TOPOMIND: UNIFYING LANE TOPOLOGY REASONING AND UNDERSTANDING IN DRIVING SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Lane topology reasoning provides geometric and relational constraints essential for autonomous driving. Existing methods, however, treat it primarily as a perception problem, detecting lane elements and then inferring connectivity without understanding topology-relevant scene context. We present TOPOMIND, to our knowledge the first unified framework for lane topology perception and topology-oriented scene understanding. TOPOMIND integrates lane-topology perception and topology-oriented scene understanding within a shared BEV–language architecture, where a large language model jointly processes metric BEV and text tokens and Semantic-Geometric Query Refinement couples the resulting multimodal context with geometry-anchored lane queries for topology reasoning. To supervise this capability, we assemble a topology-oriented VQA corpus comprising approximately 1.19 million question–answer pairs, covering road layout, lane markings, traffic controls, nearby road users, and coarse ego-centric directional cues, providing auxiliary language supervision over topology-relevant scene semantics. Extensive experiments on OpenLane-V2 and OmniDrive-nuScenes demonstrate that TOPOMIND achieves state-of-the-art performance on both lane topology perception and topology-oriented understanding, while consistently surpassing perception-only topology models. These results demonstrate the effectiveness of jointly learning topology-oriented language supervision and lane topology perception within a unified representation.

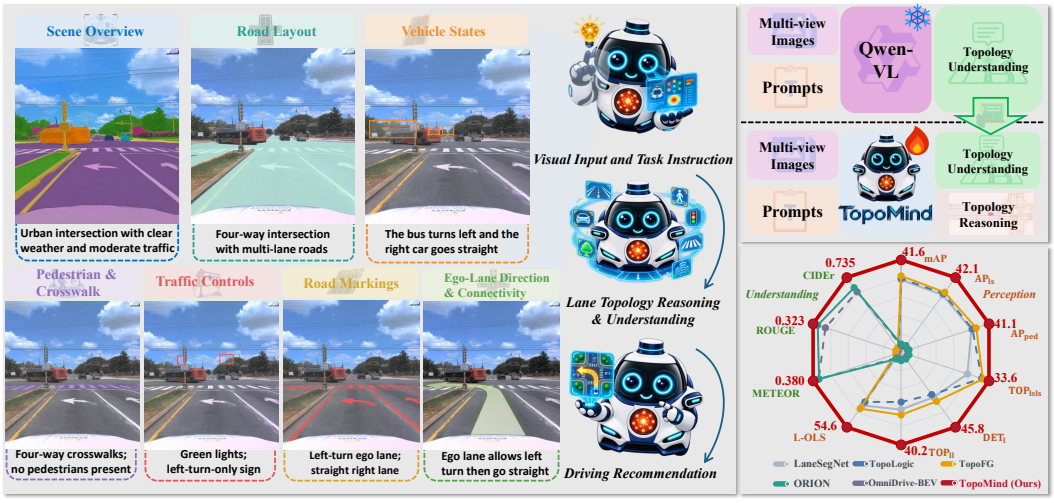


Figure 1: **Motivation and overview of TOPOMIND.** The left panel illustrates topology-oriented scene understanding covering road structure and topology-relevant scene semantics. The upper-right panel shows how VQA pairs generated by a frozen Qwen-VL teacher supervise TOPOMIND, which supports scene understanding and lane-topology perception within a unified framework. The lower-right comparison shows that TOPOMIND, the first lane-topology reasoning model with both capabilities, outperforms perception-only lane-topology methods and existing driving-scene understanding methods.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Safe driving requires more than detecting individual lane markings; it relies on a structured understanding of how lanes constrain feasible vehicle motion. The centerline geometry of the ego lane provides a local reference for maintaining the current direction of travel, while its connectivity to downstream lanes allows the driver to anticipate how the road continues through intersections and other complex layouts. These complementary requirements motivate lane topology reasoning, which jointly estimates lane geometry, boundary semantics, and directed lane-to-lane connectivity from onboard observations (Wang et al., 2023; Li et al., 2024b). By providing explicit structural constraints to downstream prediction and planning modules, lane topology reasoning narrows the space of feasible maneuvers and supports safer driving decisions.

Existing lane topology reasoning methods predominantly follow a *first-detect-then-reason* paradigm: they first detect geometric lane elements and subsequently infer pairwise topology from the resulting instance representations. Recent studies strengthen this paradigm through endpoint-aware alignment and distance-guided relation modeling (Fu et al., 2024; 2026b), as well as temporal BEV and query propagation (Yang et al., 2025c;a). In parallel, the autonomous driving community is increasingly exploring interpretable systems and flexible human–vehicle interaction. Vision–language models support this direction by drawing on pretrained world knowledge and language-based reasoning to describe complex driving scenes and answer context-dependent questions (Sima et al., 2024; Wang et al., 2025; Zhou et al., 2025). Lane topology reasoning, however, has largely remained confined to geometric perception: existing models are not designed to explain the surrounding road context or answer topology-aware questions, such as which lane markings are nearby or which direction the ego lane permits. Bridging this gap is non-trivial because visual geometry and linguistic semantics are represented in fundamentally different modalities. This naturally raises a central question: *how can scene understanding and lane-topology perception be seamlessly integrated into a unified model?*

Driven by these motivations, we propose TOPOMIND, a unified perception-and-understanding model for lane topology reasoning. Figure 1 illustrates the motivation and capabilities of TOPOMIND. We assemble a topology-oriented VQA corpus of approximately 1.19M question–answer pairs covering road layout, lane markings, traffic controls, nearby road users, and coarse ego-centric driving directions. The resulting language supervision captures scene semantics and local directional cues relevant to topology reasoning, rather than encoding the complete lane graph or pairwise lane-to-lane connectivity. A frozen Qwen-VL teacher generates additional VQA pairs from multi-view images and prompts to supervise the understanding branch of TOPOMIND. To bridge the representation gap between linguistic semantics and geometry-anchored lane queries, we further introduce semantic-geometric query refinement. This module refines reference-anchored lane queries using the shared multimodal context before topology decoding. Consequently, TOPOMIND supports language-based scene understanding and lane-topology prediction within a unified framework. Experiments demonstrate consistent improvements over perception-only lane-topology methods on OpenLane-V2 and existing driving-scene understanding methods on OmniDrive-nuScenes (Wang et al., 2025).

Our main contributions are summarized as follows:

- We propose TOPOMIND, a unified perception-and-understanding framework that couples LLM-based environmental understanding with geometry-aware lane-topology prediction. To the best of our knowledge, TOPOMIND is the first framework to jointly formulate language-based scene understanding and lane-topology perception within a single model.
- We assemble a topology-oriented VQA corpus of approximately 1.19M question–answer pairs that provides auxiliary language supervision over topology-relevant scene semantics and coarse ego-centric directional cues. We further develop Semantic-Geometric Query Refinement to couple this semantic context with spatial geometry and reference-anchored lane queries.
- Extensive experiments demonstrate the effectiveness of TOPOMIND, yielding consistent improvements in both lane-topology perception and scene understanding.

2 Related Work

2.1 Vision-Language Models for Driving

Large-scale vision–language pretraining equips VLMs with transferable knowledge for understanding dynamic driving scenes. Connecting this understanding to prediction and planning supports generalization and interpretable decisions (Sima et al., 2024; Shao et al., 2024; Xu et al., 2024; Wang et al., 2025). VQA provides an interface for supervising driving-scene understanding. DriveLM (Sima et al., 2024) organizes perception, prediction, and planning questions into Graph VQA for interactive driving. OmniDrive (Wang et al., 2025) uses counterfactual question–answer supervision to connect 3D understanding with future outcomes and planning. DriveGPT4 (Xu et al., 2024) instruction-tunes a video–language model to answer questions, explain behavior, and predict control signals. LMDrive (Shao et al., 2024) combines language instructions and multimodal sensor features for closed-loop driving. CoT reasoning makes intermediate driving decisions explicit. ReasonPlan (Liu et al., 2025) combines next-scene prediction with a supervised decision CoT. AutoVLA (Zhou et al., 2026) learns trajectory-only and CoT-enhanced modes, using reinforcement fine-tuning to avoid unnecessary reasoning. Chameleon (Zhang et al., 2025) handles routine lane-topology relations with synthesized programs and difficult cases with VLM reasoning. However, autoregressive intermediate text introduces latency that can impede real-time decisions, motivating latent CoT. LCDrive (Tan et al., 2026) interleaves action proposals with latent world-model tokens for faster inference. LaST-VLA (Luo et al., 2026a) aligns latent spatio-temporal reasoning with geometric constraints and world-model dynamics.

Unlike language-centric models that formulate driving outputs as text or use only front-view video, TopoMind jointly optimizes understanding and geometric topology prediction from surround-view BEV features. The metric BEV preserves cross-view geometry, while SGQR aligns language context with spatially anchored lane queries.

