

LaneDiffusion++: Feature-Level Generative Priors for Structured Lane Graph Learning

Zijie Wang, Wei Zhang, Weiming Zhang, Xiao Tan, Jie Feng, Xiaonan Luo, Guanbin Li

Abstract—Learning structured lane topology from visual observations is challenging due to structural ambiguity, occlusions, and incomplete lane markings in complex urban environments. Most existing centerline graph learning methods rely on deterministic regression, which implicitly assumes a single optimal topology for each scene and therefore struggles under uncertain visual conditions. In this work, we propose LaneDiffusion++, a unified feature-level generative framework that reformulates centerline graph learning as the conditional restoration of prior-enhanced Bird’s Eye View (BEV) features, rather than directly generating vectorized graphs in a discrete output space. Specifically, LaneDiffusion++ first injects structural lane priors into intermediate BEV features through a Lane Prior Injection Module (LPIM), and then employs a Lane Prior Diffusion Module (LPDM) to perform conditional generative modeling in feature space, progressively restoring topology-aware representations while capturing multiple plausible lane configurations. To further exploit the stochasticity of diffusion inference, we introduce a Lane Hypothesis Aggregation Module (LHAM), which explicitly consolidates multiple diffusion-generated feature hypotheses into a stable topology-aware representation before decoding. Extensive experiments on large-scale benchmarks demonstrate consistent improvements in both geometric precision and topological consistency over strong deterministic and generative baselines. These results suggest that feature-level generative modeling, together with uncertainty-aware hypothesis aggregation, provides an effective paradigm for structured topology reasoning in autonomous perception.

Index Terms—Autonomous driving, lane centerline graph learning, structured topology learning, diffusion models, generative modeling, bird’s eye view perception

I. INTRODUCTION

STRUCTURED scene understanding is a fundamental problem in spatial intelligence and autonomous perception systems. In autonomous driving, environmental structure is commonly represented as graphs defined over semantic primitives, where nodes encode geometric entities and edges capture logical connectivity. Among various structural representations, lane centerline graphs provide a compact yet expressive abstraction of drivable topology, serving as a critical interface between perception and downstream motion planning and control

Zijie Wang and Guanbin Li are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, 510006, China. They are also with Shenzhen Loop Area Institute, Shenzhen, 518038, China. Guanbin Li is also with Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, China.

Wei Zhang, Weiming Zhang and Xiao Tan are with Baidu Inc., Shenzhen, 518066, China.

Jie Feng is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi’an 710071, China.

Xiaonan Luo is with Guilin University of Electronic Technology, Guilin, 541004, China.

Corresponding Author is Guanbin Li (liguanbin@mail.sysu.edu.cn).

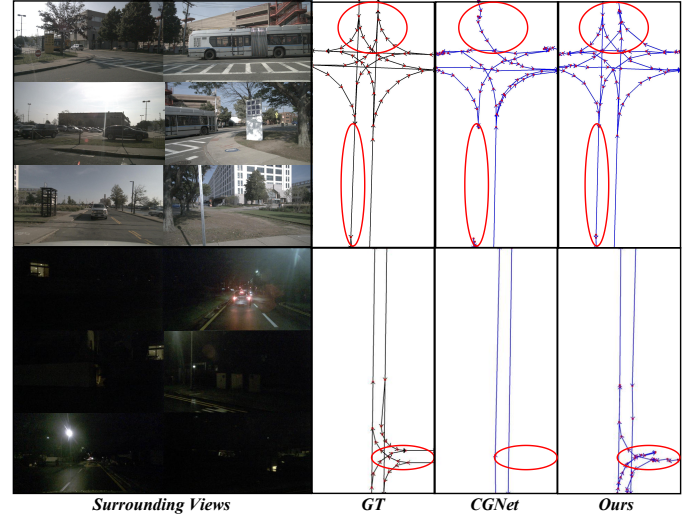


Fig. 1. **Motivation.** In challenging scenarios with occlusion or ambiguous visual evidence, deterministic centerline graph learning methods (e.g. CGNet) may commit to an incorrect topology. In contrast, a generative formulation can model multiple plausible structural interpretations and provide more robust topology reasoning.

modules [1], [2]. Accurate recovery of such graphs requires not only precise localization of geometric lane segments but also robust reasoning about their long-range connectivity under occlusion, truncation and ambiguous visual evidence.

Recent advances in Bird’s Eye View (BEV)-based perception have significantly improved map element detection and vectorized representation learning [3]–[5]. Nevertheless, most existing centerline graph learning approaches adopt deterministic regression paradigms [6], directly regressing vectorized lane segments and subsequently inferring connectivity relationships. While effective under clear observations, such deterministic formulations implicitly assume that the underlying topology is uniquely determined by local visual observations. In real-world driving environments, however, this assumption is often violated. Lane markings may be partially missing, heavily occluded by dynamic objects or visually ambiguous at intersections and merges. As a result, the same observation may admit multiple plausible topological interpretations. From this perspective, centerline graph learning is inherently a *structured uncertainty reasoning* problem, and should be modeled as estimating a conditional distribution over possible graph structures, rather than predicting a single deterministic output.

Generative modeling provides a natural way to capture such ambiguity. In particular, diffusion models have shown strong capability in modeling complex multi-modal distributions [7].

A straightforward idea is therefore to directly generate vectorized lane graphs using diffusion [8]. However, this is intrinsically difficult because vectorized topology resides in a discrete structured space with variable output cardinality, permutation ambiguity, and unstable combinatorial optimization. These properties make direct generative modeling of graph outputs considerably harder than standard continuous generation tasks. This observation motivates a different question: instead of generating discrete topology directly, can we perform generative reasoning in a continuous representation space where structural semantics are easier to preserve and refine?

In this paper, we answer this question with *LaneDiffusion++*, a feature-level generative framework for structured centerline graph learning. The key idea is to shift generative modeling from the vector output space to the BEV feature space. Specifically, we first inject lane structural priors into intermediate BEV features through a *Lane Prior Injection Module* (LPIM), obtaining prior-enhanced representations that encode geometric and topological cues. We then learn a conditional diffusion process with a *Lane Prior Diffusion Module* (LPDM) to restore these prior-enhanced features from the original BEV features. In this way, centerline graph learning is reformulated as *conditional feature restoration* rather than direct graph synthesis, enabling more stable generative reasoning over complex lane topology.

A central observation of this work is that diffusion inference naturally produces multiple stochastic feature hypotheses, each corresponding to a plausible structural interpretation of the same ambiguous scene. While such stochasticity is often treated merely as sampling variation, we show that it can be exploited as a valuable source of structural uncertainty for topology reasoning. To this end, we introduce the *Lane Hypothesis Aggregation Module* (LHAM), which explicitly aggregates multiple diffusion-generated hypotheses in an uncertainty-aware manner. By modeling cross-hypothesis statistics and fusing them with the conditional BEV representation, LHAM emphasizes structurally consistent patterns while suppressing noisy or unreliable activations. This design turns stochastic diffusion sampling from a source of instability into a mechanism for robust structured inference, leading to more stable and topology-consistent predictions in challenging scenarios such as occlusions, weak lane evidence and complex junctions.

Furthermore, *LaneDiffusion++* adopts a structured diffusion transition defined between semantically meaningful BEV feature states, instead of the conventional formulation that progressively corrupts data into Gaussian noise. This feature-state transition preserves structural information throughout the diffusion process and allows significantly shorter sampling chains, yielding a favorable trade-off between inference efficiency and prediction performance. Extensive experiments on nuScenes [9] and Argoverse2 [10] demonstrate that *LaneDiffusion++* consistently improves both geometric accuracy and topological consistency over strong deterministic and generative baselines. Beyond quantitative gains, our analyses show that feature-level generative modeling is particularly effective under structural ambiguity, and that uncertainty-aware hypothesis aggregation is critical for converting stochastic diffusion

diversity into robust topology estimation. In summary, our contributions are threefold:

- We reformulate centerline graph learning as a conditional generative topology inference problem in BEV feature space, providing a principled perspective for structured perception under ambiguous observations.
- We propose *LaneDiffusion++*, a unified feature-level generative framework that combines prior injection, conditional diffusion, and structured feature-state transitions for robust lane graph learning.
- We introduce an uncertainty-aware Lane Hypothesis Aggregation Module that explicitly consolidates multiple stochastic diffusion hypotheses into stable and topology-consistent representations, and validate its effectiveness through extensive experiments and analyses.

This work is an extension of our previous conference paper published at International Conference on Computer Vision (ICCV 2025) [11], with substantial enhancements in methodology, analysis and experimental validation. Compared with the conference version, which established the feature-level prior diffusion paradigm, this extended version makes three substantial advances. First, we provide a more general and rigorous formulation of feature-level generative topology learning. Second, we newly introduce LHAM to explicitly exploit stochastic diffusion hypotheses for robust structured inference. Third, we expand the experimental evaluation with additional ablation and robustness studies, and provide deeper theoretical and empirical analyses of structured diffusion transitions and aggregation-based feature refinement for topology reasoning.

II. RELATED WORKS

A. Lane Graph Learning

Lane graph learning aims to recover structured lane centerlines together with their topological relationships from visual observations. Early studies [12]–[14] mainly focused on aerial or satellite imagery, where road structures are typically coarse and large-scale but lack fine-grained geometric details due to limited resolution, perspective ambiguity and frequent occlusions. Although these methods demonstrated the feasibility of graph-based road extraction, their representations and sensing settings are not directly suitable for high-precision online perception in autonomous driving.

With the development of onboard perception systems, recent works have shifted toward online lane graph construction from vehicle-mounted sensors. Several approaches infer lane centerlines directly from monocular or multi-view images. Methods such as STSU [15] and its subsequent extensions [16]–[21] predict lane structures from monocular observations, but their limited field of view often leads to incomplete topology reasoning, especially in long-range or highly occluded scenarios. To alleviate this issue, surround-view BEV-based methods have been introduced to provide more complete spatial context. For example, TopoNet [22] adopts surround-view modeling with explicit graph representations, while LaneGAP [23] and CGNet [6] improve structural continuity through path linking, connectivity reasoning, and segment refinement.

