7
MapUnveiler [29]	110	71.0	69.1	71.8	70.6	-
PrevPredMap [28]	110	71.2	70.0	72.8	71.3	-
MapTracker [10]	72	80.0	74.1	74.1	76.1	17.6
MapExpert [45]	100	79.4	73.9	76.2	76.5	-
HisTrackMap [27]	72	79.8	**74.5**	75.4	76.6	-
RelMap-TF (Ours)	72	**81.1**	73.6	**76.6**	**77.1**	16.4

introduces an additional prediction branch to generate multiple hypotheses per instance, which improves convergence and robustness. $\mathcal{L}_{\text{dense}}$ includes BEV and PV segmentation losses that guide the model to predict foreground regions across multiple views.

RelMap-TF. For temporal modeling, we follow MapTracker [10] and adopt three objectives: BEV segmentation loss $\mathcal{L}_{\text{BEV}}$, vector tracking loss $\mathcal{L}_{\text{track}}$, and transformation consistency loss $\mathcal{L}_{\text{trans}}$. The total loss is:

$$\mathcal{L}_{\text{TF}} = \mathcal{L}_{\text{BEV}} + \mathcal{L}_{\text{track}} + \lambda_s \mathcal{L}_{\text{trans}}, \tag{10}$$

where λ_s is a balance weight.

Here, $\mathcal{L}_{\text{BEV}}$ combines pixel-wise Focal and Dice losses over rasterized BEV masks. $\mathcal{L}_{\text{track}}$ enforces temporal consistency via hierarchical matching of classification and geometric alignment terms between consecutive frames. Finally, $\mathcal{L}_{\text{trans}}$ ensures that query updates in latent space preserve instance geometry and class semantics over time.

IV. EXPERIMENTS

A. Datasets and Evaluation

Datasets. We evaluate our approach on two widely used datasets for HD vectorized map construction: the nuScenes dataset [47] and the Argoverse 2 dataset [48]. The nuScenes dataset provides a 360-degree field-of-view (FOV) coverage of the ego-vehicle using six surrounding cameras, capturing diverse urban driving scenarios with 2D vectorized annotations for critical map elements. Argoverse 2 comprises data from seven cameras and provides high-fidelity 3D vectorized map annotations.

Evaluation Metrics. Following previous works, we evaluate our method on three types of map instances: lane dividers, pedestrian crossings, and road boundaries. To establish correspondences between predicted and ground truth map elements, we use Chamfer Distance (CD) as the matching criterion under three thresholds: 0.5, 1.0, and 1.5 meters. We compute Average Precision (AP) at each threshold, and report the final mean Average Precision (mAP) as the average over all classes and distance thresholds.

B. Implementation Details

All models in this paper are trained on four NVIDIA L40S GPUs with a ResNet-50 [46] backbone.

RelMap-SF. We follow the implementation of MapQR [16], using $N_{\text{ins}} = 100$ instance queries, each predicting $N_p = 20$ points. The Transformer decoder is configured with $N_{\text{head}} = 8$ attention heads and an embedding dimension of $D = 256$. For training, we use the AdamW optimizer with a cosine annealing learning rate schedule starting at 6×10^{-4}. The model is trained for 24 and 110 epochs on nuScenes and 6 epochs on Argoverse 2.

RelMap-TF. We adopt the training setup of MapTracker [10]. During training, four historical frames are randomly sampled from the ten preceding frames (5 seconds). The model is trained for 72 epochs on nuScenes using a three-stage training pipeline and the AdamW optimizer with an initial learning rate of 5×10^{-4}.