2.2 Lane Topology Reasoning

Online HD-map construction recovers road geometry from onboard observations, reducing dependence on costly offline map maintenance. VectorMapNet (Liu et al., 2023) predicts sparse BEV polylines, while MapTR and MapTRv2 (Liao et al., 2022; 2025) represent map elements as permutation-equivalent point sets. StreamMapNet (Yuan et al., 2024) and stream query denoising (Wang et al., 2024) exploit temporal information for map consistency. These methods primarily reconstruct individual map elements without explicitly modeling directed lane connectivity.

Lane topology reasoning extends mapping to structured road graphs. OpenLane-V2 (Wang et al., 2023) evaluates lane and traffic-element perception alongside their relations. TopoMLP (Wu et al., 2024) detects elements before predicting topology, whereas LaneSegNet (Li et al., 2024b) unifies centerlines, boundaries, semantics, and connectivity. Later methods improve graph interaction (Li et al., 2026b; Lv et al., 2025; Fu et al., 2024; 2026b), endpoint geometry and fine-grained representations (Xu et al., 2026; Luo et al., 2026b), connected-lane or sequence modeling (Yang et al., 2025b; Li et al., 2026a; Fu et al., 2026a), and temporal propagation (Yang et al., 2025c;a). These methods focus on geometric detection and relations; TopoMind additionally jointly learns topology-oriented scene understanding and lane perception through a shared BEV–language representation.

3 Method

3.1 Problem Formulation

Lane topology reasoning aims to recover a structured local road graph from onboard observations, analogous to detecting the lane geometry represented in an HD map (Wang et al., 2023; Li et al., 2024b). In this work, we focus on two lane representations: centerlines and lane segments. Let $L = \{L_i\}_{i=1}^{N_l}$ denote the ground-truth lane set and $\mathbf{A} \in \{0,1\}^{N_l \times N_l}$ its directed adjacency matrix. Given calibrated surround-view images $\mathcal{I} = \{I_i\}_{i=1}^{N_v}$ and topology-oriented text prompts $\mathcal{T} = \{T_i\}_{i=1}^{N_q}$, our model predicts a set of lane candidates $\hat{L} = \{\hat{L}_j\}_{j=1}^{N_{\text{qry}}}$, their connectivity probabilities

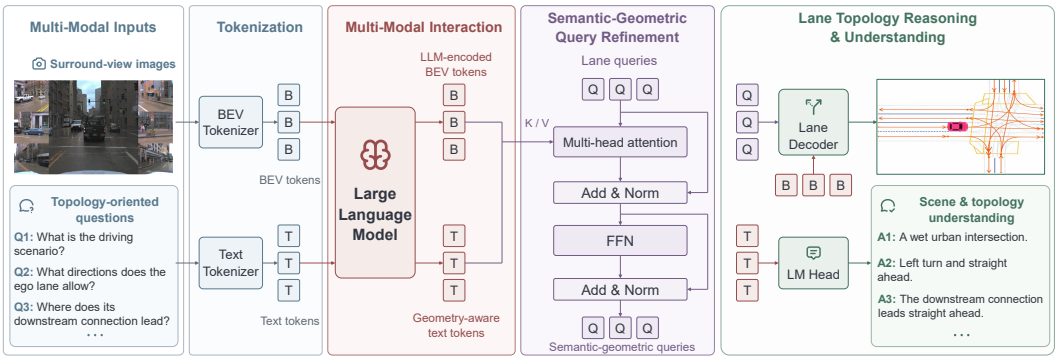


Figure 2: Overall architecture of TOPOMIND. Multi-view observations are encoded as metric BEV tokens and jointly processed with topology-aware VQA tokens by an LLM. The understanding branch generates scene descriptions and topology-aware answers, while the perception branch recovers BEV features from the shared LLM states. Semantic-geometric query refinement then injects textual context and BEV geometry into spatially anchored lane queries before lane-topology decoding.

$\hat{\mathbf{A}} \in [0,1]^{N_{\text{qry}} \times N_{\text{qry}}}$, and scene-understanding responses $\hat{\mathcal{U}} = \{\hat{U}_i\}_{i=1}^{N_q}$:

$$(\hat{L}, \hat{\mathbf{A}}, \hat{\mathcal{U}}) = \mathcal{F}_\theta(\mathcal{I}, \mathcal{T}), \tag{1}$$

where N_v, N_q, and N_{qry} denote the numbers of camera views, text prompts, and lane queries, respectively. Each lane L_i is represented by an ordered sequence of 3D points in ego-centric coordinates, $\mathbf{P}_i = [\mathbf{p}_{i,1}, \ldots, \mathbf{p}_{i,K}]$ with $\mathbf{p}_{i,k} = (x_{i,k}, y_{i,k}, z_{i,k})$, together with its semantic attributes. A centerline contains only the central lane curve, whereas a lane segment comprises its centerline and left and right boundaries, denoted by $L_i = (\mathbf{P}_i^l, \mathbf{P}_i^c, \mathbf{P}_i^r)$. The lane-segment category distinguishes elements such as drivable lanes and pedestrian crossings, while each boundary is labeled as solid, dashed, or invisible. In the ground-truth adjacency matrix, $A_{ij} = 1$ indicates that L_i directly connects to L_j along the direction of travel; correspondingly, $\hat{A}_{ij}$ is the predicted connection probability between $\hat{L}_i$ and $\hat{L}_j$.

Figure 2 summarizes the architecture: the LLM jointly processes BEV and text tokens, SGQR refines lane queries for geometric and topological decoding, and the language-modeling head generates understanding responses.

3.2 TOPOLOGY-ORIENTED SCENE UNDERSTANDING

Lane topology is not determined by geometry alone. Road layout, lane markings, traffic controls, nearby road users, and the direction of the ego lane all provide contextual cues for interpreting feasible lane connections. Inspired by prior driving-scene VQA benchmarks and holistic vision–language datasets (Qian et al., 2024; Wang et al., 2025), we use Qwen3-VL-32B-Instruct (Bai et al., 2025) as a frozen teacher to generate topology-oriented VQA supervision. Importantly, this supervision is designed to encode scene semantics and local ego-centric directional cues; it does not describe the complete lane graph or provide pairwise lane-to-lane adjacency labels. Full lane geometry and directed connectivity are supervised separately by the perception objectives. To reduce front–rear ambiguity, we summarize the synchronized camera observations into front-view and rear-view panoramic composites. VQA generation follows a coarse-to-fine procedure. The teacher first summarizes the global driving context and then constructs both template-driven and scene-adaptive question–answer pairs to provide language supervision over topology-relevant scene semantics, including road structure, traffic controls, surrounding road users, and coarse ego-centric directional cues. As a qualitative sanity check, we manually inspect a subset of the generated VQA pairs and correct obvious inaccuracies, inconsistencies, or insufficiently grounded responses when encountered. Detailed view construction, auxiliary annotation prompts, question templates, filtering instructions, and quality-control procedures are provided in Appendix F. The resulting pairs $\{(T_i, U_i)\}_{i=1}^{N_q}$ supervise the understanding branch of TOPOMIND. This hybrid design provides consistent coverage of key topology-relevant scene semantics through fixed templates while preserving scene-specific

diversity through adaptive questions. As shown in Figure 3, the resulting VQA supervision covers rich semantic content and diverse categories.

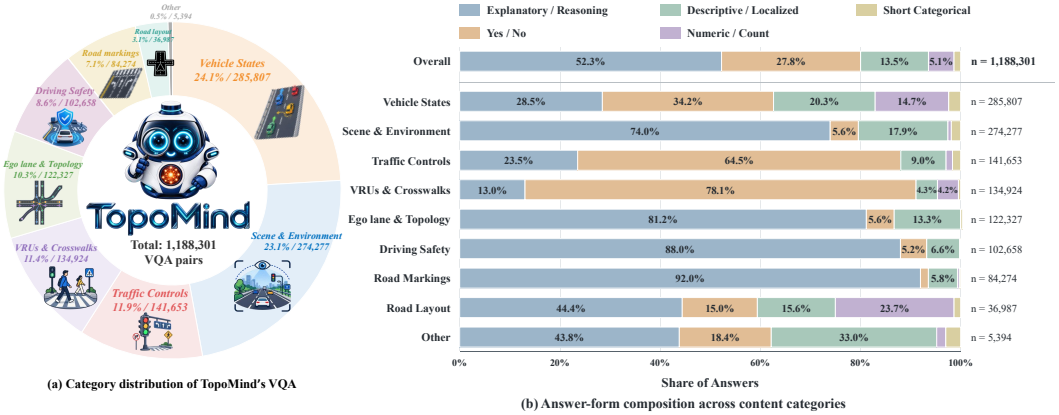


Figure 3: Statistics of the topology-oriented VQA supervision. (a) Distribution across scene-understanding categories. (b) Answer-form composition within each category.