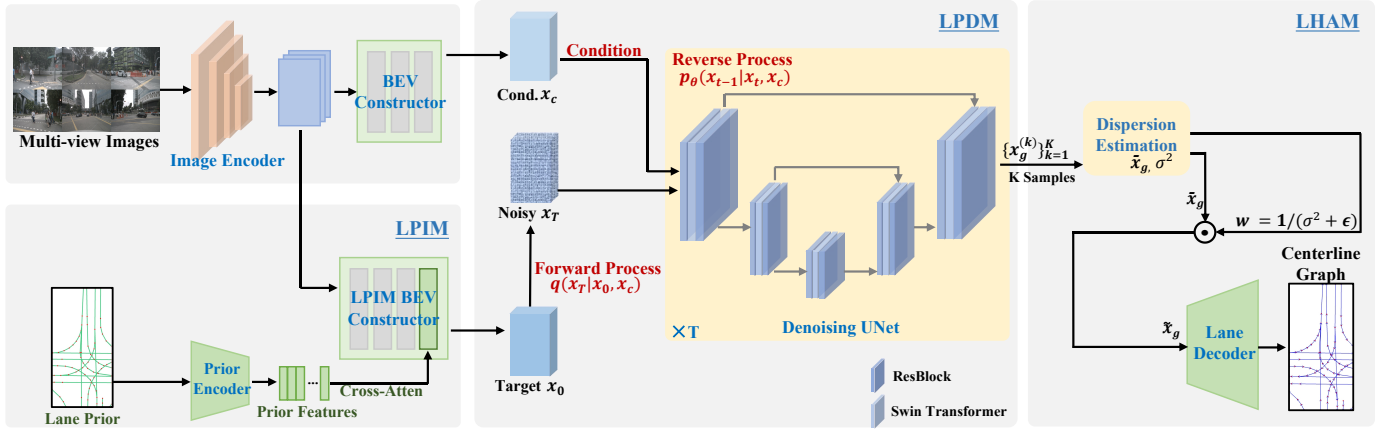


Fig. 2. Overview of LaneDiffusion++. The framework consists of three key components: (i) a Lane Prior Injection Module (LPIM), which injects lane priors into BEV features to construct the diffusion target; (ii) a Lane Prior Diffusion Module (LPDM), which models the prior-enhanced BEV feature via a diffusion process conditioned on the original BEV feature; and (iii) the proposed Lane Hypothesis Aggregation Module (LHAM), which operates during inference to aggregate multiple stochastic diffusion predictions in an uncertainty-aware manner, producing more stable lane topology estimation. The modular design of LaneDiffusion++ makes it a flexible add-on for BEV feature-based architectures. Ground-truth lane priors are used only during training to construct x_0 . During inference, LaneDiffusion++ takes only multi-view images as input and restores topology-aware features from x_c .

More recently, auto-regressive formulations [24]–[26] have modeled lane graph construction as a sequential generation process, enabling more explicit topology prediction and stronger structural dependency modeling. These methods move beyond local detection by introducing structured decoding order and graph-aware generation strategies. Nevertheless, most existing lane graph learning approaches remain fundamentally deterministic and aim to predict a single optimal topology for each scene. Such formulations can be suboptimal under ambiguous observations, where lane markings may be incomplete, occluded or visually weak, and multiple plausible structural interpretations may coexist. In this sense, the challenge of lane graph learning is not only geometric reconstruction but also uncertainty-aware topology reasoning. This motivates exploring generative formulations that explicitly model the multi-modality of lane structures instead of committing to a single deterministic prediction.

B. Prior Information

Incorporating prior information has been widely studied to enhance perception and map understanding in autonomous driving. A common strategy is to introduce auxiliary cues that supplement incomplete visual evidence. Temporal priors are leveraged in methods such as StreamMapNet [27], which propagates scene representations across frames to improve temporal consistency and map stability. Spatial priors are explored through collaborative perception frameworks [28]–[31], where multiple agents share complementary observations for more complete scene understanding. Other studies introduce priors from heterogeneous external sources, including 2D lane annotations [32], satellite imagery [33], historical maps [34], and standard-definition maps [35]–[37], all of which provide additional structural guidance for lane reasoning. Trajectory-based priors derived from large-scale driving data have also been explored to improve map construction and topology estimation [38]. These approaches demonstrate that prior in-

formation is highly valuable for resolving structural ambiguity beyond what is directly observable in the current frame.

However, most existing prior-guided BEV perception methods introduce auxiliary cues as additional information sources, which may increase system complexity and introduce alignment, synchronization, or availability issues. Moreover, these priors are usually used to enhance perception features directly, rather than being formulated as latent structural targets for generative learning. In contrast, LaneDiffusion++ uses ground-truth lane priors only during training to construct prior-enhanced BEV features as diffusion targets. During inference, no external map prior or ground-truth lane graph is required. LPDM restores topology-aware features solely from image-derived BEV features. Therefore, our method turns lane priors into feature-space generative supervision, rather than treating them as deployment-time auxiliary inputs.

C. Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) [7] have recently emerged as powerful generative models and have been widely adopted in a variety of vision tasks [39]–[47], including image synthesis [48], segmentation [49], and object detection [50]. Beyond pixel-level generation, diffusion models have shown strong capability in modeling complex distributions and multi-modal outputs, making them particularly suitable for problems involving ambiguity and uncertainty.

In autonomous driving, diffusion-based approaches have been explored in motion prediction [51] and scene generation [52]–[55], where generative modeling helps capture diverse future outcomes. More recently, diffusion models have also been introduced to map understanding tasks. For example, DiffMap [56] formulates map segmentation as a generative process, while PolyDiffuse [8] explores diffusion for vectorized map reconstruction.

Despite these advances, existing diffusion-based mapping methods typically formulate diffusion either over rasterized map representations or directly over vectorized outputs. For

structured lane graph learning, direct diffusion over vectorized topology remains challenging due to variable output cardinality, permutation ambiguity, and the discrete combinatorial nature of graph prediction. As a result, the benefits of diffusion-based uncertainty modeling may be difficult to fully realize in the final output space. In contrast, LaneDiffusion++ performs diffusion in the BEV feature space, where structural lane priors can be injected, preserved, and progressively refined in a continuous representation. This feature-level formulation is more compatible with modern BEV perception pipelines and provides a more stable substrate for topology-aware generative reasoning. Furthermore, unlike prior diffusion-based mapping approaches that mainly focus on generation itself, our extended framework explicitly studies how multiple stochastic diffusion hypotheses can be aggregated for robust structured inference, which is crucial for stable topology estimation under ambiguous observations.

III. METHODOLOGY

In this section, we present LaneDiffusion++, a unified framework for structured centerline graph learning through conditional generative modeling in BEV feature space. We first define the lane graph prediction task (Sec. III-A) and introduce a feature-level generative formulation (Sec. III-B) that models structural uncertainty in continuous latent representations rather than directly in discrete graph space. Based on this formulation, LaneDiffusion++ consists of three components: (i) the Lane Prior Injection Module (LPIM), which constructs structurally grounded diffusion targets (Sec. III-D); (ii) the Lane Prior Diffusion Module (LPDM), which performs conditional generative restoration of prior-enhanced BEV features (Sec. III-E); and (iii) the Lane Hypothesis Aggregation Module (LHAM), which consolidates multiple stochastic feature hypotheses into a stable topology-aware representation (Sec. III-F). Finally, we describe the staged optimization strategy used to train the overall framework (Sec. III-G).

A. Task Definition

The centerline graph learning task aims to recover structured lane topology from the multi-view images captured by a vehicle's surround-view cameras. Specifically, the objective is two-fold: first, to detect lane centerline segments in the BEV representation without overlap; second, to infer the connectivity among these segments to construct a coherent lane graph. Depending on the evaluation protocol and downstream usage (see Sec. IV-B), we consider two complementary levels of structural granularity: a segment-level graph and a point-level graph. At the segment level, the graph is represented as

$$\mathbf{G} = (\mathbf{V}, \mathbf{E}), \quad (1)$$

where each vertex in $\mathbf{V} = \{\mathbf{P}_i\}_{i=1}^{|\mathbf{V}|}$ corresponds to an individual lane segment. Each segment $\mathbf{P}_i$ is represented as an ordered point sequence

$$\mathbf{P}_i = [\mathbf{p}_{i,1}, \mathbf{p}_{i,2}, \dots, \mathbf{p}_{i,N_i}], \mathbf{p}_{i,j} = (x_{i,j}, y_{i,j}) \in \mathbb{R}^2, \quad (2)$$

and the edge set $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ encodes the connectivity among these segments. For finer structural modeling, we additionally define a point-level graph

$$\ddot{\mathbf{G}} = (\ddot{\mathbf{V}}, \ddot{\mathbf{E}}), \quad (3)$$

where $\ddot{\mathbf{V}}$ contains all sampled points along lane polylines and $\ddot{\mathbf{E}}$ represents the point-wise connectivity relationships. The segment-level graph captures global lane topology, while the point-level graph provides a more detailed representation of local geometric and structural continuity. These two graph representations serve as the structured supervision targets for lane decoding, whereas the proposed generative modeling is performed in the intermediate BEV feature space.

B. Feature-level Generative Topology Learning

Learning structured lane topology from visual observations is fundamentally different from conventional dense prediction tasks. Unlike pixel-wise regression or object detection with fixed-dimensional outputs, centerline graph learning requires predicting a structured graph composed of geometric primitives and their connectivity. As defined in Sec. III-A, the desired output is a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ denotes lane centerline segments and $\mathbf{E}$ encodes their topological relations.

Most existing approaches formulate this problem as deterministic prediction and directly learn a mapping

$$\mathbf{G} = f_{\theta}(I), \quad (4)$$

where I denotes the multi-view input images. While effective in clear scenes, this formulation implicitly assumes that each observation corresponds to a single optimal topology. In real driving environments, however, this assumption is often violated. Lane markings may be partially missing, severely occluded, or visually ambiguous at merges, splits, and intersections. Under such conditions, multiple plausible structural interpretations may explain the same observation. Therefore, centerline graph learning is inherently a structured uncertainty reasoning problem, and a more principled view is to model the conditional distribution $p(\mathbf{G} | I)$, rather than regressing a single deterministic graph.

Directly modeling $p(\mathbf{G} | I)$ in the vectorized graph space is challenging for three reasons. First, graph outputs have variable cardinality, making the generative process difficult to define in a fixed-dimensional output space. Second, the ordering of graph elements is often ambiguous, leading to permutation sensitivity during learning. Third, the graph space is discrete and combinatorial, which makes stable continuous optimization and uncertainty modeling considerably more difficult than in standard dense generation tasks. These challenges suggest that, although the final goal remains graph prediction, the uncertainty associated with topology may be better modeled in a continuous latent space than directly in the discrete output space. Based on this observation, we reformulate topology learning as conditional generative modeling in BEV feature space. Let $z \in \mathbb{R}^{H \times W \times C}$ denote the BEV feature representation extracted from the input images I . Instead of directly generating vectorized graphs, we construct a prior-enhanced feature representation

$$\tilde{z} = z + \mathcal{P}(\mathbf{G}), \quad (5)$$

where $\mathcal{P}(\mathbf{G})$ denotes a structural prior encoding function that injects geometric and topological cues derived from the ground-truth graph $\mathbf{G}$ into the BEV feature map. Under this formulation, structural topology is embedded into a spatially aligned continuous representation, enabling the model to perform generative reasoning in feature space.