C. Main Results

Quantitative Results on nuScenes. We evaluate both single-frame and temporal variants of our RelMap framework on the nuScenes [47] validation set. As shown in Table I, RelMap-SF consistently achieves the highest mAP among single-frame methods. At 24 epochs, it reaches an mAP of 68.4, outperforming MapTRv2 by 6.9 points. At 110 epochs, it achieves an mAP of 74.0, exceeding MapQR [16] by 1.4 points and outperforming MapTRv2 by 5.3 points. These substantial gains underscore the efficacy of incorporating spatial and semantic priors. Compared to PriorMapNet [25], which relies on offline priors that require dataset-specific recomputation, our learnable priors within the Transformer model adapt dynamically during training, eliminating the need for precomputed priors while still achieving a 1.3-point higher mAP than PriorMapNet. For temporal map construction, RelMap-TF achieves a new state-of-the-art performance with an mAP of 77.1, outperforming leading methods including MapTracker [10], HisTrackMap [27], and MapExpert [45]. To better assess generalization capabilities, we further evaluate RelMap-TF on the geographically disjoint

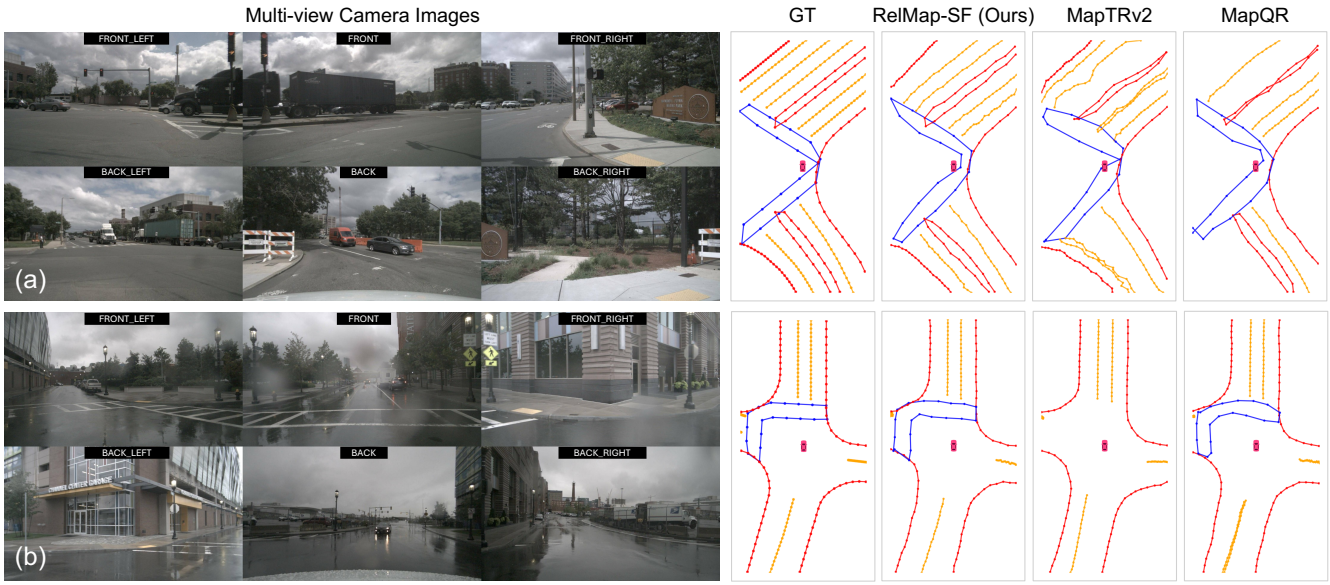

Fig. 4. Qualitative results of RelMap-SF on the nuScenes dataset. The red elements represent road boundaries, yellow elements represent lane dividers, and blue elements represent pedestrian crossings. Our model consistently outperforms other models in terms of spatial relations and shape prediction.