3.3 BEV and Text Tokenizer

To preserve cross-view geometric relationships while retaining rich semantic cues for the LLM, we encode the multi-view observations with a BEV tokenizer. Given the surround-view images $\mathcal{I} = \{I_i\}_{i=1}^{N_v}$, a ResNet-50 backbone (He et al., 2016) and FPN (Lin et al., 2017) first extract multi-scale image features. BEVFormer (Li et al., 2024c) then aggregates these features in a shared bird's-eye-view coordinate system:

$$\mathbf{F}_i^{\text{mv}} = \text{FPN}(\text{ResNet}(I_i)), \qquad \mathbf{F}_{\text{bev}} = \text{BEVFormer}\Big(\{\mathbf{F}_i^{\text{mv}}\}_{i=1}^{N_v}\Big) \in \mathbb{R}^{H\times W\times C}, \tag{2}$$

where H and W denote the spatial resolution of the encoded scene and C is the BEV channel dimension. Thus, $\mathbf{F}_{\text{bev}}$ jointly preserves metric geometry across camera views and the semantic information required for scene understanding.

Directly feeding all BEV locations to the LLM would incur excessive memory and attention costs. We therefore apply $r \times r$ average pooling to $\mathbf{F}_{\text{bev}} \in \mathbb{R}^{H\times W\times C}$ and use a channel projection module to map its channels from C to the LLM hidden dimension d_{LLM}. This changes the feature dimensions to $\frac{H}{r} \times \frac{W}{r} \times d_{\text{LLM}}$. Flattening the spatial grid then produces the BEV token sequence $\mathbf{X}^{\text{bev}} \in \mathbb{R}^{N_b\times d_{\text{LLM}}}$, where $N_b = HW/r^2$, enabling the BEV and text sequences to be processed in a shared embedding space.

For a topology-oriented text prompt $T_i \in \mathcal{T}$ associated with the current scene, the LLM tokenizer segments the raw text into N_i^{text} subword units and maps them to the vocabulary-index sequence $\mathbf{z}_i^{\text{text}} = [z_{i,1}, \ldots, z_{i,N_i^{\text{text}}}]$, where each $z_{i,j}$ indexes an entry in the LLM vocabulary $\mathcal{V}$. The LLM token-embedding matrix $\mathbf{E}_{\text{tok}} \in \mathbb{R}^{|\mathcal{V}|\times d_{\text{LLM}}}$ then maps these discrete indices to the continuous text-embedding sequence $\mathbf{X}_i^{\text{text}} \in \mathbb{R}^{N_i^{\text{text}}\times d_{\text{LLM}}}$:

$$\begin{aligned} \mathbf{z}_i^{\text{text}} &= \text{Tokenizer}_{\text{LLM}}(T_i) = [z_{i,1}, \ldots, z_{i,N_i^{\text{text}}}], \qquad z_{i,j} \in \{1, \ldots, |\mathcal{V}|\}, \\ \mathbf{X}_i^{\text{text}} &= [\mathbf{E}_{\text{tok}}(z_{i,1}); \ldots; \mathbf{E}_{\text{tok}}(z_{i,N_i^{\text{text}}})] \in \mathbb{R}^{N_i^{\text{text}}\times d_{\text{LLM}}}. \end{aligned} \tag{3}$$

During autoregressive generation, the text sequence additionally includes the available answer prefix, supplied by teacher forcing during training or previously generated tokens at inference. In the joint-encoding equations below, $\mathbf{X}^{\text{text}}$ and $\mathbf{H}^{\text{text}}$ denote the text embeddings and states of the supplied dialogue context.

3.4 Bridging Perception and Understanding

To couple topology-oriented language representations with lane perception, we jointly encode BEV and text features and use them to refine spatially anchored lane queries.

LLM Embedding. The shared LLM processes concatenated BEV and text tokens. Following Qwen3-VL (Bai et al., 2025), M-RoPE preserves the two-dimensional BEV grid and sequential text positions during multimodal interaction:

$$[\mathbf{H}^{\text{bev}}; \mathbf{H}^{\text{text}}] = \text{LLM}\big([\mathbf{X}^{\text{bev}}; \mathbf{X}^{\text{text}}]\big)\,. \tag{4}$$

The latent states $\mathbf{H}^{\text{bev}} \in \mathbb{R}^{N_b \times d_{\text{LLM}}}$ and $\mathbf{H}^{\text{text}}$ encode spatial and linguistic context, respectively. We project the BEV states to C channels, restore the $\frac{H}{r} \times \frac{W}{r}$ grid, and apply bilinear upsampling followed by convolutional refinement. The recovered features $\widetilde{\mathbf{F}}_{\text{bev}} \in \mathbb{R}^{H \times W \times C}$ provide dense BEV representations for subsequent geometric perception.

Understanding. The understanding branch generates scene descriptions and answers to topology-oriented questions. The LM head predicts each answer autoregressively:

$$p(\hat{U}_i \mid \mathcal{I}, T_i) = \prod_{m=1}^{N_i} p\big(\hat{u}_{i,m} \mid \hat{u}_{i,<m}, \mathbf{X}^{\text{bev}}, \mathbf{X}_i^{\text{text}}\big)\,, \tag{5}$$

where N_i is the answer length. At each decoding step, the LM head maps the final text state, conditioned on the question and answer prefix, to the next-token distribution. During training, teacher forcing is used and only assistant-answer tokens are supervised:

$$\mathcal{L}_{\text{vqa}} = -\frac{1}{\sum_{i=1}^{N_q} |\mathcal{S}_i|} \sum_{i=1}^{N_q} \sum_{m \in \mathcal{S}_i} \log p\big(u_{i,m} \mid u_{i,<m}, \mathbf{X}^{\text{bev}}, \mathbf{X}_i^{\text{text}}\big)\,, \tag{6}$$

where $\mathcal{S}_i$ contains the supervised answer-token positions for question i; the loss averages over all non-masked answer tokens.

Semantic-geometric Query Refinement. SGQR bridges the shared LLM states and geometric lane queries $\mathbf{Q} \in \mathbb{R}^{N_{\text{qry}} \times d_q}$. Before lane decoding, we construct a semantic-geometric context from recovered BEV features and language states:

$$\mathbf{Z}^{\text{SG}} = \text{LN}\Big([\phi_b(\widetilde{\mathbf{F}}_{\text{bev}}); \text{Pool}(\phi_t(\mathbf{H}^{\text{text}}))]\Big)\,. \tag{7}$$

Here, ϕ_b projects and arranges the BEV grid as spatial tokens, while ϕ_t maps the dialogue states to the same query dimension d_q. The lane queries attend to this context as

$$\mathbf{Q}^{\text{SG}} = \mathbf{Q} + \text{MHA}\big(\mathbf{Q}, \mathbf{Z}^{\text{SG}}, \mathbf{Z}^{\text{SG}}\big)\,. \tag{8}$$

where MHA denotes multi-head cross-attention, with $\mathbf{Q}$ serving as the queries and the semantic-geometric context $\mathbf{Z}^{\text{SG}}$ as the keys and values. For clarity, Equation 8 summarizes the core attention update, omitting the standard normalization and feed-forward sublayers shown in Figure 2. Consequently, $\mathbf{Q}^{\text{SG}}$ integrates lane-specific geometric features with topology-oriented language context.

Perception. The lane decoder further updates the refined queries using the recovered BEV features. We reuse $\mathbf{Q}^{\text{SG}}$ to denote the decoder-updated queries, from which geometry, semantic attributes, and directed connectivity are predicted:

$$\begin{aligned} \mathbf{Q}^{\text{SG}} &= \text{LaneDecoder}\Big(\mathbf{Q}^{\text{SG}}, \widetilde{\mathbf{F}}_{\text{bev}}\Big)\,, \\ \hat{L} = f_{\text{lane}}(\mathbf{Q}^{\text{SG}}), &\qquad \hat{A}_{ij} = \sigma\big(f_{\text{rel}}(\mathbf{Q}_i^{\text{SG}}, \mathbf{Q}_j^{\text{SG}})\big)\,. \end{aligned} \tag{9}$$

Here, f_{lane} comprises MLP heads for 3D point regression and semantic classification. The MLP relation head f_{rel} predicts pairwise connectivity logits, which the sigmoid function σ converts to probabilities $\hat{A}_{ij}$, yielding the lane graph $(\hat{L}, \hat{\mathbf{A}})$.

Following prior lane-topology methods (Li et al., 2024b; Fu et al., 2024), Hungarian bipartite matching first establishes one-to-one assignments between the predicted lane queries and ground-truth lanes. The perception branch is then supervised by a lane detection loss and a topology loss for their connectivity:

$$\mathcal{L}_{\text{perception}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{topo}}. \tag{10}$$

The complete objective combines perception with the autoregressive understanding loss as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{perception}} + \lambda_{\text{vqa}}\mathcal{L}_{\text{vqa}}$, where λ_{vqa} balances the two branches. Detailed definitions are provided in Appendix B.

4 EXPERIMENTS

4.1 DATASETS AND METRICS

OpenLane-V2. We evaluate TOPOMIND on both subsets of OpenLane-V2 (Wang et al., 2023), each comprising 1,000 driving scenes with synchronized multi-view images sampled at 2 Hz and topology annotations. Subset-A is derived from Argoverse 2 and provides seven camera views, whereas Subset-B is derived from nuScenes and provides six. We focus on lane-to-lane topology reasoning: lane-segment topology is evaluated on Subset-A, while centerline topology is evaluated on both subsets. This protocol tests the same model formulation across distinct data sources, camera layouts, and lane representations.