The learning objective then becomes modeling the conditional distribution

$$p_{\theta}(\tilde{z} | z), \quad (6)$$

which represents the distribution of structurally consistent BEV features conditioned on the observed scene representation. In other words, instead of directly generating a graph, the model learns to restore topology-aware latent features whose decoded outputs correspond to plausible lane structures. The final graph prediction is still obtained through deterministic decoding on the recovered feature representation. This design does not replace graph prediction itself; rather, it shifts uncertainty modeling from the discrete graph space to an intermediate continuous latent space where structural information can be more stably represented, perturbed, and restored.

This feature-level generative formulation offers several advantages. First, continuous BEV features avoid the combinatorial complexity of direct graph generation. Second, structural priors can be injected in a differentiable and spatially aligned manner, providing stable supervision for generative learning. Third, stochastic generative inference yields multiple feature hypotheses that can be consolidated by LHAM for robust prediction. This connection is especially important in ambiguous scenes, where robust prediction should depend not on a single stochastic sample, but on the structural consensus among multiple plausible hypotheses. This observation motivates the explicit hypothesis consolidation mechanism introduced later in Sec. III-F.

Building upon this formulation, LaneDiffusion++ implements conditional diffusion in the prior-enhanced BEV feature space to approximate $p_{\theta}(\tilde{z} | z)$. The recovered feature representation is then decoded into vectorized lane centerlines and their connectivity relationships, forming a unified pipeline for uncertainty-aware structured topology learning.

C. Framework Overview

The overall architecture of LaneDiffusion++ is illustrated in Fig. 2. The framework is designed to be seamlessly integrated with modern BEV-based perception pipelines. Given multi-view image features extracted by a backbone encoder, a BEV construction module first projects them into a unified BEV representation. On top of this representation, LaneDiffusion++ performs structured topology reasoning through three complementary stages: structural prior construction, stochastic generative restoration, and uncertainty-aware hypothesis aggregation.

The first component, the *Lane Prior Injection Module* (LPIM), injects structural lane priors into intermediate BEV features and constructs a prior-enhanced feature representation. This representation serves as the diffusion target and provides a structured latent anchor for generative learning.

TABLE I
NOTATION OF MAIN BEV FEATURE TENSORS USED IN
LANEDIFFUSION++.

Symbol	Description
z	Image-derived BEV feature in the general formulation.
$\tilde{z}$	Prior-enhanced BEV feature in the general formulation.
$\mathbf{x}_c$	Original BEV feature used as the LPDM condition.
$\mathbf{x}_0$	Prior-enhanced BEV feature used as the LPDM target.
$\mathbf{x}_t$	Intermediate feature state at diffusion step t .
$\mathbf{x}_{\text{res}}$	Residual feature between $\mathbf{x}_c$ and $\mathbf{x}_0$.
$\mathbf{x}_g$	Feature restored by one LPDM reverse trajectory.
$\bar{\mathbf{x}}_g$	Mean of multiple restored feature hypotheses.
σ^2	Cross-hypothesis dispersion map in LHAM.
$\mathbf{w}$	Dispersion-aware confidence weight in LHAM.
$\tilde{\mathbf{x}}_g$	Aggregated feature produced by LHAM.

The second component, the *Lane Prior Diffusion Module* (LPDM), models the conditional distribution of prior-enhanced features given the original BEV representation. Rather than directly synthesizing vectorized lane graphs, LPDM performs conditional feature restoration in BEV space, enabling the model to capture multiple plausible topology hypotheses while avoiding the instability of direct graph-space generation.

The third component, the *Lane Hypothesis Aggregation Module* (LHAM), consolidates multiple stochastic diffusion-generated feature hypotheses into a stable consensus representation. By explicitly exploiting cross-hypothesis agreement, LHAM transforms diffusion diversity into robust topology inference, especially under occlusions, weak lane evidence, and complex junction structures.

Finally, the aggregated feature representation is passed to the lane decoding head, which outputs the segment-level graph $\mathbf{G}$ and the point-level graph $\tilde{\mathbf{G}}$ defined in Sec. III-A. Together, these components form a unified framework that models centerline graph learning as conditional generative inference in BEV feature space. For readability, Tab. I summarizes the main BEV feature tensors used throughout the methodology section.

D. Lane Prior Injection Module (LPIM)

To enable feature-level generative topology learning, it is necessary to construct a structurally meaningful latent target for diffusion modeling. The role of LPIM is not merely to enhance perception features, but to map explicit lane topology into a BEV-space representation suitable for generative restoration. To this end, we introduce the *Lane Prior Injection Module* (LPIM), which embeds structural lane priors derived from the ground-truth centerline graph into intermediate BEV features. The resulting prior-enhanced BEV feature serves as the diffusion target for the generative process. LPIM consists of three steps: prior knowledge selection, prior knowledge encoding, and prior-feature injection.

a) *Prior Knowledge Selection*: A key design choice concerns the source of structural priors used to guide feature-level diffusion. Some previous methods utilize external signals such as map-based priors [37] or crowdsourced trajectories [38].

Although these sources can provide useful structural cues, they may also introduce noise, coverage gaps or alignment errors. In our setting, the purpose of LPIM is to construct a high-fidelity structural target for diffusion learning. Therefore, during training, we directly use the ground-truth lane graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ defined in Sec. III-A. Each lane segment $\mathbf{P}_i \in \mathbf{V}$ is represented as an ordered polyline

$$\mathbf{P}_i = [\mathbf{p}_{i,1}, \dots, \mathbf{p}_{i,N_i}], \quad \mathbf{p}_{i,j} = (x_{i,j}, y_{i,j}), \quad (7)$$

which provides accurate geometric and topological supervision in BEV space. Importantly, these ground-truth lane priors are used only during training to construct the generative target for LPDM. They are not used during inference, where the model takes only the image-derived BEV feature $\mathbf{x}_c$ as input and restores the topology-aware feature $\mathbf{x}_g$ through LPDM. Thus, LaneDiffusion++ requires no external map prior, ground-truth lane graph, or test-time topology supervision during deployment.

b) Prior Knowledge Encoding: Given the ground-truth lane centerlines $\{\mathbf{P}_i\}_{i=1}^{|\mathbf{V}|}$, we first convert them into a unified representation suitable for feature injection. Each centerline polyline is uniformly resampled to a fixed number of points N :

$$\mathbf{P}_i = [\mathbf{p}_{i,1}, \dots, \mathbf{p}_{i,N}], \quad \mathbf{p}_{i,j} = (x_{i,j}, y_{i,j}). \quad (8)$$

We then apply multi-frequency sinusoidal positional encoding [57] to each point in order to capture both coarse global geometry and fine local variations along lane structures. The encoded point sequence is further processed by a *Prior Encoder*, implemented as a Transformer encoder with L multi-head self-attention layers. This design is particularly suitable for lane topology modeling because structural relations among centerlines are often non-local: distant lane segments may still be topologically related through merges, splits, and long-range continuity. By applying self-attention over the encoded lane representations, the Prior Encoder captures such relational context before injecting it into BEV space. The resulting output is a set of lane-aware structural features $\mathbf{F}_{\text{prior}} \in \mathbb{R}^{|\mathbf{V}| \times H}$, where H denotes the hidden dimension and each row represents the global structural embedding of one lane centerline.

c) Prior Knowledge Injection: The encoded structural features are injected into intermediate BEV features through cross-attention. Let $\mathbf{F}_{\text{bev}} \in \mathbb{R}^{H \times W \times C}$ denote the BEV feature representation produced by the BEV constructor. During injection, the BEV features act as queries, while the encoded lane priors serve as keys and values. This design allows each BEV location to selectively attend to the most relevant structural priors and to incorporate lane geometry and topology information in a spatially aligned manner. We insert this prior-feature injection step after each spatial cross-attention stage in the BEV construction module, yielding a modified backbone referred to as the *LPIM BEV Constructor*. The output of LPIM is a prior-enhanced BEV feature representation that explicitly encodes lane geometry and topology. This representation serves as the generative target of LPDM and provides a structured latent anchor for subsequent conditional diffusion learning.

E. Lane Prior Diffusion Module (LPDM)

Given the prior-enhanced BEV feature constructed by LPIM, we introduce the *Lane Prior Diffusion Module* (LPDM) to perform conditional generative modeling in feature space. The goal of LPDM is to learn the distribution of structurally grounded BEV features conditioned on the original BEV representation. Unlike conventional output-space generation, the target and condition in our setting already lie in the same semantic BEV feature domain. Therefore, the problem is more naturally interpreted as *structured feature restoration* rather than unconditional synthesis from pure noise. Let $\mathbf{x}_0$ denote the prior-enhanced BEV feature produced by LPIM, and let $\mathbf{x}_c$ denote the original BEV feature extracted from the perception backbone. LPDM learns a conditional diffusion process to model $p_{\theta}(\mathbf{x}_0 | \mathbf{x}_c)$.

a) Preliminaries of DDPMs: Denoising Diffusion Probabilistic Models (DDPMs) [7] define a forward noising process and a learned reverse denoising process. Given clean data $\mathbf{x}_0$, the standard forward process progressively adds Gaussian noise:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (9)$$

where β_t is the variance schedule. The corresponding marginal distribution can be written as

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (10)$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. Equivalently,