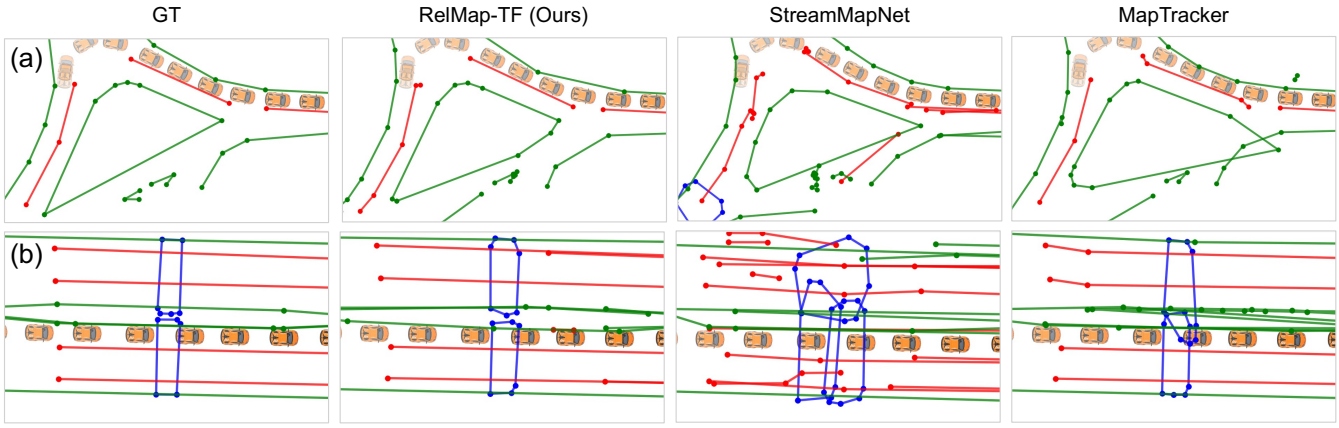

Fig. 5. Qualitative results of RelMap-TF on the nuScenes dataset. The green elements represent road boundaries, red elements represent lane dividers, and blue elements represent pedestrian crossings. Our model produces more accurate and spatially consistent map elements.

train/validation split proposed by StreamMapNet [9], which mitigates the risk of inflated scores due to memorization in the original split. As shown in Table II, RelMap-TF achieves an mAP of 42.3, demonstrating the model's robustness and improved generalization to unseen environments.

These results underscore the effectiveness of our proposed spatial relations and semantic priors in enhancing map construction. The consistent improvements across both single-frame and temporal variants demonstrate that our priors are generalizable and can be integrated into Transformer-based decoders. Furthermore, in terms of computational efficiency, RelMap achieves comparable frames per second (FPS) compared to existing methods, ensuring a balance between accuracy and inference speed.

Quantitative Results on Argoverse 2. We further evaluate the RelMap-SF model on the Argoverse 2 dataset, which provides 3D $(x, y, z$ coordinates) vectorized map data, allowing assessment in both 2D and 3D. As shown in Table III, RelMap-SF achieves an mAP of 69.9 in 2D and 68.5 in 3D, surpassing MapTRv2 by 2.5 and 3.8 points, respectively. These results demonstrate the strong performance of our approach across different spatial representations and datasets.

TABLE II

RESULTS ON THE GEOGRAPHICALLY DISJOINT TRAIN AND VALIDATION SPLIT OF NUSCENES PROPOSED BY STREAMMAPNET [9].

Methods	Epoch	AP_{ped}	AP_{div}	AP_{bou}	mAP
StreamMapNet [9]	110	31.6	28.1	40.7	33.5
MapTracker [10]	72	45.9	30.0	45.1	40.3
MapExpert [45]	100	46.7	**34.1**	45.1	42.0
RelMap-TF (Ours)	72	**46.8**	32.5	**47.6**	**42.3**

Qualitative Results. We present qualitative comparisons on the nuScenes dataset in Fig. 4 and Fig. 5. Compared to MapTRv2 [23] and MapQR [16], our RelMap-SF model produces more accurate, regular, and spatially coherent predictions in Fig. 4. As shown in Fig. 5, RelMap-TF yields more accurate shapes and spatial arrangements than MapTracker [10] and StreamMapNet [9]. These results confirm that the spatial and

TABLE III

RESULTS ON ARGOVERSE 2 VALIDATION SET.