OmniDrive-nuScenes. For flexible driving-scene understanding on Subset-B, we use the question–answer supervision provided by OmniDrive-nuScenes (Wang et al., 2025). In addition, we generate topology-oriented VQA supervision for both OpenLane-V2 Subset-A and Subset-B, enabling topology-oriented understanding to be learned across both subsets. This joint setting assesses whether language supervision benefits perception while retaining flexible question answering.

Metrics. For lane-segment perception, we follow LaneSegNet (Li et al., 2024b) and report AP_{ls} for lane segments and AP_{ped} for pedestrian crossings, with $\mathrm{mAP} = (\mathrm{AP}_{ls} + \mathrm{AP}_{ped})/2$. Lane-to-lane connectivity is evaluated using TOP_{lsls}. For centerline perception, we report the lane detection score DET_l and lane-to-lane topology score TOP_{ll}. Their aggregate lane-only score is denoted by L-OLS and computed as $\text{L-OLS} = (\mathrm{DET}_l + \sqrt{\mathrm{TOP}_{ll}})/2$. Appendix C provides the complete metric definitions, and Appendix A reports the implementation and training details.

4.2 MAIN RESULTS

Lane-segment topology reasoning. Table 1 compares TOPOMIND with existing lane-segment perception methods on OpenLane-V2 Subset-A. TOPOMIND achieves the best performance across all four metrics, surpassing the strongest prior results by 7.2 mAP points and 2.3 TOP_{lsls} points. The pronounced gains on both lane segments and pedestrian crossings are consistent with topology-oriented language supervision providing a useful auxiliary learning signal within the shared representation. The balanced improvements in AP_{ls} and AP_{ped} further show that the gain is not driven by a single map-element category.

Table 1: Lane-segment topology reasoning on OpenLane-V2 Subset-A. All methods are trained for 24 epochs.

Method	Venue	mAP	AP_{ls}	AP_{ped}	TOP_{lsls}
StreamMapNet (Yuan et al., 2024)	WACV24	20.3	22.1	18.6	13.2
TopoNet (Li et al., 2026b)	SCIS26	23.0	23.9	22.0	-
SQD-MapNet (Wang et al., 2024)	ECCV24	26.0	27.1	24.9	16.6
MapTR (Liao et al., 2022)	ICLR23	27.0	25.9	28.1	-
MapTRv2 (Liao et al., 2025)	IJCV25	28.5	26.6	30.4	-
LaneSegNet (Li et al., 2024b)	ICLR24	32.6	32.3	32.9	25.4
TopoLogic (Fu et al., 2024)	NeurIPS24	33.2	33.0	33.4	30.8
Topo2Seq (Yang et al., 2025b)	AAAI25	33.6	33.7	33.5	26.9
TopoFG (Xu et al., 2026)	AAAI26	34.4	33.8	35.1	31.3
TOPOMIND (Ours)	-	**41.6**	**42.1**	**41.1**	**33.6**
Improvement	-	**+7.2**↑	**+8.3**↑	**+6.0**↑	**+2.3**↑

Centerline topology reasoning. Table 2 reports single-frame centerline results on both OpenLane-V2 subsets. TOPOMIND consistently leads all centerline metrics, improving L-OLS by 6.0 points on Subset-A and 10.1 points on Subset-B. Its consistent advantage across two subsets with different scene distributions and camera configurations demonstrates the effectiveness of the unified representation: it recovers centerline geometry more accurately while preserving the directed connectivity required for topology reasoning. The gains in both DET_l and TOP_{ll} indicate that improved localization is accompanied by more reliable directed relations.

Table 2: Centerline topology reasoning on OpenLane-V2. All methods use 24 epochs except TopoFormer (40); § marks our reimplementation because runnable official code is unavailable.

Method	Venue	Subset-A			Subset-B		
		DET_l	TOP_{ll}	L-OLS	DET_l	TOP_{ll}	L-OLS
MapTR (Liao et al., 2022)	ICLR23	17.7	5.9	21.0	15.2	-	-
TopoNet (Li et al., 2026b)	SCIS26	28.6	10.9	30.8	24.3	6.7	25.1
TopoMLP (Wu et al., 2024)	ICLR24	28.5	21.7	37.5	26.0	21.0	35.9
LaneSegNet (Li et al., 2024b)	ICLR24	31.1	25.3	40.7	27.5	24.9	38.7
TopoLogic (Fu et al., 2024)	NeurIPS24	29.9	23.9	39.4	25.9	21.6	36.2
TopoPoint (Fu et al., 2026b)	NeurIPS25	31.4	28.7	42.5	31.2	28.3	42.2
TopoFormer§ (Lv et al., 2025)	CVPR25	34.7	24.1	41.9	34.8	23.2	41.5
TopoFG (Xu et al., 2026)	AAAI26	33.8	30.8	44.6	30.0	27.2	41.1
CoPo (Luo et al., 2026b)	KDD26	33.8	29.2	43.9	32.6	31.8	44.5
TopoMLP++ (Wu et al., 2026)	TPAMI26	29.1	22.0	38.0	27.6	22.0	37.3
TopoMind (Ours)	-	**42.5**	**34.3**	**50.6**	**45.8**	**40.2**	**54.6**
Improvement	-	**+7.8**↑	**+3.5**↑	**+6.0**↑	**+11.0**↑	**+8.4**↑	**+10.1**↑

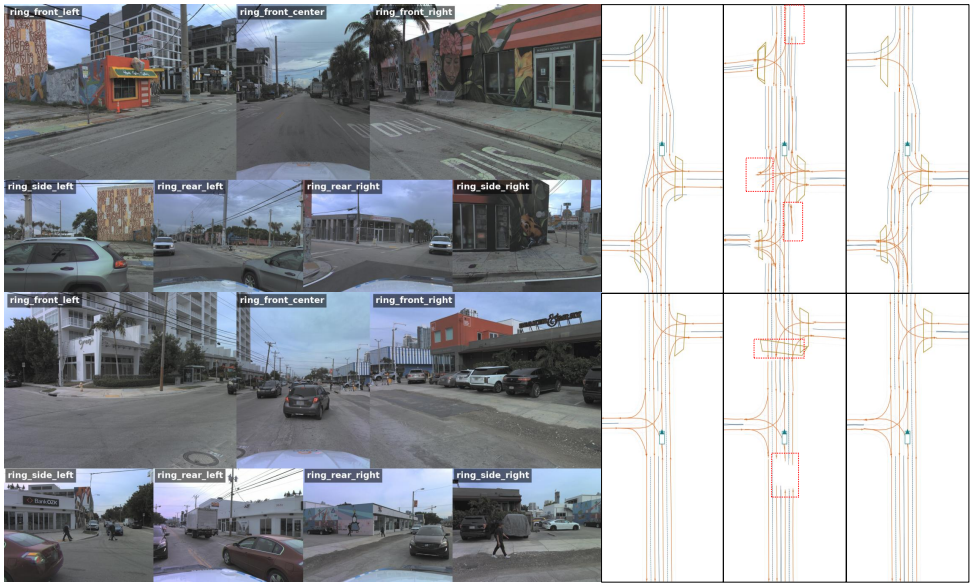


Figure 4: **Effect of topology-oriented understanding on OpenLane-V2 Subset-A.** For each scene, surround-view observations appear on the left; topology maps show the ground truth, the variant without understanding, and TopoMind from left to right. Without understanding, predictions contain incomplete lane geometry and missing or incorrect connections.

4.3 Ablation Studies

Table 3: Ablation of understanding and query refinement on OpenLane-V2 Subset-A.

Variant	mAP↑	AP_{ls} ↑	AP_{ped} ↑	TOP_{lsls} ↑
w/o understanding	36.9	35.8	37.9	28.4
LLM only (w/o text)	38.6	37.8	39.4	31.3
w/o SGQR	39.6	39.9	39.1	32.9
Direct query concat.	40.4	40.8	40.0	32.9
Full model	**41.6**	**42.1**	**41.1**	**33.6**

Table 4: Unified learning on OpenLane-V2 Subset-B (perception) and OmniDrive-nuScenes (understanding). All models are trained for 24 epochs.

Training setup	Perception / Understanding		
	DET_l/METEOR	TOP_{ll}/ROUGE	L-OLS/CIDEr
Understanding only	–/0.371	–/0.321	–/0.725
Perception only	36.7/–	31.4/–	46.3/–
Unified model	**45.8/0.380**	**40.2/0.323**	**54.6/0.735**

Component and fusion ablations. Table 3 ablates topology-oriented understanding and SGQR on Subset-A. Removing understanding reduces mAP and TOP_{lsls} by 4.7 and 5.2 points, respectively. Retaining the LLM without text recovers 1.7 mAP and 2.9 topology points over this variant, but remains 3.0 and 2.3 points below the full model, indicating that the gain depends on topology-oriented textual context rather than the LLM alone. Removing SGQR decreases mAP by 2.0 points. Direct query concatenation, which appends lane queries to the BEV and text token sequence for joint LLM processing, remains 1.2 mAP points and 0.7 topology points below SGQR, showing the advantage of explicit semantic-geometric refinement. The matched topology scores of the non-SGQR variants also expose the limitation of unstructured concatenation.