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (11)$$

The reverse process learns a parameterized distribution

$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t), \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_t, t)). \quad (12)$$

b) Residual Feature-state Diffusion: Although standard DDPMs provide strong generative flexibility, directly diffusing from a Gaussian prior is unnecessarily expensive for our feature-level restoration setting. In our case, the target feature $\mathbf{x}_0$ and the conditional feature $\mathbf{x}_c$ already reside in the same semantically meaningful space. Thus, instead of modeling a long noise-to-data generation trajectory, we construct a compact diffusion process between two structured BEV feature states. This design is not only more efficient, but also better aligned with the task itself, since the generative objective is to recover a structurally enriched feature from a correlated conditional feature. Following this intuition, LPDM adopts a conditional residual diffusion strategy inspired by ResShift [58]. We define the residual difference

$$\mathbf{x}_{\text{res}} = \mathbf{x}_c - \mathbf{x}_0. \quad (13)$$

The forward process gradually shifts $\mathbf{x}_0$ toward $\mathbf{x}_c$ over a T -step Markov chain:

$$q(\mathbf{x}_t | \mathbf{x}_0, \mathbf{x}_c) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0 + \eta_t \mathbf{x}_{\text{res}}, \kappa^2 \eta_t \mathbf{I}), \quad (14)$$

where η_t controls the interpolation ratio between $\mathbf{x}_0$ and $\mathbf{x}_c$, and κ controls the noise magnitude. The schedule $\{\eta_t\}_{t=1}^T$ is monotonically increasing with $\eta_1 \rightarrow 0$ and $\eta_T \rightarrow 1$, so that

Algorithm 1: Training of LPDM

Input: Diffusion target set $\mathcal{X}_0$, condition feature set $\mathcal{X}_{\text{cond}}$, diffusion step number T , maximum iteration number max_iter

```

1  $\text{cur\_iter} := 1$ ;
2 while  $\text{cur\_iter} \leq \text{max\_iter}$  do
3   Sample  $\mathbf{x}_0 \sim \mathcal{X}_0$  and  $\mathbf{x}_c \sim \mathcal{X}_{\text{cond}}$ ;
4   Sample  $t \sim \text{Uniform}(\{1, 2, \dots, T\})$ ;
5   Sample  $\mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0, \mathbf{x}_c)$  according to Eq. 14;
6   Take a gradient descent step on  $\nabla \mathcal{L}_{\text{diff}}$  in Eq. 19;
7    $\text{cur\_iter} := \text{cur\_iter} + 1$ ;
8 end
```

the diffusion trajectory progressively moves from the prior-enhanced feature toward the conditional feature state. In the reverse process, LPDM learns

$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_c) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, \mathbf{x}_c, t), \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_t, \mathbf{x}_c, t)). \quad (15)$$

The denoising network f_{θ} is implemented as a Swin Transformer based UNet [59]. Following the standard ELBO derivation of diffusion models [7], [60], the objective minimizes the KL divergence between the true posterior and the learned reverse process:

$$\theta^* = \arg \min_{\theta} \sum_t D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0, \mathbf{x}_c) \| p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_c)). \quad (16)$$

Under Eq. 14, the posterior admits a closed-form solution:

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0, \mathbf{x}_c) = \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{\eta_{t-1}}{\eta_t} \mathbf{x}_t + \frac{\gamma_t}{\eta_t} \mathbf{x}_0, \kappa^2 \frac{\eta_{t-1}}{\eta_t} \gamma_t \mathbf{I}\right), \quad (17)$$

where $\gamma_t = \eta_t - \eta_{t-1}$. Since the covariance term is independent of the learnable predictor, we fix the variance and parameterize the mean as

$$\boldsymbol{\mu}_{\theta} = \frac{\eta_{t-1}}{\eta_t} \mathbf{x}_t + \frac{\gamma_t}{\eta_t} f_{\theta}(\mathbf{x}_t, \mathbf{x}_c, t). \quad (18)$$

This yields the simplified training objective

$$\theta^* = \arg \min_{\theta} \mathcal{L}_{\text{diff}} = \arg \min_{\theta} w_t \|f_{\theta}(\mathbf{x}_t, \mathbf{x}_c, t) - \mathbf{x}_0\|_2^2. \quad (19)$$

After T reverse steps, LPDM produces a generated BEV feature $\mathbf{x}_g$ that approximates the prior-enhanced representation. From a probabilistic perspective, the residual feature-state diffusion can be viewed as a stochastic restoration process in BEV feature space. Unlike standard DDPMs that generate data from an uninformative Gaussian prior, LPDM restores a topology-aware feature representation from the conditional BEV feature $\mathbf{x}_c$ through a short reverse chain. Different random initializations and noise samples during the reverse process lead to different restored features $\mathbf{x}_g$, which serve as plausible topology-aware feature hypotheses for the same input scene. These hypotheses are structural realizations in the continuous BEV feature space, rather than explicit samples of final vectorized lane graphs. This interpretation provides the probabilistic basis for LHAM, which consolidates multiple LPDM-restored feature hypotheses into a stable representation for final graph decoding.

Algorithm 2: Sampling of LPDM

Input: Condition feature $\mathbf{x}_c$, diffusion step number T

Output: Generated feature $\mathbf{x}_g$

```

1  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_c, \kappa^2 \eta_T \mathbf{I})$ ;
2 for  $t := T, T-1, \dots, 1$  do
3   Sample  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  if  $t > 1$ , else  $\epsilon := \mathbf{0}$ ;
4    $\boldsymbol{\mu} := \frac{\eta_{t-1}}{\eta_t} \mathbf{x}_t + \frac{\gamma_t}{\eta_t} f_{\theta}(\mathbf{x}_t, \mathbf{x}_c, t)$  according to Eq. 18;
5    $\mathbf{x}_{t-1} := \boldsymbol{\mu} + \kappa \sqrt{\frac{\eta_{t-1} \gamma_t}{\eta_t}} \epsilon$ ;
6 end
7  $\mathbf{x}_g := \mathbf{x}_{t=0}$ ;
8 return  $\mathbf{x}_g$ ;
```

F. Lane Hypothesis Aggregation Module (LHAM)

Compared with the conference version [11], LHAM is newly introduced in this extended version to exploit the multiple structurally plausible BEV feature hypotheses produced by stochastic LPDM reverse diffusion. As discussed in Sec. III-E, these hypotheses correspond to different feature-space restorations of the same image-derived BEV representation, rather than explicit samples of final vectorized lane graphs. They capture the uncertainty of lane topology under ambiguous observations. However, plausible diversity does not automatically imply reliable prediction. Individual diffusion samples may still be affected by stochastic variation and may not be equally trustworthy across spatial locations. Therefore, robust topology inference should not depend on any single realization, but on the structural consensus across multiple generated hypotheses. To address this issue, we introduce the *Lane Hypothesis Aggregation Module* (LHAM), which consolidates multiple diffusion-generated hypotheses into a consensus feature representation.

a) *Diffusion Hypothesis Sampling*: Given the conditional feature $\mathbf{x}_c$, we sample K independent reverse diffusion trajectories with different noise initializations. Each trajectory produces a generated BEV feature hypothesis $\mathbf{x}_g^{(k)}$, where $k = 1, \dots, K$. These hypotheses form the set $\mathcal{X}_g = \{\mathbf{x}_g^{(k)}\}_{k=1}^K$, which approximates the conditional feature distribution learned by LPDM. The diversity among these samples reflects structural ambiguity in the scene, especially in regions with occlusion, weak lane evidence, or complex topology interactions.

b) *Consistency-aware Aggregation*: To derive a stable representation from these stochastic hypotheses, LHAM measures their cross-hypothesis agreement and uses it to guide feature fusion. Given K diffusion-generated features $\{\mathbf{x}_g^{(k)}\}_{k=1}^K$, the sample mean is computed element-wise over the hypothesis dimension:

$$\bar{\mathbf{x}}_g = \frac{1}{K} \sum_{k=1}^K \mathbf{x}_g^{(k)}. \quad (20)$$

The cross-hypothesis dispersion is then estimated element-wise as

$$\sigma^2 = \frac{1}{K} \sum_{k=1}^K \left(\mathbf{x}_g^{(k)} - \bar{\mathbf{x}}_g\right)^2. \quad (21)$$

TABLE II

SUMMARY OF THE THREE-STAGE OPTIMIZATION PROCEDURE. ONLY THE MODULES LISTED IN THE “OPTIMIZED MODULES” COLUMN ARE UPDATED IN EACH STAGE, WHILE THE REMAINING COMPONENTS ARE FIXED OR UNUSED. $\mathbf{x}_c$ DENOTES THE ORIGINAL BEV FEATURE, $\mathbf{x}_0$ DENOTES THE PRIOR-ENHANCED BEV FEATURE PRODUCED BY LPIM, AND $\tilde{\mathbf{x}}_g$ DENOTES THE AGGREGATED DIFFUSION-RESTORED FEATURE.

Stage	Optimized Modules	Input	Target	Loss	Output
I: Lane Prior Initialization	LPIM + auxiliary lane head	$\mathbf{x}_c$, lane priors	centerline GT	$\mathcal{L}_{\text{lane}}$	$\mathbf{x}_0$
II: Conditional Diffusion Learning	LPDM	$\mathbf{x}_c, \mathbf{x}_0, t$	$\mathbf{x}_0$	$\mathcal{L}_{\text{diff}}$	$p_{\theta}(\mathbf{x}_0 \mathbf{x}_c)$
III: Task-specific Adaptation	fusion layer + lane decoder	$\mathbf{x}_c, \tilde{\mathbf{x}}_g$	centerline GT	$\mathcal{L}_{\text{lane}}$	final lane graph prediction

Here, σ^2 has the same spatial and channel dimensions as the BEV feature, and each element measures the disagreement among diffusion hypotheses at a specific BEV location and feature channel. An inverse-dispersion confidence weight is derived as

$$\mathbf{w} = \frac{1}{\sigma^2 + \epsilon}, \quad (22)$$

where ϵ is a small constant for numerical stability. The resulting weight is normalized to keep its average scale stable and clipped to avoid excessively large responses in low-dispersion regions. The dispersion-aware generated feature is obtained by

$$\tilde{\mathbf{x}}_g = \mathbf{w} \odot \bar{\mathbf{x}}_g, \quad (23)$$

where $\odot$ denotes element-wise multiplication. This weighting emphasizes feature responses that are consistent across stochastic diffusion hypotheses and suppresses responses with high cross-hypothesis disagreement. Finally, the aggregated feature $\tilde{\mathbf{x}}_g$ is fused with the conditional BEV feature $\mathbf{x}_c$ through a lightweight feature fusion layer, producing the refined representation used by the lane decoder. In this way, LHAM transforms stochastic diffusion diversity into robust structured inference. Rather than treating multi-sample diffusion as a test-time enhancement, it serves as a core mechanism for consolidating multiple plausible topology hypotheses into a stable topology-aware feature representation.

G. Optimization

The training of LaneDiffusion++ follows a progressive three-stage optimization strategy that decouples structural prior construction, generative modeling, and task-specific adaptation. This staged design stabilizes diffusion learning and prevents interference between deterministic lane supervision and stochastic generative training. Tab. II summarizes the optimization interface of the three stages, including the optimized modules, inputs, supervision targets, losses, and outputs.