Methods	Dim	AP_{ped}		AP_{div}		AP_{bou}		mAP	
MapTRv2 [23]	2 | 3	62.9	60.7	72.1	68.9	67.1	64.5	67.4	64.7
MapQR [16]	2 | 3	64.3	60.1	72.3	71.2	68.1	66.2	68.2	65.9
HIMap [8]	2 | 3	**69.0**	**66.7**	69.5	68.3	**70.3**	**70.3**	69.6	68.4
RelMap-SF (Ours)	2 | 3	65.3	64.4	**74.2**	**72.7**	70.1	68.3	**69.9**	**68.5**

semantic priors remain beneficial in the temporal setting, enabling more stable and coherent predictions across frames.

D. Ablation Studies

We conduct ablation studies on the nuScenes dataset to assess the effectiveness of individual components and to validate key design choices. All experiments are based on the RelMap-SF model and trained for 24 epochs.

Effectiveness of Key Components. We first investigate the impact of the Class-aware Spatial Relation Prior and the MoE-based Semantic Prior by progressively introducing them into the baseline. As shown in Table IV, incorporating the Spatial Relation Prior improves mAP by 0.6, demonstrating its ability to enhance the model's capacity to capture spatial dependencies among map elements. Adding Class-aware Modulation further enhances mAP by 0.9, demonstrating the advantage of learning spatial relations in a class-specific manner for improved relational reasoning. Finally, using the MoE-based Semantic Prior yields an additional 1.3-point mAP improvement, highlighting its effectiveness in capturing class-specific feature variations. By routing map instances to specialized expert decoders, this prior enables more adaptive and expressive feature decoding, leading to better overall performance.

MoE Design Variants. To explore different MoE configurations, we experiment with varying the number of experts and the routing mechanisms. We implement a *Vanilla MoE* following a standard MoE design, where a gating network selects the top-1 expert among three candidates for each instance. We also evaluate an MoE design based on MapExpert [45] (*MapExpert MoE*), which expands the number of experts to eight and computes a weighted combination of the top-2 expert outputs. Additionally, we implement a variant of our MoE design that incorporates a shared expert (*MoE w/ Shared Expert*) [42]. As shown in Table V, our MoE design, which utilizes only three experts and does not require an additional routing network, achieves the highest mAP. While MapExpert MoE has a more flexible expert selection mechanism, its performance is sensitive to the auxiliary expert balance loss, necessitating careful hyperparameter tuning. Our approach eliminates the need for additional loss terms, reducing the complexity of tuning while maintaining strong performance. Adding shared expert results in an accuracy drop, indicating that shared feature decoding may weaken the effectiveness of our class-specific prior.

TABLE IV

EFFECTIVENESS OF KEY COMPONENTS IN RELMAP-SF. †: REPRODUCED RESULT OF MAPQR [16] USING THE OFFICIAL CODE.

Class-aware Spatial Relation Prior		Semantic Prior	mAP
Relation Encoding	*Class-aware Mod.*	(MoE-based)	
-	-	-	65.6†
✓	-	-	66.2
✓	✓	-	67.1
-	-	✓	67.3
✓	✓	✓	68.4

TABLE V

INFLUENCE OF MOE DESIGN VARIANTS IN RELMAP-SF

	# of Experts	Separate Routing	mAP
Vanilla MoE	3	✓	66.3
MapExpert MoE	8	✓	65.9
MoE w/ shared expert	3 + 1 (shared)		66.7
Ours	3		**67.3**

V. CONCLUSIONS

We introduce **RelMap**, a framework for online vectorized map construction that incorporates a *Class-aware Spatial Relation Prior* and a *MoE-based Semantic Prior* to enhance spatial coherence and class-specific feature refinement. We instantiate our framework in two variants: RelMap-SF for single-frame map construction and RelMap-TF for temporal map construction. RelMap models achieve state-of-the-art performance on the nuScenes and Argoverse 2 datasets, highlighting the effectiveness and generalization of the proposed priors. In future work, we plan to explore the integration of external structured knowledge, such as Standard Definition maps, to further guide relational and semantic priors, especially for improving long-range and temporal consistency.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (Award No. 2346267), and the FHWA Center for Excellence on New Mobility and Automated Vehicles.