Unified perception and understanding. Table 4 evaluates centerline perception on OpenLane-V2 Subset-B and scene understanding on OmniDrive-nuScenes. Joint optimization improves all six metrics over the separate-model setting, with gains of 9.1 points in DET_l, 8.8 points in TOP_{ll}, and 8.3 points in L-OLS, together with improvements of 0.009 METEOR, 0.002 ROUGE, and 0.010 CIDEr. These results indicate that topology-oriented language supervision and centerline perception provide complementary training signals when jointly optimized within the unified model.

4.4 Qualitative Results

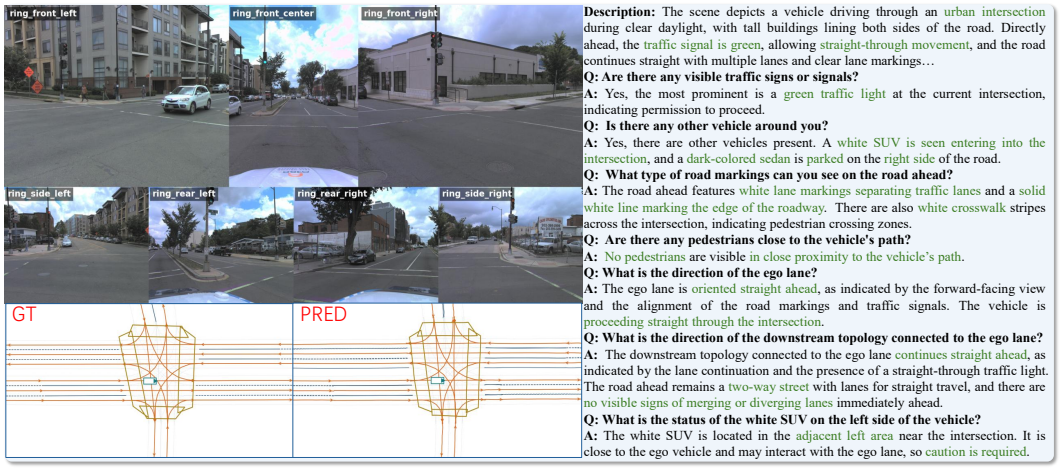


Figure 5: **Unified perception and understanding results of TopoMind.** The predicted lane graph closely follows the ground truth. Meanwhile, the generated responses correctly identify topology-relevant scene content, including the traffic signal, surrounding vehicles, road markings, ego-lane direction, and downstream connectivity, demonstrating that TopoMind supports both scene understanding and lane topology reasoning.

Figures 4 and 5 qualitatively show that topology-oriented understanding improves lane-graph completeness and that TopoMind provides grounded scene interpretations together with accurate topology predictions. This agreement with the quantitative results is consistent with the benefit of topology-oriented language supervision for structured lane-graph prediction.

5 Conclusion

In this paper, we presented TopoMind, a unified model for lane topology reasoning and topology-oriented scene understanding. Built on a shared BEV–language architecture, TopoMind integrates structured lane-graph prediction and language-based scene understanding from surround-view observations. We constructed topology-oriented VQA supervision over topology-relevant scene semantics, and introduced Semantic-Geometric Query Refinement to couple linguistic context and spatial geometry with lane queries. Experiments on OpenLane-V2 and OmniDrive-nuScenes demonstrate that TopoMind achieves state-of-the-art lane topology perception and improves scene understanding, extending the capabilities of conventional perception-only topology models.

AI USE STATEMENT

We used OpenAI Codex to assist with research-code implementation and debugging, language editing and polishing of the manuscript, LaTeX typesetting, and figure creation and refinement. All AI-generated or AI-assisted code and text were reviewed, revised, and validated by the authors before inclusion. The authors inspected the code and verified its behavior through experimental checks, checked the edited text and figures for technical accuracy and faithful representation of the reported work, and made all final decisions regarding the methodology, experimental design, analysis, and conclusions. The authors take full responsibility for the implementation, reported results, accuracy, originality, and final content of this submission.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. ORION: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24823–24834, 2025.

Jiahui Fu, Zehao Huang, Han Li, Naiyan Wang, and Si Liu. Generative lane topology reasoning via autoregressive model with geometry prior. *arXiv preprint arXiv:2606.31814*, 2026a.

Yanping Fu, Wenbin Liao, Xinyuan Liu, Hang Xu, Yike Ma, Yucheng Zhang, and Feng Dai. Topologic: An interpretable pipeline for lane topology reasoning on driving scenes. *Advances in Neural Information Processing Systems*, 37:61658–61676, 2024.

Yanping Fu, Xinyuan Liu, Tianyu Li, Yike Ma, Yucheng Zhang, and Feng Dai. Topopoint: Enhance topology reasoning via endpoint detection in autonomous driving. *Advances in Neural Information Processing Systems*, 38:118229–118250, 2026b.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Han Li, Yulu Gao, Si Liu, Yuhang Wang, Bo Liu, and Beipeng Mu. Unified modeling of lane and lane topology for driving scene reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026a.

Tianyu Li, Peijin Jia, Bangjun Wang, Li Chen, Kun Jiang, Junchi Yan, and Hongyang Li. Lanesegnet: Map learning with lane segment perception for autonomous driving. In *International Conference on Learning Representations*, volume 2024, pp. 1708–1724, 2024b.

Tianyu Li, Li Chen, Huijie Wang, Yang Li, Jiazhi Yang, Xiangwei Geng, Hang Xu, Chunjing Xu, Junchi Yan, Ping Luo, et al. Graph-based topology reasoning for driving scenes. *Science China Information Sciences*, 69(5):152103, 2026b.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024c.

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

Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, 133(3):1352–1374, 2025.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 936–944. Ieee, 2017.

Xueyi Liu, Zuodong Zhong, Yuxin Guo, Yun-Fu Liu, Zhiguo Su, Qichao Zhang, Junli Wang, Yinfeng Gao, Yupeng Zheng, Qiao Lin, et al. Reasonplan: Unified scene prediction and decision reasoning for closed-loop autonomous driving. *arXiv preprint arXiv:2505.20024*, 2025.

Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International conference on machine learning*, pp. 22352–22369. PMLR, 2023.

Yuechen Luo, Fang Li, Shaoqing Xu, Yang Ji, Zehan Zhang, Bing Wang, Yuannan Shen, Jianwei Cui, Long Chen, Guang Chen, et al. Last-vla: Thinking in latent spatio-temporal space for vision-language-action in autonomous driving. *arXiv preprint arXiv:2603.01928*, 2026a.

Yueru Luo, Changqing Zhou, Yiming Yang, Erlong Li, Chao Zheng, Shuguang Cui, and Zhen Li. Geometry-guided representations for coherent lane and traffic topology reasoning in driving scenes. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3504–3515, 2026b.

Changsheng Lv, Mengshi Qi, Liang Liu, and Huadong Ma. T 2 sg: Traffic topology scene graph for topology reasoning in autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17197–17206. IEEE, 2025.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4542–4550, 2024. doi: 10.1609/aaai.v38i5.28253.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15120–15130. IEEE, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krähenbühl, Marco Pavone, and Boris Ivanovic. Latent chain-of-thought world modeling for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39724–39733, 2026.

Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, et al. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems*, 36:18873–18884, 2023.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22442–22452. IEEE, 2025.

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

Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. Topomlp: A simple yet strong pipeline for driving topology reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 45604–45615, 2024.

Dongming Wu, Wencheng Han, Cheng-Zhong Xu, and Jianbing Shen. Topomlp++: Towards strong and scalable lane topology reasoning in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Guoqing Xu, Yiheng Li, and Yang Yang. Fine-grained representation for lane topology reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11214–11222, 2026.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.

Yiming Yang, Hongbin Lin, Yueru Luo, Suzhong Fu, Chao Zheng, Xinrui Yan, Shuqi Mei, Kun Tang, Shuguang Cui, and Zhen Li. Fastopowm: Fast-slow lane segment topology reasoning with latent world models. *arXiv preprint arXiv:2507.23325*, 2025a.

Yiming Yang, Yueru Luo, Bingkun He, Erlong Li, Zhipeng Cao, Chao Zheng, Shuqi Mei, and Zhen Li. Topo2seq: Enhanced topology reasoning via topology sequence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9318–9326, 2025b.

Yiming Yang, Yueru Luo, Bingkun He, Hongbin Lin, Suzhong Fu, Chao Zheng, Zhipeng Cao, Erlong Li, Chao Yan, Shuguang Cui, et al. Topostreamer: Temporal lane segment topology reasoning in autonomous driving. *arXiv preprint arXiv:2507.00709*, 2025c.

Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7341–7350. IEEE, 2024.