Stage-I: Lane Prior Initialization. In the first stage, we train LPIM to produce structurally grounded BEV feature representations. The optimization follows the lane learning objective used in the CGNet baseline, denoted as $\mathcal{L}_{\text{lane}}$:

$$\begin{aligned} \mathcal{L}_{\text{lane}} = & \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{poly}} + \lambda_3 \mathcal{L}_{\text{topo}} \\ & + \lambda_4 \mathcal{L}_{\text{dir}} + \lambda_5 \mathcal{L}_{\text{bezier}} + \lambda_6 \mathcal{L}_{\text{ja}}, \end{aligned} \quad (24)$$

where the terms correspond to lane classification, polyline regression, topology reasoning, direction consistency, Bézier curve regularization, and junction alignment, respectively. An auxiliary lane prediction head is used in this stage to provide

lane-level supervision for LPIM, and is not used in the subsequent diffusion learning stage. This stage establishes a stable prior-enhanced BEV feature target for subsequent diffusion learning.

Stage-II: Conditional Diffusion Learning. After LPIM converges, its parameters are frozen and LPDM is trained to model the conditional distribution of prior-enhanced BEV features. Specifically, LPDM learns the generative process $p_{\theta}(\mathbf{x}_0 | \mathbf{x}_c)$, where $\mathbf{x}_0$ is the prior-enhanced feature produced by LPIM and $\mathbf{x}_c$ is the original BEV feature extracted by the backbone. The diffusion model is optimized using $\mathcal{L}_{\text{diff}}$ in Eq. 19. Freezing LPIM in this stage prevents target drift and allows LPDM to focus on learning a stable conditional generative mapping. The detailed training procedure is summarized in Alg. 1.

Stage-III: Task-specific Adaptation. In the final stage, both LPIM and LPDM are fixed, and LPDM is switched to sampling mode as described in Alg. 2. For each input, multiple diffusion trajectories are generated and consolidated by LHAM to produce an aggregated feature $\tilde{\mathbf{x}}_g$. This feature is fused with the conditional BEV feature $\mathbf{x}_c$ and passed to the lane decoder. During this stage, only the decoder and fusion layer are optimized using $\mathcal{L}_{\text{lane}}$. This design allows the decoding head to adapt to the diffusion-enhanced representation without disturbing the already learned generative prior. As a result, the three stages progressively establish a stable structural target, learn a robust conditional generative model, and finally adapt the prediction head to topology-aware diffusion features.

IV. EXPERIMENTS

A. Dataset

We evaluate the proposed method on two large-scale autonomous driving benchmarks that provide vectorized lane topology annotations:

nuScenes. The nuScenes [9] dataset contains 1,000 driving scenes collected in complex urban environments, each lasting approximately 20 seconds. Key-frame annotations are provided at 2 Hz, and each sample includes synchronized multi-camera images covering a full 360° horizontal field of view around the ego vehicle. Following previous works, we adopt the official train/validation split for evaluation.

Argoverse2. Argoverse2 [10] contains 1,000 driving logs collected in diverse urban scenarios. Each log provides 15 seconds of multi-view RGB sequences captured at 20 Hz from 7 cameras, together with a high-definition vectorized map. We follow the official split to ensure fair comparison.

Notably, both datasets provide vectorized lane centerline annotations with explicit topological connectivity, which directly aligns with our structured centerline graph prediction task. These datasets also contain challenging scenarios involving occlusions, missing lane markings, and complex intersections, making them suitable for evaluating generative topology reasoning under structural ambiguity.

B. Evaluation Metrics

We adopt the evaluation protocol of CGNet [6] to comprehensively assess predicted lane graphs from both geometric and topological perspectives. Specifically, we report two complementary groups of metrics: *segment-level metrics*, which evaluate lane segment reconstruction and inter-segment connectivity at a coarse structural level, and *point-level metrics*, which measure fine-grained geometric precision and local topology consistency.

1) **Segment-level Metrics:** Segment-level metrics operate on the graph $G = (V, E)$, where vertices V represent lane centerline segments and edges $E \subseteq V \times V$ encode connectivity between segments. This group evaluates whether lane segments are correctly reconstructed and whether their inter-segment connectivity is preserved, providing a coarse but planning-relevant view of lane graph quality.

a) **Intersection-over-Union (IoU)** [61] measures the spatial overlap between predicted and ground-truth centerline segments. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ denote the rasterized representations of the predicted and ground-truth segments. The score is defined as

$$\text{IoU}(\mathcal{B}_1, \mathcal{B}_2) = \frac{\mathcal{B}_1 \cap \mathcal{B}_2}{\mathcal{B}_1 \cup \mathcal{B}_2}. \quad (25)$$

b) **mAP_{cf}** [3] evaluates segment reconstruction quality based on Chamfer distance. Mean average precision is computed over multiple matching thresholds $T = \{0.5, 1.0, 1.5\}$:

$$\text{mAP}_{cf} = \frac{1}{|T|} \sum_{t \in T} \text{AP}_t. \quad (26)$$

c) **DET_l** [2] evaluates segment detection accuracy using discrete Fréchet distance. The score is averaged over thresholds $T = \{1.0, 2.0, 3.0\}$:

$$\text{DET}_l = \frac{1}{|T|} \sum_{t \in T} \text{AP}_t. \quad (27)$$

d) **TOP_{ll}** [2] measures the correctness of predicted connectivity relationships between lane segments. Let $\hat{V}'$ denote the matched vertices between the predicted graph $\hat{G}$ and the ground-truth graph G . The topology score is defined as

$$\text{TOP}_{ll} = \frac{1}{|V|} \sum_{v \in V} \frac{\sum_{n' \in N'(v)} P(\hat{n}') \mathbb{1}(\hat{n}' \in N(v))}{|N(v)|}, \quad (28)$$

where $N(v)$ denotes the ordered neighbor list of v , and $P(\hat{n}')$ is the confidence score of edge $\hat{n}'$.

2) **Point-level Metrics:** Point-level metrics operate on the fine-grained graph $\tilde{G} = (\tilde{V}, \tilde{E})$, where $\tilde{V}$ contains interpolated points sampled from centerline polylines and $\tilde{E}$ represents their connectivity. This group provides a finer-grained evaluation of centerline accuracy and topology consistency, covering local geometry, local connectivity, junction regions, path-level consistency, and junction detection. Compared with segment-level metrics, these metrics are more sensitive to subtle geometric errors and local structural inconsistencies around lane continuity and junction regions.

a) **GEO F1** [14] measures geometric accuracy of predicted centerlines. All polylines are interpolated such that adjacent vertices are spaced at 0.25 m, and predicted and ground-truth points are considered matched if their distance is below 0.5 m. Precision, recall, and F1 are computed as

$$\text{Precision}_{\text{GEO}} = \frac{|\tilde{V}_m|}{|\hat{\tilde{V}}|}, \quad \text{Recall}_{\text{GEO}} = \frac{|\tilde{V}_m|}{|\tilde{V}|}, \quad (29)$$

$$\text{F1}_{\text{GEO}} = \frac{2 \text{Precision}_{\text{GEO}} \text{Recall}_{\text{GEO}}}{\text{Precision}_{\text{GEO}} + \text{Recall}_{\text{GEO}}}. \quad (30)$$

b) **TOPO F1** [14] extends GEO F1 by additionally evaluating local connectivity consistency. For each matched vertex pair $(v, \hat{v})$, local subgraphs are constructed within an 8 m radius, and the GEO metric is computed on the resulting subgraphs.

c) **JTOPO F1** [23] focuses on topology quality around junction points, defined as vertices with in-degree or out-degree greater than one. It is particularly sensitive to complex intersection and merging structures.

d) **Average Path Length Similarity (APLS)** [62] evaluates shortest-path consistency between predicted and ground-truth graphs:

$$\text{APLS} = 1 - \frac{1}{N} \sum \min \left\{ 1, \frac{|d(a, b) - d(a', b')|}{d(a', b')} \right\}, \quad (31)$$

where $d(a, b)$ and $d(a', b')$ denote shortest-path distances in the ground-truth and predicted graphs, respectively.

e) **SDA** [63] evaluates junction detection accuracy using Hungarian matching within a circular neighborhood around each ground-truth junction. A predicted junction is counted as correct if the matching distance is below 1.0 m, and an F1 score is reported accordingly.

Together, the segment-level and point-level groups provide complementary evaluation focuses: the former emphasizes coarse lane segment reconstruction and inter-segment connectivity, while the latter emphasizes fine-grained geometry, local topology, junction quality, and path consistency.

C. Implementation Details

All models are implemented in PyTorch and trained on 8 NVIDIA Tesla A800 GPUs. Unless otherwise specified, the batch size is set to 2 per GPU. The BEV perception range is $[-15.0 \text{ m}, 15.0 \text{ m}]$ along the X-axis and $[-30.0 \text{ m}, 30.0 \text{ m}]$ along the Y-axis.

Baseline Architecture. Following prior work, we reproduce the state-of-the-art CGNet [6] as the primary baseline. The

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON nuSCENES USING POINT-LEVEL AND SEGMENT-LEVEL METRICS.

Method	Epoch	Point-level Metrics					Segment-level Metrics			
		GEO F1 $\uparrow$	TOPO F1 $\uparrow$	JTOPO F1 $\uparrow$	APLS $\uparrow$	SDA $\uparrow$	IoU $\uparrow$	mAP _{cf} $\uparrow$	DET _t $\uparrow$	TOP _{ll} $\uparrow$
STSU [15]	200	33.0	20.6	13.9	11.0	7.0	–	–	–	–
HDMaNet [61]	30	45.5	20.0	14.8	25.9	0.5	–	–	–	–
VectorMapNet [3]	24	48.4	38.1	27.9	10.3	7.2	29.8	25.5	16.1	0.5
TopoNet [22]	24	50.8	39.6	31.9	26.2	5.6	51.3	31.6	17.9	0.8
MapTR [4]	24	53.3	39.7	32.4	26.8	7.6	53.7	33.1	18.9	1.0
CGNet [6]	24	54.7	42.2	34.1	30.7	8.8	56.3	35.2	22.0	1.3
LaneDiffusion [11]	24	58.9	46.8	38.8	37.1	10.6	58.6	41.6	28.8	3.4
LaneDiffusion++	24	59.3	47.4	39.6	37.9	11.0	59.0	42.3	29.6	3.8

image backbone adopts ResNet-50 [64] with a Feature Pyramid Network (FPN) [65] for multi-scale feature extraction. A geometry-aware transformer module (GKT) [66] projects image features into BEV space, producing feature maps of size $\mathbb{R}^{256 \times 200 \times 100}$. The centerline decoder employs deformable attention [67] for flexible interactions between object queries and BEV features. For fair comparison, all CGNet-based reproduced models and ablation variants, including CGNet, LaneDiffusion, and LaneDiffusion++, share the same image backbone, FPN, GKT-based BEV constructor, BEV feature resolution, and deformable-attention centerline decoder unless otherwise specified. They also use the same input image resolution, BEV perception range, data preprocessing and augmentation protocol. The differences among these variants are limited to the proposed LPIM, LPDM, LHAM, and their corresponding training stages.