REFERENCES

[1] L. Gao, X. Xia, Z. Zheng, H. Xiang, Z. Meng, X. Han, Z. Zhou, Y. He, Y. Wang, Z. Li, *et al.*, "Cooperative localization in transportation 5.0," *IEEE Transactions on Intelligent Vehicles*, 2024.

[2] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6531–6543, 2022.

[3] Z. Zhou, H. Xiang, Z. Zheng, S. Z. Zhao, M. Lei, Y. Zhang, T. Cai, X. Liu, J. Liu, M. Bajji, *et al.*, "V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction," *arXiv preprint arXiv:2412.01812*, 2024.

[4] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.

[5] Z. Huang, H. Liu, and C. Lv, "Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 3903–3913.

[6] X. Liu, S. Wang, W. Li, R. Yang, J. Chen, and J. 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. 14 812–14 821.

[7] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "Maptr: Structured modeling and learning for online vectorized hd map construction," *arXiv preprint arXiv:2208.14437*, 2022.

[8] Y. Zhou, H. Zhang, J. Yu, Y. Yang, S. Jung, S.-I. Park, and B. Yoo, "Himap: Hybrid representation learning for end-to-end vectorized hd map construction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 396–15 406.

[9] T. Yuan, Y. Liu, Y. Wang, Y. Wang, and H. 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.

[10] J. Chen, Y. Wu, J. Tan, H. Ma, and Y. Furukawa, "Maptracker: Tracking with strided memory fusion for consistent vector hd mapping," in *European Conference on Computer Vision*. Springer, 2024, pp. 90–107.

[11] W. Ding, L. Qiao, X. Qiu, and C. 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.

[12] A. Shi, Y. Cai, X. Chen, J. Pu, Z. Fu, and H. Lu, "Globalmapnet: An online framework for vectorized global hd map construction," *arXiv preprint arXiv:2409.10063*, 2024.

[13] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 760–13 769.

[14] S. Chen, T. Cheng, X. Wang, W. Meng, Q. Zhang, and W. Liu, "Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer," *arXiv preprint arXiv:2206.04584*, 2022.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[16] Z. Liu, X. Zhang, G. Liu, J. Zhao, and N. Xu, "Leveraging enhanced queries of point sets for vectorized map construction," in *European Conference on Computer Vision*. Springer, 2024, pp. 461–477.

[17] Y. Cai, W. Dong, Z. Liu, H. Wang, and L. Chen, "Homap: End-to-end vectorized hd map construction with high-order modeling," *IEEE Transactions on Intelligent Vehicles*, pp. 1–11, 2024.

[18] J. Yang, M. Jiang, S. Yang, X. Tan, Y. Li, E. Ding, H. Wang, and J. Wang, "Mgmapnet: Multi-granularity representation learning for end-to-end vectorized hd map construction," *arXiv preprint arXiv:2410.07733*, 2024.

[19] S. Li and H. Li, "Deep generative modeling based on vae-gan for 3d indoor scene synthesis," *International Journal of Computer Games Technology*, vol. 2023, no. 1, p. 3368647, 2023.

[20] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3588–3597.

[21] X. Hou, M. Liu, S. Zhang, P. Wei, B. Chen, and X. Lan, "Relation detr: Exploring explicit position relation prior for object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 89–105.

[22] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[23] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Maptrv2: An end-to-end framework for online vectorized hd map construction," *International Journal of Computer Vision*, pp. 1–23, 2024.

[24] Z. Zhang, Y. Zhang, X. Ding, F. Jin, and X. Yue, "Online vectorized hd map construction using geometry," in *European Conference on Computer Vision*. Springer, 2024, pp. 73–90.

[25] R. Wang, X. Lu, X. Liu, X. Zou, T. Cao, and Y. Li, "Priormapnet: Enhancing online vectorized hd map construction with priors," *arXiv preprint arXiv:2408.08802*, 2024.