Zongzheng Zhang, Xinrun Li, Sizhe Zou, Guoxuan Chi, Siqi Li, Xuchong Qiu, Guoliang Wang, Guantian Zheng, Leichen Wang, Hang Zhao, et al. Chameleon: Fast-slow neuro-symbolic lane topology extraction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3752–3758. IEEE, 2025.

Xin Zhou, Dingkang Liang, Sifan Tu, Xiwu Chen, Yikang Ding, Dingyuan Zhang, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27817–27827. IEEE, 2025.

Zewei Zhou, Tianhui Cai, Seth Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *Advances in Neural Information Processing Systems*, 38:27920–27956, 2026.

Appendix

Contents

The appendix is organized into seven major parts.

A Implementation Details 13

B Loss Functions 14

C Metric Definitions 15

D Additional Perception and Understanding Results 16

E Additional Qualitative Results 16

F VQA Generation Prompts 17

G Limitations and Future Work 23

A Implementation Details

Framework refactoring. We refactor the original MMCV/MMDetection3D-based project into a standalone PyTorch implementation integrated with Hugging Face Transformers. The original detection stack couples model construction, data processing, training hooks, and mixed-precision handling to OpenMMLab-specific interfaces, increasing the integration effort for modern LLM components. Our refactoring replaces these dependencies throughout the data pipeline, BEV encoder, lane decoder, prediction heads, losses, and training loop. In particular, deformable attention uses the Transformers implementation, while a native PyTorch trainer handles distributed training, automatic mixed precision, optimization, and checkpointing. Transformers and PEFT provide LLM loading, tokenization, and LoRA adaptation. This separation reduces dependence on the legacy detection stack and facilitates integration with modern PyTorch APIs and evolving language-model backbones, while retaining the geometric and topological prediction interfaces needed by our task.

Accelerated language inference. We additionally implement an optional vLLM backend for Stage 2 question answering. It accepts the learned multimodal prompt embeddings, loads the checkpoint's exported LoRA adapter, and batches questions with support for tensor parallelism. BEV refinement and lane-topology decoding remain in the PyTorch perception branch. This design makes optimized autoregressive inference available to the unified model without requiring its geometric decoder to be reformulated as a language-generation module.

Model configuration. We use ResNet-50 with FPN as the image encoder and BEVFormer to aggregate calibrated surround-view features. The metric region spans 102.4×51.2 meters and is encoded as a 100×200 BEV grid with 256 channels. Before LLM processing, 4×4 average pooling reduces the grid to 25×50; a 1,024-dimensional interface projection followed by the LLM input projection produces 1,250 BEV tokens in the Qwen3-VL hidden space. The perception branch uses 200 lane queries. On Subset-A, each predicted line is represented by 10 ordered 3D points, whereas each Subset-B centerline is represented by 11 ordered 2D points; a zero height coordinate is appended only when exporting Subset-B predictions to the benchmark format. In Stage 2, we use Qwen3-VL-2B-Instruct with rank-128 LoRA, and set the VQA loss weight to $\lambda_{\text{vqa}} = 1.0$. VQA supervision is generated offline with the larger, frozen Qwen3-VL-32B-Instruct teacher through vLLM; the teacher is used only for annotation generation and is distinct from the Qwen3-VL-2B-Instruct model trained within TopoMind.

Two-stage training. All Subset-A and Subset-B models follow the two-stage protocol in Table 5 on eight NVIDIA A100 GPUs. Both stages use one sample per GPU, yielding a global batch size of 8. Stage 1 learns the single-frame geometric and topological perception model from surround-view images. Stage 2 initializes the perception modules from the Stage 1 checkpoint, introduces the LLM and SGQR, and jointly optimizes lane perception, lane-to-lane topology, and topology-oriented VQA

understanding. Both stages use AdamW with weight decay 0.01 and gradient clipping at a maximum norm of 35. The base learning rate is 2×10^{-4} for trainable parameters, while the image backbone uses a multiplier of 0.1, corresponding to 2×10^{-5}.

Table 5: Configuration of the two-stage training protocol. The global batch size is 8, with one sample per GPU on eight GPUs.

Configuration	Stage 1	Stage 2
Initialization	ImageNet backbone	Stage 1 checkpoint
Optimizer	AdamW	AdamW
Base learning rate	2×10^{-4}	2×10^{-4}
Backbone learning rate	2×10^{-5}	2×10^{-5}
Epochs	12	12
LR scheduler	Cosine annealing Linear, 500 iterations	Cosine annealing Linear, 500 iterations
Warmup	initial ratio $1/3$	initial ratio $1/3$
Minimum LR ratio	10^{-3}	10^{-3}
Global batch size	8	8
Hardware	$8\times$ NVIDIA A100	$8\times$ NVIDIA A100
Training tasks	Lane detection and lane topology	Lane detection, lane topology, and VQA understanding

B Loss Functions

Perception objective. Following the LaneSegNet assigner adopted in our implementation (Li et al., 2024b), the N_{qry} predictions $\hat{L}$ are first assigned one-to-one to the N_l ground-truth lanes L. For a ground-truth lane L_i and a prediction $\hat{L}_j$, the matching objective is

$$\begin{aligned} \mathcal{C}_{\text{match}}(L_i, \hat{L}_j) &= \lambda_{\text{cls}}\mathcal{C}_{\text{cls}}(L_i, \hat{L}_j) + \lambda_{\text{vec}}\mathcal{C}_{\text{vec}}(L_i, \hat{L}_j) \\ &\quad + \lambda_{\text{seg}}\big(\lambda_{\text{ce}}\mathcal{C}_{\text{ce}}(L_i, \hat{L}_j) + \lambda_{\text{dice}}\mathcal{C}_{\text{dice}}(L_i, \hat{L}_j)\big), \\ \hat{\pi} &= \arg\min_{\pi \in \Pi(N_l, N_{\text{qry}})} \sum_{i=1}^{N_l} \mathcal{C}_{\text{match}}(L_i, \hat{L}_{\pi(i)}). \end{aligned} \tag{11}$$

Here, $\mathcal{C}_{\text{cls}}$ is the focal classification cost for the lane category, $\mathcal{C}_{\text{vec}}$ is the ℓ_1 cost between the ordered 3D point representations, and $\mathcal{C}_{\text{ce}}$ and $\mathcal{C}_{\text{dice}}$ measure the cross-entropy and Dice discrepancies between the predicted and ground-truth lane masks. $\Pi(N_l, N_{\text{qry}})$ denotes the set of injective assignments from ground-truth lanes to predictions. Lane-boundary types and lane connectivity are supervised after the assignment and therefore do not enter $\mathcal{C}_{\text{match}}$. After matching, the perception losses are defined as

$$\begin{aligned} \mathcal{L}_{\text{det}} &= \lambda_{\text{vec}}\mathcal{L}_{\text{vec}} + \lambda_{\text{seg}}\big(\lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}\big) + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{type}}\mathcal{L}_{\text{type}}, \\ \mathcal{L}_{\text{topo}} &= \lambda_{\text{top}}\ \text{Focal}(\hat{\mathbf{A}}^{\hat{\pi}}, \mathbf{A}). \end{aligned} \tag{12}$$

Here, $\mathcal{L}_{\text{vec}}$ is the ℓ_1 loss on the ordered 3D point sequences $\mathbf{P}_i$ and $\hat{\mathbf{P}}_{\hat{\pi}(i)}$; for lane segments it is applied to the left boundary, centerline, and right boundary. $\mathcal{L}_{\text{cls}}$ is a focal loss for lane or pedestrian-crossing classification, $\mathcal{L}_{\text{type}}$ is cross-entropy on left- and right-boundary types, and the cross-entropy and Dice terms supervise the lane-segment mask. Classification also supervises unmatched queries as background, whereas geometric, boundary-type, and mask supervision applies to matched predictions. For centerline training, only the prediction terms associated with centerline geometry and semantics are retained. The aligned adjacency matrix is defined by $(\hat{\mathbf{A}}^{\hat{\pi}})_{ij} = \hat{A}_{\hat{\pi}(i),\hat{\pi}(j)}$, so $\mathcal{L}_{\text{topo}}$ applies focal loss to aligned predicted and ground-truth connections. Under the notation in Equation 12, we use $\lambda_{\text{cls}} = 1.5$, $\lambda_{\text{vec}} = 0.025$, $\lambda_{\text{seg}} = 1$, $\lambda_{\text{ce}} = 3$, $\lambda_{\text{dice}} = 3$, $\lambda_{\text{type}} = 0.1$, and $\lambda_{\text{top}} = 5$. The Hungarian matcher uses the same effective classification, vector, cross-entropy-mask, and Dice weights. These coefficients are absorbed into $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{topo}}$ in Equation 10.

Understanding objective. For the i-th topology-oriented question, teacher forcing conditions every answer token $u_{i,m}$ on its ground-truth prefix $u_{i,<m}$ and the multimodal scene tokens. Equation 6 is

therefore the mean autoregressive negative log-likelihood over non-masked tokens, evaluated only at assistant-answer positions $m \in \mathcal{S}_i$; question, system-prompt, padding, and visual-token positions are masked from the loss.