Training Strategy. Training follows the three-stage schedule summarized in Tab. II and described in Sec. III-G. LPIM is first trained for 24 epochs using AdamW [68] with an initial learning rate of 6×10^{-4} and cosine annealing. LPDM is then trained from scratch for 24 epochs using AdamW with a reduced learning rate of 6×10^{-5} . Finally, the centerline decoder is trained for 24 epochs with AdamW (initial learning rate 6×10^{-4}) while keeping LPDM fixed. The weighting factors in Eq. 24 are empirically set to $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\} = \{2, 5, 1, 0.005, 0.01, 0.1\}$.

Lane Prior Injection Module. For LPIM, each scene contains a variable number of lane priors M , and each polyline prior is represented by $N = 20$ sampled points. The BEV feature dimension is $H = 256$, the positional embedding dimension is $d = 32$, and the prior Transformer encoder uses $L = 6$ layers.

Diffusion Configuration. LPDM operates in feature space rather than directly in the vectorized output space. The noise variance hyper-parameter is set to $\kappa = 2.0$, and the growth rate controlling $\sqrt{\eta_t}$ is set to $p = 0.3$. Unless otherwise specified, the number of diffusion steps is set to $T = 15$, which provides a favorable balance between generation quality and computational cost. Together, $\{T, \kappa, p\}$ define the main LPDM sampling configuration, corresponding to the diffusion step number, noise magnitude, and interpolation schedule, respectively.

Inference and Sample Aggregation. During inference,

we use $K = 5$ diffusion trajectories by default to obtain multiple plausible feature hypotheses for each scene. These hypotheses are consolidated by the proposed LHAM rather than naive feature averaging. Specifically, LHAM computes cross-hypothesis statistics, applies a normalized and clipped inverse-dispersion confidence weight with stability constant $\epsilon = 1 \times 10^{-5}$, and fuses the weighted feature with the conditional BEV feature using the Concat+ED setting. LHAM introduces no additional denoising steps beyond the sampled LPDM trajectories. Therefore, the dominant inference overhead comes from LPDM sampling, while the aggregation operation itself remains lightweight.

D. Quantitative Comparison with State-of-the-Art

a) Results on nuScenes: Table III reports quantitative comparisons on the nuScenes validation set using both point-level and segment-level metrics. The proposed method achieves the best performance across all reported metrics, outperforming both deterministic and generative baselines. Compared with previous deterministic methods, the gains are particularly pronounced on topology-sensitive metrics such as TOPO F1, JTOPO F1, APLS, SDA, and TOP_{ll}. These metrics evaluate complementary aspects of lane graph quality, including local connectivity consistency, topology around junction regions, shortest-path consistency, junction detection accuracy, and segment-level connectivity correctness. The consistent improvements across these metrics are aligned with our motivation: feature-level generative modeling is especially beneficial for reducing structural ambiguity and preserving coherent lane connectivity, rather than only improving local geometric overlap. For example, compared with the strongest deterministic baseline CGNet, LaneDiffusion improves TOPO F1 from 42.2 to 46.8 and APLS from 30.7 to 37.1, indicating better topology recovery under challenging conditions such as occlusion and incomplete lane markings. LaneDiffusion++ further improves over the conference version by introducing the proposed LHAM. By aggregating multiple stochastic feature hypotheses before decoding, LHAM produces a more stable structural representation and yields consistent improvements across both point-level and segment-level metrics. These results verify that uncertainty-aware hypothesis aggregation can effectively convert diffusion diversity into performance gains for structured lane graph prediction.

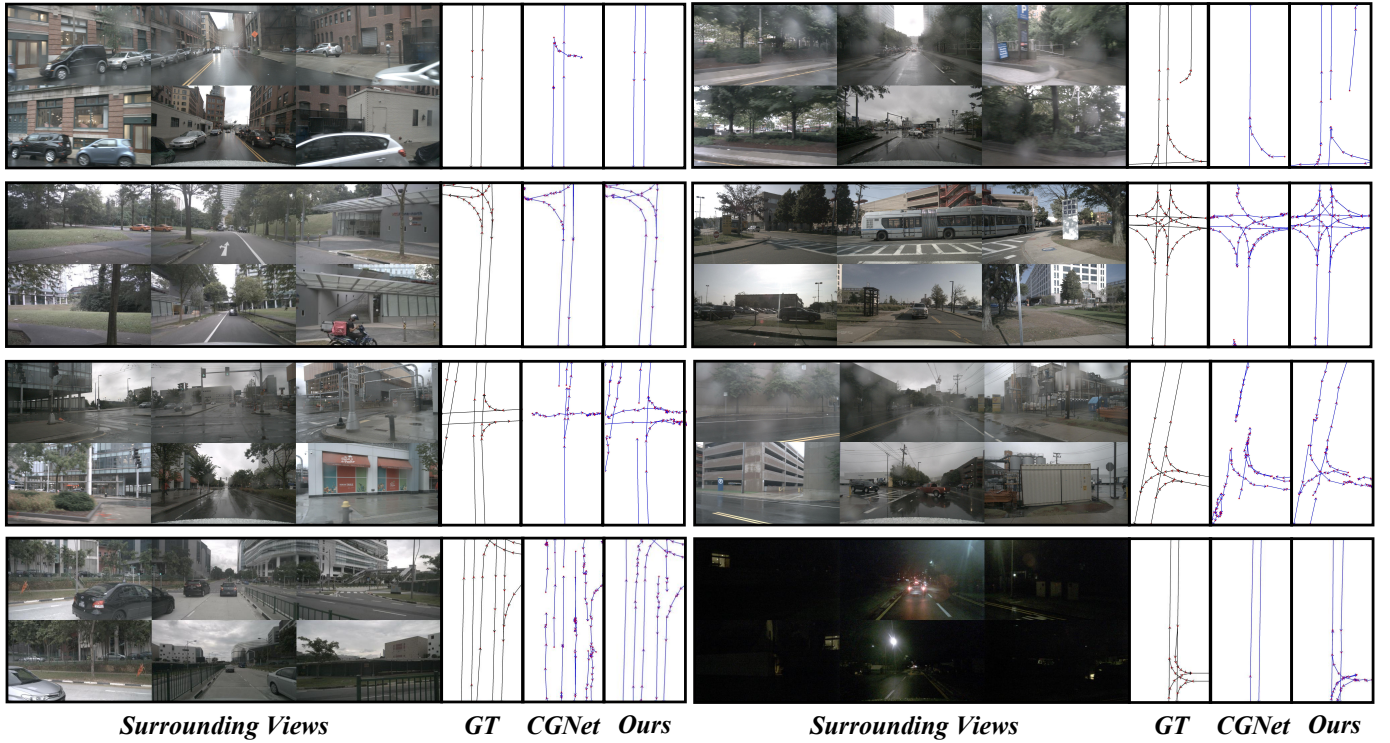


Fig. 3. Qualitative comparisons under challenging scenarios, including complex intersections, partially occluded lanes and low-visibility conditions.

TABLE IV

COMPARISON WITH STATE-OF-THE-ART METHODS ON ARGOVERSE2 VAL. ALL METHODS ARE TRAINED FOR 6 EPOCHS.

Method	TOPO F1 $\uparrow$	JTOPO F1 $\uparrow$	APLS $\uparrow$	SDA $\uparrow$	TOP _{ll} $\uparrow$
TopoNet [22]	30.2	23.7	15.3	7.7	0.3
MapTR [4]	42.8	33.5	22.3	13.6	0.5
CGNet [6]	44.5	34.6	23.6	13.7	0.5
LaneDiffusion [11]	47.0	36.4	25.8	14.4	1.8
LaneDiffusion++	47.5	37.0	26.5	14.7	2.0

b) Results on Argoverse2: Table IV shows the results on the Argoverse2 dataset. LaneDiffusion++ again achieves the best performance on all reported metrics and consistently outperforms both deterministic baselines and the conference version. The improvements are observed across both point-level and segment-level topology-oriented metrics, suggesting that the structural benefits are not limited to a single evaluation criterion. The consistent improvements across both datasets confirm the robustness and cross-dataset generalization ability of the proposed generative topology learning framework. Notably, the improvement margins are larger for topology-aware metrics than purely geometric ones, which further supports the motivation of introducing generative feature modeling for structured reasoning.

E. Qualitative Analysis

Fig. 3 presents qualitative comparisons under challenging scenarios, including complex intersections, partially occluded lanes, and low-visibility conditions. Deterministic baselines often produce fragmented lane segments, missing connections,

TABLE V

ABLATION STUDY ON THE DIFFUSION STEP NUMBER T AND THE LPDM TRAINING PARADIGM ON nuScenes. “SAMPLING” DENOTES TRAINING WITH RANDOMLY SAMPLED DIFFUSION TIMESTEPS, WHILE “FULL-CHAIN” SUPERVISES ALL DENOISING STEPS ALONG THE FULL DIFFUSION CHAIN. INFERENCE TIME IS THE MEASURED END-TO-END RUNTIME ON AN NVIDIA A800 GPU.

T	Paradigm	Time	TOPO F1 $\uparrow$	JTOPO F1 $\uparrow$	APLS $\uparrow$	SDA $\uparrow$
5	Full-chain	~3s	45.6	37.5	35.5	9.5
	Sampling	~3s	44.8	36.4	33.1	9.2
15	Sampling	~12s	46.8	38.8	37.1	10.6
30	Sampling	~28s	46.0	37.6	36.3	10.2

or incomplete junction structures when visual evidence is weak. In contrast, the proposed method recovers more coherent lane topology by reasoning over plausible structural priors in BEV feature space. Even in ambiguous regions with partially missing lane markings, the generated features provide stronger structural cues for consistent lane continuation and connectivity prediction.