[26] S. Wang, F. Jia, W. Mao, Y. Liu, Y. Zhao, Z. Chen, T. Wang, C. Zhang, X. Zhang, and F. Zhao, "Stream query denoising for vectorized hd-map construction," in *European Conference on Computer Vision*. Springer, 2024, pp. 203–220.

[27] J. Yang, S. Yang, X. Tan, and H. Wang, "Histrackmap: Global vectorized high-definition map construction via history map tracking," *arXiv preprint arXiv:2503.07168*, 2025.

[28] N. Peng, X. Zhou, M. Wang, X. Yang, S. Chen, and G. Chen, "Prevpredmap: Exploring temporal modeling with previous predictions for online vectorized hd map construction," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 8134–8143.

[29] N. Kim, H. Seong, D. Ji, and S. Jang, "Unveiling the hidden: Online vectorized hd map construction with clip-level token interaction and propagation," 2024. [Online]. Available: https://arxiv.org/abs/2411.11002

[30] N. Peng, X. Zhou, M. Wang, G. Chen, and W. Xu, "Uni-prevpredmap: Extending prevpredmap to a unified framework of prior-informed modeling for online vectorized hd map construction," *arXiv preprint arXiv:2504.06647*, 2025.

[31] Z. Jiang, Z. Zhu, P. Li, H.-a. Gao, T. Yuan, Y. Shi, H. Zhao, and H. Zhao, "P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors," *IEEE Robotics and Automation Letters*, 2024.

[32] K. Z. Luo, X. Weng, Y. Wang, S. Wu, J. Li, K. Q. Weinberger, Y. Wang, and M. Pavone, "Augmenting lane perception and topology understanding with standard definition navigation maps," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 4029–4035.

[33] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5177–5186.

[34] C. Jiang, H. Xu, X. Liang, and L. Lin, "Hybrid knowledge routed modules for large-scale object detection," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[35] T. Liu, Z. Qian, J. Berrevoets, and M. van der Schaar, "Goggle: Generative modelling for tabular data by learning relational structure," in *The Eleventh International Conference on Learning Representations*, 2023.

[36] X. Hao, D. Huang, J. Lin, and C.-Y. Lin, "Relation-enhanced detr for component detection in graphic design reverse engineering," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 4785–4793.

[37] J. Shin, H. Jeong, F. Rameau, and D. Kum, "Instagram: Instance-level graph modeling for vectorized hd map learning," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[38] Z. Xu, K.-Y. K. Wong, and H. Zhao, "Insmapper: Exploring inner-instance information for vectorized hd mapping," in *European Conference on Computer Vision*. Springer, 2024, pp. 296–312.

[39] Y. Luo, C. Zhou, Y. Yang, E. Li, C. Zheng, S. Mei, S. Cui, and Z. Li, "Reltopo: Enhancing relational modeling for driving scene topology reasoning," *arXiv preprint arXiv:2506.13553*, 2025.

[40] H. Hu, J. Xu, F. Wang, T. Li, Y. Wang, L. Hu, and Z. Zhang, "Fastmap: Fast queries initialization based vectorized hd map reconstruction framework," *arXiv preprint arXiv:2503.05492*, 2025.

[41] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[42] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, *et al.*, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.

[43] X. Wu, S. Huang, and F. Wei, "Mixture of lora experts," *arXiv preprint arXiv:2404.13628*, 2024.

[44] S. Dou, E. Zhou, Y. Liu, S. Gao, J. Zhao, W. Shen, Y. Zhou, Z. Xi, X. Wang, X. Fan, *et al.*, "Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment," *arXiv preprint arXiv:2312.09979*, vol. 4, no. 7, 2023.

[45] D. Zhang, D. Chen, P. Zhi, Y. Chen, Z. Yuan, C. Li, R. Zhou, Q. Zhou, *et al.*, "Mapexpert: Online hd map construction with simple and efficient sparse map element expert," *arXiv preprint arXiv:2412.12704*, 2024.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[48] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," *arXiv preprint arXiv:2301.00493*, 2023.