Stage-wise objectives. Stage 1 optimizes $\mathcal{L}_{\text{perception}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{topo}}$, exactly as defined in Equation 10. Stage 2 additionally introduces topology-oriented language supervision and optimizes $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{perception}} + \lambda_{\text{vqa}}\mathcal{L}_{\text{vqa}}$, matching the complete objective stated in the main text. This joint objective allows language supervision to shape the shared LLM representation while retaining direct geometric and graph supervision for perception.

C Metric Definitions

We follow the official OpenLane-V2 protocol (Wang et al., 2023) for centerline detection and lane-to-lane topology, and the LaneSegNet protocol (Li et al., 2024b) for lane-segment perception.

Centerline detection. Let $\mathbf{P}_i$ and $\hat{\mathbf{P}}_j$ denote the ordered 3D point sequences of ground-truth centerline L_i and predicted centerline $\hat{L}_j$ after evaluation preprocessing. Their lengths need not equal the model's output point count K; the Subset-B prediction formatter resamples centerlines to 11 points. Their geometric discrepancy is the discrete Fréchet distance

$$D_{\text{F}}(\mathbf{P}_i, \hat{\mathbf{P}}_j) = \min_{\gamma \in \Gamma(\mathbf{P}_i, \hat{\mathbf{P}}_j)} \max_{(\mathbf{p}, \hat{\mathbf{p}}) \in \gamma} \|\mathbf{p} - \hat{\mathbf{p}}\|_2, \tag{13}$$

where $\Gamma(\mathbf{P}_i, \hat{\mathbf{P}}_j)$ is the set of monotone couplings between the two point sequences. Matching requires a distance strictly below t, after the benchmark's distance-dependent relaxation for remote lanes. The score DET_l averages 11-point interpolated AP over $\Theta_l = \{1.0, 2.0, 3.0\}$ meters. Let $p_t(\rho)$ denote precision at recall ρ:

$$\text{DET}_l = \frac{1}{|\Theta_l|} \sum_{t \in \Theta_l} \text{AP}_t, \qquad \text{AP}_t = \frac{1}{11} \sum_{j=0}^{10} \max_{\rho \geq j/10} p_t(\rho). \tag{14}$$

The maximum is zero when no attained recall satisfies the threshold. The evaluator multiplies the geometric distance by $\max(0.5, 1 - 0.005 d_{\text{ego}})$, where d_{ego} is the ground-truth centerline's nearest distance to the ego origin.

Lane-segment and pedestrian-crossing detection. Following the representation in Section 3.1, a ground-truth lane segment is $L_i = (\mathbf{P}_i^l, \mathbf{P}_i^c, \mathbf{P}_i^r)$ and its prediction is $\hat{L}_j = (\hat{\mathbf{P}}_j^l, \hat{\mathbf{P}}_j^c, \hat{\mathbf{P}}_j^r)$. LaneSegNet measures their distance by

$$\begin{aligned} D_{ls}(L_i, \hat{L}_j) = \frac{1}{2}\Big[& D_{\text{F}}(\mathbf{P}_i^c, \hat{\mathbf{P}}_j^c) + \text{Chamfer}(\mathbf{P}_i^l, \hat{\mathbf{P}}_j^l) \\ & + \text{Chamfer}(\mathbf{P}_i^r, \hat{\mathbf{P}}_j^r)\Big]. \end{aligned} \tag{15}$$

The boundary terms measure geometric agreement separately for the left and right boundaries, whereas the centerline Fréchet term retains lane direction. The evaluator applies the same distance-dependent relaxation and 11-point AP calculation. The resulting AP_{ls} is averaged over matching thresholds $\{1.0, 2.0, 3.0\}$ meters. Pedestrian crossings are non-directional and are therefore matched solely by Chamfer distance at thresholds $\{0.5, 1.0, 1.5\}$ meters to obtain AP_{ped}. We report their mean as $\text{mAP} = (\text{AP}_{ls} + \text{AP}_{ped})/2$.

Lane topology. OpenLane-V2 evaluates directed connectivity on geometrically matched graphs. For each matching threshold, predictions are aligned to ground-truth lanes using the official evaluator's unmatched-lane convention. For lane L_i and direction $d \in \{\text{in}, \text{out}\}$, let $\mathcal{N}_i^d$ be the ground-truth neighbors. Predicted neighbors $\hat{\mathcal{N}}_i^d$ retain connections with probability strictly greater than 0.5 and are ranked by confidence. With $\text{Prec}_i^d(k)$ and $\text{rel}_i^d(k)$ denoting rank-k precision and relevance,

respectively, the per-lane, per-direction AP is

$$\mathrm{AP}_i^d = \begin{cases} 1, & |\mathcal{N}_i^d| = |\hat{\mathcal{N}}_i^d| = 0, \\ 0, & \text{exactly one neighbor set is empty}, \\ \dfrac{1}{|\mathcal{N}_i^d|} \displaystyle\sum_{k=1}^{|\hat{\mathcal{N}}_i^d|} \mathrm{Prec}_i^d(k)\, \mathrm{rel}_i^d(k), & \text{otherwise.} \end{cases} \tag{16}$$

The final TOP averages these values over both directions, all ground-truth lanes in nonempty frames, and the corresponding geometric matching thresholds. For an unmatched lane, the evaluator treats its incident connections as incorrect by assigning above-threshold scores to absent ground-truth edges and zero to present edges. TOP_{ll} applies this definition to the directed centerline graph, while TOP_{lsls} applies it to the directed lane-segment graph. Finally, we summarize the centerline-only performance by $\text{L-OLS} = (\mathrm{DET}_l + \sqrt{\mathrm{TOP}_{ll}})/2$, where both inputs are normalized to $[0, 1]$ before aggregation; the tables report all three scores on a 0–100 scale.

D Additional Perception and Understanding Results

Table 6 combines centerline topology results on OpenLane-V2 Subset-B with scene-understanding results on OmniDrive-nuScenes. TopoMind supports both tasks within one model and achieves the best results across the six reported metrics. The understanding baselines may support other driving tasks; *Unsupported* refers specifically to the lane-topology or language outputs evaluated here.

Table 6: Centerline topology perception on OpenLane-V2 Subset-B and scene understanding on OmniDrive-nuScenes. LLM Params denotes language-model size. *Unsupported* indicates that a method does not provide the corresponding evaluated outputs; – denotes an unavailable score or an unspecified/not applicable parameter count. § follows the reimplementation convention in Table 2.

Method	Venue	LLM Params	Centerline perception			Understanding		
			DET_l ↑	TOP_{ll} ↑	L-OLS ↑	METEOR↑	ROUGE↑	CIDEr↑
Lane-topology perception methods								
MapTR (Liao et al., 2022)	ICLR23	–	15.2	–	–	*Unsupported*		
TopoNet (Li et al., 2026b)	SCIS26	–	24.3	6.7	25.1			
TopoMLP (Wu et al., 2024)	ICLR24	–	26.0	21.0	35.9			
LaneSegNet (Li et al., 2024b)	ICLR24	–	27.5	24.9	38.7			
TopoLogic (Fu et al., 2024)	NeurIPS24	–	25.9	21.6	36.2			
TopoPoint (Fu et al., 2026b)	NeurIPS25	–	31.2	28.3	42.2			
TopoFormer§ (Lv et al., 2025)	CVPR25	–	34.8	23.2	41.5			
TopoFG (Xu et al., 2026)	AAAI26	–	30.0	27.2	41.1			
CoPo (Luo et al., 2026b)	KDD26	–	32.6	31.8	44.5			
TopoMLP++ (Wu et al., 2026)	TPAMI26	–	27.6	22.0	37.3			
Driving-scene understanding methods								
LLaVA-OV (Li et al., 2024a)	TMLR25	7B	*Unsupported*			–	0.221	0.284
OmniDrive-BEV (Wang et al., 2025)	CVPR25	7B				0.356	0.278	0.595
ORION (Fu et al., 2025)	ICCV25	7B				0.354	0.306	0.635
Unified lane-topology perception and understanding								
TopoMind (Ours)	–	2B	**45.8**	**40.2**	**54.6**	**0.380**	**0.323**	**0.735**
Improvement	–	–	**+11.0** ↑	**+8.4** ↑	**+10.1** ↑	**+0.024** ↑	**+0.017** ↑	**+0.100** ↑

E Additional Qualitative Results

Figure 6 provides additional qualitative comparisons between the no-language variant and the full model on OpenLane-V2 Subset-B.



Figure 6: **Qualitative comparison on OpenLane-V2 Subset-B.** Each row shows synchronized surround-view observations on the left and centerline topology maps on the right. From left to right, the maps are the ground truth (GT), the no-language variant, and the full model. Red dashed boxes highlight missing or incorrect centerlines and connectivity in the no-language predictions. The full model more closely matches the ground truth, illustrating that topology-oriented language supervision improves both centerline geometry and topology reasoning.