F. Ablation Analysis

In this section, we conduct ablation studies on the nuScenes benchmark to investigate the impact of key design choices and hyper-parameters. Unless otherwise stated, all models are trained for 24 epochs under the same setting to ensure fair comparison.

a) Effect of Diffusion Step Number: The diffusion step number T controls the granularity of the progressive restoration process in LPDM. As shown in Tab. V, increasing T from

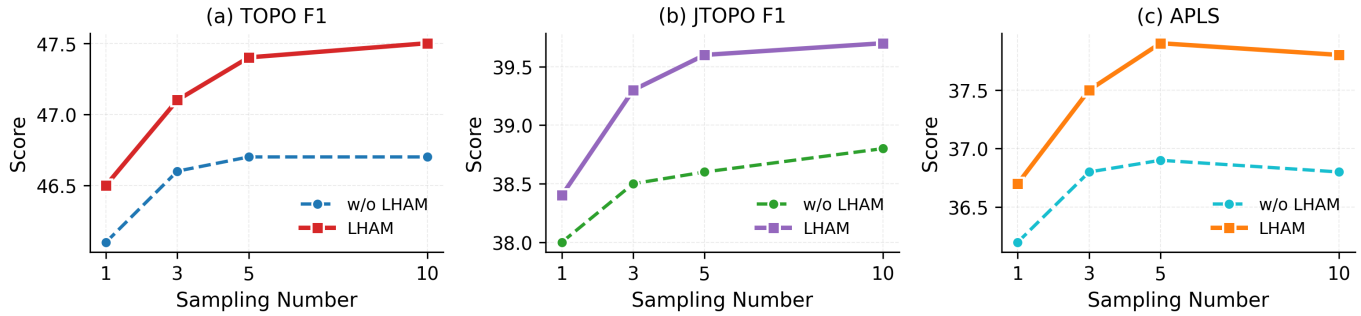


Fig. 4. Effect of diffusion sampling number during inference on nuScenes. Increasing the number of samples improves performance until saturation, while LHAM consistently stabilizes the predictions across different sampling settings.

TABLE VI

ABLATION STUDY ON THE FEATURE REFINEMENT MECHANISM IN LHAM. “CONCAT” AND “ADD” DENOTE FEATURE FUSION VIA CONCATENATION AND ELEMENT-WISE ADDITION, RESPECTIVELY. “FC” REPRESENTS A CONVOLUTION-BASED REFINEMENT BLOCK, WHILE “ED” DENOTES AN ENCODER-DECODER REFINEMENT MODULE.

Method	TOPO F1 ↑	JTOPO F1 ↑	APLS ↑	SDA ↑
No Refinement	43.9	35.9	33.3	9.2
Concat + FC	45.6	37.0	35.7	9.8
Concat + ED	46.8	38.8	37.1	10.6
Add + FC	45.3	37.2	35.7	10.2
Add + ED	46.8	38.5	36.9	10.6

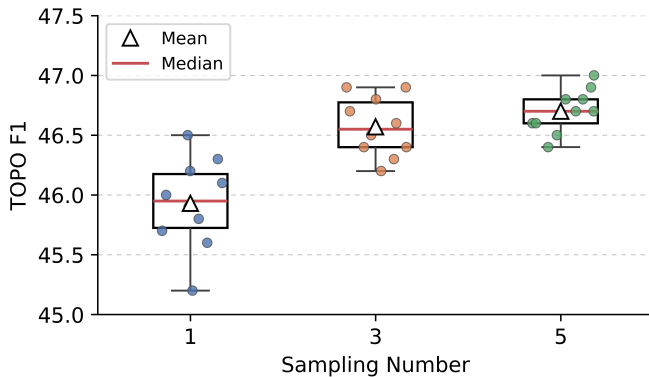


Fig. 5. Prediction variance analysis under different diffusion sampling numbers. Higher sampling numbers slightly improve the mean performance and significantly reduce the variance of TOPO F1.

5 to 15 consistently improves performance across topology-aware metrics, indicating that a moderate restoration chain helps refine structural priors and capture lane connectivity. We additionally report the end-to-end inference time on an NVIDIA A800 GPU to characterize the accuracy-efficiency trade-off. Further increasing T to 30 brings higher computational cost but no additional performance gain, suggesting that excessively long diffusion chains may introduce redundant refinement and accumulated noise. Therefore, we set $T = 15$ as the default configuration, which offers a favorable balance between prediction accuracy and inference efficiency.

b) Training Paradigm for LPDM: We compare two training paradigms for LPDM: the *Sampling* paradigm and the *Full-chain* paradigm. The Sampling paradigm randomly samples diffusion timesteps during training, while the Full-chain paradigm supervises all denoising steps within a full diffusion chain. As shown in Tab. V, the Full-chain paradigm slightly outperforms Sampling when the diffusion chain is short ($T = 5$). However, its memory consumption increases linearly with T , making it impractical for longer chains. Therefore, we adopt the Sampling paradigm for larger diffusion step numbers, which supports the default LPDM training strategy by providing a better trade-off between performance and computational cost.

c) Feature Refinement in LHAM: Tab. VI compares different feature fusion strategies in the LHAM refinement stage. Without refinement, directly feeding the aggregated feature to the decoder yields clearly inferior performance, indicating that the diffusion-restored feature cannot be used as a standalone replacement for the conditional BEV feature. This suggests that LHAM needs to align the generated topology-aware cues with the image-derived BEV evidence before decoding. Simple fusion strategies such as concatenation or element-wise addition followed by a convolution block already provide noticeable improvements, showing that explicit feature fusion is necessary for exploiting LPDM hypotheses. Replacing the convolution block with an encoder-decoder refinement module further improves the results and achieves the best overall performance. This indicates that multi-scale contextual refinement is beneficial for reconciling diffusion-generated structural priors with deterministic BEV features, especially for topology-aware prediction. Overall, the ablation shows that the effectiveness of LHAM comes not only from using multiple generated hypotheses, but also from how these hypotheses are aligned and fused with the original BEV representation.

d) Effect of Sampling Number: Fig. 4 studies the effect of the number of diffusion samples during inference. Increasing the sampling number improves performance up to a certain point, since additional samples provide richer topology hypotheses for aggregation. However, the marginal gain gradually saturates as the number of samples becomes large. In practice, using 5 samples already yields a stable representation. Moreover, when LHAM is enabled, the perfor-

mance improvement becomes more consistent across different sampling settings, indicating that uncertainty-aware aggregation helps the model make better use of multiple stochastic hypotheses.

e) Diffusion Behavior Analysis: The stochastic sampling process in diffusion models naturally produces multiple plausible feature hypotheses for the same scene. To further analyze this behavior, we measure prediction variance under different sampling numbers. As shown in Fig. 5, increasing the sampling number slightly improves the mean TOPO F1, while significantly reduces its variance across runs. This observation suggests that diffusion inference naturally captures multiple plausible structural hypotheses for ambiguous scenes. The proposed LHAM module further exploits this property by consolidating these hypotheses into a more stable consensus representation, thereby improving prediction consistency beyond single-sample inference.

G. Practical Scope and Limitations

Although LaneDiffusion++ improves topology-aware lane graph prediction, several practical limitations remain. First, the proposed LPDM sampling introduces additional inference latency compared with deterministic regression, although the residual feature-state diffusion adopts a relatively short sampling chain. Future work may explore faster sampling strategies or lightweight feature restoration networks to further improve efficiency. Second, LaneDiffusion++ operates on BEV features and therefore depends on the quality and resolution of the underlying BEV representation. Stronger BEV encoders or temporal aggregation may improve robustness under weak visual evidence, occlusion, or long-range perception errors. Third, LPIM relies on accurate lane graph annotations during training to construct prior-enhanced feature targets, making annotation quality important for learning reliable structural priors. Finally, this work focuses on lane centerline topology, while extending feature-level generative topology learning to broader structured map elements, such as road boundaries, crosswalks, and other traffic-related primitives, remains an important future direction.

V. CONCLUSION

In this paper, we presented LaneDiffusion++, a feature-level generative framework for structured centerline graph learning. Instead of directly generating vectorized lane graphs in a discrete output space, the proposed method reformulates lane topology reasoning as the conditional restoration of prior-enhanced BEV features. This formulation provides a more stable and effective way to model structural uncertainty, while preserving both geometric fidelity and topological consistency in the final decoded graphs. A key insight of this work is that structured topology ambiguity is better handled in an intermediate feature space than directly in graph space. Moreover, while diffusion inference naturally produces multiple plausible structural hypotheses, reliable topology prediction requires these hypotheses to be explicitly consolidated rather than used independently. By introducing uncertainty-aware hypothesis aggregation, LaneDiffusion++ converts diffusion diversity into

a stable topology-aware representation, which is particularly beneficial under occlusion, weak lane evidence, and complex intersections. Extensive experiments on nuScenes and Argoverse2 demonstrate that the proposed framework consistently outperforms strong deterministic and generative baselines, with especially clear advantages on topology-sensitive metrics. These results suggest that feature-level generative modeling, together with uncertainty-aware hypothesis aggregation, provides a promising paradigm for structured topology reasoning in autonomous perception. In future work, we will explore more efficient generative designs, tighter integration between uncertainty modeling and structured decoding, and broader extensions of feature-level generative reasoning to richer scene understanding tasks.

ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China (2024YFB3908503, 2024YFB3908500), in part by the National Natural Science Foundation of China (NO. 62322608 and NO. 62495081), and in part by the Shenzhen Loop Area Institute under grant FPF10120260001. The authors would also like to thank National Supercomputer Center in Guangzhou for providing high performance computational resources.

REFERENCES

- [1] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8340–8350.
- [2] H. Wang, T. Li, Y. Li, L. Chen, C. Sima, Z. Liu, B. Wang, P. Jia, Y. Wang, S. Jiang *et al.*, "Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [3] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Vectormapnet: End-to-end vectorized hd map learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 352–22 369.
- [4] 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.
- [5] 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.
- [6] Y. Han, K. Yu, and Z. Li, "Continuity preserving online centerline graph learning," in *European Conference on Computer Vision*. Springer, 2024, pp. 342–359.
- [7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [8] J. Chen, R. Deng, and Y. Furukawa, "Polydiffuse: Polygonal shape reconstruction via guided set diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [9] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: 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.
- [10] 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," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- [11] Z. Wang, W. Zhang, X. Tan, H. Liu, Y. Wang, and G. Li, "Lanediffusion: Improving centerline graph learning via prior injected bev feature generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 27 052–27 062.

- [12] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, "Roadtracer: Automatic extraction of road networks from aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4720–4728.
- [13] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1715–1724.
- [14] S. He and H. Balakrishnan, "Lane-level street map extraction from aerial imagery," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2080–2089.
- [15] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, "Structured bird's-eye-view traffic scene understanding from onboard images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 661–15 670.
- [16] Y. He, M. Jiang, X. Ye, L. Du, Z. Zou, W. Zhang, X. Tan, and E. Ding, "Repainting and imitating learning for lane detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5508–5516.
- [17] Y. B. Can, A. Liniger, D. P. Paudel, and L. Van Gool, "Improving online lane graph extraction by object-lane clustering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8591–8601.
- [18] D. Xia, W. Zhang, X. Liu, W. Zhang, C. Gong, J. Huang, M. Yang, and D. Yang, "Dumapnet: An end-to-end vectorization system for city-scale lane-level map generation," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6015–6024.
- [19] S. Huang, Z. Shen, Z. Huang, Y. Liao, J. Han, N. Wang, and S. Liu, "Anchor3dlane++: 3d lane detection via sample-adaptive sparse 3d anchor regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [20] D. Xia, W. Zhang, X. Liu, W. Zhang, C. Gong, X. Tan, J. Huang, M. Yang, and D. Yang, "Ldmapnet-u: An end-to-end system for city-scale lane-level map updating," *arXiv preprint arXiv:2501.02763*, 2025.
- [21] H. Li, S. Huang, L. Xu, Y. Gao, B. Mu, and S. Liu, "Ratopo: improving lane topology reasoning via redundancy assignment," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 777–786.
- [22] T. Li, L. Chen, H. Wang, Y. Li, J. Yang, X. Geng, S. Jiang, Y. Wang, H. Xu, C. Xu *et al.*, "Graph-based topology reasoning for driving scenes," *arXiv preprint arXiv:2304.05277*, 2023.
- [23] B. Liao, S. Chen, B. Jiang, T. Cheng, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction," in *European Conference on Computer Vision*. Springer, 2024, pp. 334–351.
- [24] J. Lu, R. Peng, X. Cai, H. Xu, H. Li, F. Wen, W. Zhang, and L. Zhang, "Translating images to road network: A non-autoregressive sequence-to-sequence approach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23–33.
- [25] R. Peng, X. Cai, H. Xu, J. Lu, F. Wen, W. Zhang, and L. Zhang, "Lanegraph2seq: Lane topology extraction with language model via vertex-edge encoding and connectivity enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4497–4505.
- [26] Y. Yang, Y. Luo, B. He, E. Li, Z. Cao, C. Zheng, S. Mei, and Z. Li, "Topo2seq: Enhanced topology reasoning via topology sequence learning," *arXiv preprint arXiv:2502.08974*, 2025.
- [27] 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.
- [28] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake lidar in 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9243–9252.
- [29] Y. Han, H. Zhang, H. Li, Y. Jin, C. Lang, and Y. Li, "Collaborative perception in autonomous driving: Methods, datasets, and challenges," *IEEE Intelligent Transportation Systems Magazine*, 2023.
- [30] Y. Li, Z. Li, N. Chen, M. Gong, Z. Lyu, Z. Wang, P. Jiang, and C. Feng, "Multiagent multitaversal multimodal self-driving: Open mars dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 041–22 051.
- [31] Z. Dan, Z. Wang, W. Zhang, X. Lin, W. Zhang, X. Tan, J. Wang, L. Lin, and G. Li, "Optimvmap: Offline vectorized map construction via optimal multi-vehicle perspectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 18 140–18 149.
- [32] H. Li, Z. Huang, Z. Wang, W. Rong, N. Wang, and S. Liu, "Enhancing 3d lane detection and topology reasoning with 2d lane priors," *arXiv preprint arXiv:2406.03105*, 2024.
- [33] X. Zhang, G. Liu, Z. Liu, N. Xu, Y. Liu, and J. Zhao, "Enhancing vectorized map perception with historical rasterized maps," in *European Conference on Computer Vision*. Springer, 2024, pp. 422–439.
- [34] W. Gao, J. Fu, Y. Shen, H. Jing, S. Chen, and N. Zheng, "Complementing onboard sensors with satellite maps: a new perspective for hd map construction," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 103–11 109.
- [35] T. Ort, J. M. Walls, S. A. Parkison, I. Gilitschenski, and D. Rus, "Maplite 2.0: Online hd map inference using a prior sd map," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8355–8362, 2022.
- [36] 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.
- [37] 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.
- [38] P. Jia, Z. Luo, T. Wen, M. Yang, K. Jiang, L. Cui, and D. Yang, "Enhancing lane segment perception and topology reasoning with crowdsourcing trajectory priors," *arXiv preprint arXiv:2411.17161*, 2024.
- [39] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.
- [40] W. Zhong, J. Lin, P. Chen, L. Lin, and G. Li, "High-fidelity and lip-synced talking face synthesis via landmark-based diffusion model," *arXiv preprint arXiv:2408.05416*, 2024.
- [41] J. Zhang, J. Wu, Y. Ren, X. Xia, H. Kuang, P. Xie, J. Li, X. Xiao, W. Huang, S. Wen *et al.*, "Unifl: Improve latent diffusion model via unified feedback learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 67 355–67 382, 2024.
- [42] C. He, Y. Shen, C. Fang, F. Xiao, L. Tang, Y. Zhang, W. Zuo, Z. Guo, and X. Li, "Diffusion models in low-level vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [43] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, W. Xiong, H. Zhang, L. Cao, and S. Chen, "Diffusion model-based image editing: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [44] J. Huang, P. Yan, J. Liu, J. Wu, Z. Wang, Y. Wang, L. Lin, and G. Li, "Dreamfuse: Adaptive image fusion with diffusion transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 17 292–17 301.
- [45] Z. Wang, W. Zhang, W. Zhang, X. Tan, W. Chen, X. Li, and G. Li, "Ardepth: Auto-regressive monocular depth estimation with progressive visual conditioning," *arXiv preprint arXiv:2607.12433*, 2026.
- [46] M. Fuest, P. Ma, M. Gui, J. Schusterbauer, V. T. Hu, and B. Ommer, "Diffusion models and representation learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.
- [47] Z. Wang, W. Zhang, W. Zhang, F. Zhang, X. Tan, Y. Qin, and G. Li, "World models as group actions," *arXiv preprint arXiv:2605.24578*, 2026.
- [48] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [49] T. Amit, T. Shaharabany, E. Nachmani, and L. Wolf, "Segdiff: Image segmentation with diffusion probabilistic models," *arXiv preprint arXiv:2112.00390*, 2021.
- [50] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19 830–19 843.
- [51] C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, D. Anguelov *et al.*, "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9644–9653.
- [52] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," *arXiv preprint arXiv:2310.02601*, 2023.
- [53] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 55–72.
- [54] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, "Drivedreamer-2: Llm-enhanced world models for diverse driving video generation," *arXiv preprint arXiv:2403.06845*, 2024.

- [55] G. Zhao, C. Ni, X. Wang, Z. Zhu, X. Zhang, Y. Wang, G. Huang, X. Chen, B. Wang, Y. Zhang *et al.*, “Drivedreamer4d: World models are effective data machines for 4d driving scene representation,” *arXiv preprint arXiv:2410.13571*, 2024.
- [56] P. Jia, T. Wen, Z. Luo, M. Yang, K. Jiang, Z. Liu, X. Tang, Z. Lei, L. Cui, B. Zhang *et al.*, “Diffmap: Enhancing map segmentation with map prior using diffusion model,” *IEEE Robotics and Automation Letters*, 2024.
- [57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [58] Z. Yue, J. Wang, and C. C. Loy, “Resshift: Efficient diffusion model for image super-resolution by residual shifting,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 13 294–13 307, 2023.
- [59] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [60] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.
- [61] Q. Li, Y. Wang, Y. Wang, and H. Zhao, “Hdmapnet: An online hd map construction and evaluation framework,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4628–4634.
- [62] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, “Spacenet: A remote sensing dataset and challenge series,” *arXiv preprint arXiv:1807.01232*, 2018.
- [63] M. Büchner, J. Zürn, I.-G. Todoran, A. Valada, and W. Burgard, “Learning and aggregating lane graphs for urban automated driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 415–13 424.
- [64] 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.
- [65] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [66] 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.
- [67] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” in *International Conference on Learning Representations*.
- [68] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.



Weiming Zhang received a M.S. degree from Beijing University of Posts and Telecommunications in 2018. She is currently a Senior Engineer in the Applied Model Technology Department of Baidu Inc. Her current research interests include object detection/segmentation, semi-supervised learning, and vision-language models (VLMs). She is committed to developing vision models and vision-language models for autonomous driving and search recommendation.



Xiao Tan received the Ph.D. degree in computer vision from the University of New South Wales, Sydney, in 2014. His research interests include computer vision, pattern recognition, and image processing. He is currently with the Department of Computer Vision, Baidu as a Senior Engineer. He served as a reviewer for ICCV, CVPR, ECCV, AAAI, IJCV, and etc.



Jie Feng received the B.S. degree from Chang'an University, Xi'an, China, in 2008, and the Ph.D. degree from Xidian University, Xi'an, in 2014. She is currently a Professor with the Laboratory of Intelligent Perception and Image Understanding, Xidian University. Her research interests include remote sensing image processing, deep learning, and machine learning.



Xiaonan Luo is currently a Professor with the School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China. His research interests include computer vision, image processing, computer graphics, and CAD.



Zijie Wang received the M.S. degree from Nanjing Tech University, Nanjing, China, in 2024. He is currently working toward a Ph.D. degree at Sun Yat-sen University, Guangzhou China and Shenzhen Loop Area Institute, Shenzhen, China. His research interests lie in deep learning and computer vision, including autonomous driving, world model, and cross-modal retrieval.



Wei Zhang received a Ph.D. degree from the University of Hong Kong in 2017. He is currently a senior Engineer in the Department of Computer Vision Technology at Baidu Inc. His current research interests include 2D/3D object detection, segmentation, and tracking. He is dedicated to the development of computer vision models for autonomous driving and intelligent transportation systems.



Guanbin Li (M'15) is currently a full professor in School of Computer Science and Engineering, Sun Yat-sen University. He received his PhD degree from the University of Hong Kong in 2016. His current research interests include computer vision, image processing, and deep learning. He is a recipient of the ICCV 2019 Best Paper Nomination Award and CVPR2024 best paper candidate. He has authorized and co-authored on more than 200 papers in top-tier academic journals and conferences. He serves as an area chair for the conference of CVPR and ICCV.