F VQA Generation Prompts

This section reports the core prompts used to generate topology-oriented scene-understanding supervision. We use the frozen Qwen3-VL-32B-Instruct teacher described in Appendix A to generate additional annotations offline, complementing the reused OmniDrive question–answer pairs. Following prior scene-aware VQA generation practice for autonomous driving (Wang et al., 2025), we first elicit a camera-aware scene overview and then construct complementary template-driven and scene-adaptive question–answer pairs. The wording below is typeset for readability while preserving the generation instructions and output constraints. As a qualitative sanity check, we manually inspect a subset of the generated VQA pairs and correct obvious inaccuracies, inconsistencies, or insufficiently grounded responses when encountered. These manual inspections are intended as qualitative sanity checks rather than a dataset-wide accuracy audit.

Corpus composition. The corpus combines teacher-generated fixed-template supervision on both OpenLane-V2 subsets, teacher-generated scene-adaptive conversations on Subset-A, and scene-adaptive question–answer pairs reused from OmniDrive-nuScenes for Subset-B. After converting the assembled conversations into individual pairs, the corpus contains 1,188,301 question–answer pairs, whose content distribution is reported in Figure 3. Stage 2 training uses only the OpenLane-V2 training splits, comprising 22,477 Subset-A frames and 27,968 Subset-B frames; validation annotations remain held out from optimization.

Camera-aware scene overview. For Subset-A, the forward composite is ordered as front-left, front-center, and front-right, while the second composite is ordered as side-left, rear-left, rear-right, and side-right. For the six-camera Subset-B, the corresponding composites contain CAM_FRONT_LEFT, CAM_FRONT, and CAM_FRONT_RIGHT, followed by CAM_BACK_LEFT, CAM_BACK, and CAM_BACK_RIGHT. The rear-facing images are horizontally mirrored during composition to make their displayed orientation consistent. The teacher receives the two composites together with a subset-specific directional instruction:

> The first panoramic composite depicts the left-front, directly front, and right-front surroundings of the ego vehicle. The second composite depicts either the available side and rear surroundings for Subset-A or the left-rear, directly rear, and right-rear surroundings for

> Subset-B. Use the supplied directions for spatial reasoning and describe all observations from the driver's perspective without mentioning image indices or camera names.
>
> First summarize the current driving scene. Describe the road condition and scene type, such as an intersection, roundabout, or parking area, together with relevant vehicles, pedestrians, traffic signs or lights, road markings, weather, and time of day. Report only content supported by the provided images and avoid speculation.

Template-driven questions. For every frame, the teacher answers the following fixed questions:

1. Can you provide a summary of the current driving scenario based on the input images?
2. Are there any visible traffic signs or signals?
3. Is there any other vehicle around you?
4. What type of road markings can you see on the road ahead?
5. Are there any pedestrians close to the vehicle's path?
6. What is the direction of the ego lane?
7. What is the direction of the downstream topology connected to the ego lane?

Ego-centric directional questions. Questions 6 and 7 provide coarse local driving-direction supervision for the current ego lane and its immediate downstream continuation. Their answers describe only ego-centric directional categories, such as straight, left, right, or their combinations. They do not encode lane coordinates, lane identities, connectivity among non-ego lanes, pairwise adjacency relations, or the complete lane graph. During offline annotation generation, structured lane information is used only to disambiguate the coarse direction of the ego lane and its immediate continuation. The resulting directional descriptions are used as auxiliary language supervision during training and are not provided to TopoMind at evaluation time. We emphasize that this ego-centric directional supervision is substantially weaker than the target of lane-topology perception: the perception branch must recover all lane instances and predict directed connectivity among all candidate lane pairs, whereas the two directional questions concern only a local ego-centered path. Full lane geometry and connectivity therefore remain directly supervised by the perception objectives rather than by the language annotations.

Scene-adaptive questions. After receiving the generated scene overview, the teacher is prompted as follows:

> Given the panoramic driving scene and its scene overview, generate diverse question–answer pairs about the visible setting, road structure, object types and counts, object actions, relative positions, and topology-relevant events. Include both direct visual questions and more complex contextual questions when they are supported by the scene. Ask only questions whose answers can be determined confidently from the supplied evidence, and do not ask about uncertain or unobservable content. Provide detailed answers when reasoning is required. Return each pair in the JSON form {`"question": "...", "answer": "..."`}.

F.1 Qualitative Examples of VQA Supervision

Figures 7–10 show four selected examples of the VQA supervision, with synchronized surround-view images and ground-truth BEV geometry for direct inspection. Each example contains a scene description, six template-driven questions, and three scene-adaptive questions. The template-driven answers cover traffic controls, nearby vehicles, road markings, pedestrians, ego-lane direction, and downstream connectivity; the adaptive questions add environmental context, infrastructure functions, or driving considerations. The examples span an all-way-stop junction, a signalized junction, a pedestrian crossing beside construction, and a rainy urban intersection. Highlighted phrases illustrate correspondences between the language annotations and observable scene content or annotated lane connectivity, showing how the supervision links scene semantics to road topology. For Subset-B, the scene-adaptive pairs are drawn from OmniDrive-nuScenes (Wang et al., 2025), following the data composition described in the main text. These selected examples provide qualitative evidence of coverage and grounding rather than a dataset-wide accuracy estimate; highlighting applies only to the marked phrases and does not certify every detail in an answer.



Description

The vehicle is navigating through a sunny, urban environment characterized by modern buildings, palm trees, and wide sidewalks. The surroundings suggest a well-maintained city district with a mix of commercial and residential architecture. The road is relatively clear, with a few parked cars along the side and minimal traffic flow ahead, indicating a calm driving situation in a developed neighborhood.

Q: Are there any visible traffic signs or signals?

A: Yes, a prominent red octagonal "STOP" sign with "ALL WAY" underneath is visible at the upcoming intersection. Additionally, street signs indicating "NE 32nd STREET" and "1st AVENUE" are mounted above the stop sign, providing clear location information. A blue parking direction sign is also visible further down the sidewalk.

Q: Is there any other vehicle around you?

A: There are several vehicles in the vicinity. A white SUV is seen crossing the intersection directly ahead. Along the right side of the road, multiple cars are parked, including a silver Range Rover and a white sedan. Further down the street, an orange vehicle is approaching in the distance.

Q: What type of road markings can you see on the road ahead?

A: The road features faded white crosswalk markings near the intersection, indicating a pedestrian crossing area. There are also yellow painted curbs and a yellow tactile paving strip at the corner, likely to guide visually impaired pedestrians. The pavement shows some wear, with cracks visible in the crosswalk area.

Q: Are there any pedestrians close to the vehicle's path?

A: A pedestrian is visible on the sidewalk to the left, pushing a shopping cart near a streetlight. They are at a safe distance from the vehicle's path and are not crossing the road. No other pedestrians are immediately near the vehicle's current trajectory.

Q: What is the direction of the ego lane?

A: The ego lane is oriented straight ahead, continuing along the current road path without any immediate turns or bends. The vehicle is proceeding forward through an intersection, maintaining its course on what appears to be a straight thoroughfare.

Q: What is the direction of the downstream topology connected to the ego lane?

A: The downstream topology connected to the current lane includes three possible directions: turning left into a lane, turning right into a lane, or continuing straight into a lane. The intersection ahead, marked by the stop sign, allows for these three navigational choices, with the straight path continuing forward along the current road.

Q: What type of buildings are visible in the surrounding environment?

A: The surrounding environment features modern multi-story buildings with glass facades and commercial storefronts. Some buildings have visible signage, such as 'LUXURYHOMES.COM', indicating a commercial or downtown urban setting.

Q: Why is it important to check the left side of the intersection before proceeding?

A: It is important to check the left side of the intersection because a pedestrian is visible walking near the sidewalk on that side. Additionally, since this is an 'ALL WAY' stop intersection, vehicles approaching from the left may also have the right-of-way if they arrived before or at the same time as the vehicle. Ensuring the left side is clear prevents collisions with pedestrians or other vehicles.

Q: What is the likely purpose of the yellow curb on the right side of the road?

A: The yellow curb on the right side of the road typically indicates a no-parking or restricted parking zone, often used to ensure clear access for emergency vehicles, buses, or to maintain traffic flow. In this context, it may also be designated for loading zones or to prevent vehicles from stopping near the intersection, which could obstruct visibility or impede traffic.

Figure 7: **Qualitative VQA supervision.** Subset-A: an all-way-stop junction with left-turn, right-turn, and straight downstream connections. Green phrases indicate statements grounded either in observable scene content or in structured lane annotations used to construct topology-oriented supervision. The BEV depicts ground-truth annotations for qualitative inspection rather than model predictions.



Description

The vehicle is navigating an urban intersection on a clear, sunny day. It appears to be approaching or moving through a signalized intersection with multiple lanes, crosswalks, and surrounding commercial buildings. The road ahead is straight, with other vehicles present, including a white delivery truck and a gray minivan, and the environment includes sidewalks, streetlights, and some greenery.

Q: Are there any visible traffic signs or signals?

A: Yes, there are several traffic signals visible. At the intersection, the traffic light for the current direction of travel is green, while the lights for the cross street are red and yellow. There are also street signs indicating "Chicon St" and "E 7th St," and a "Left Turn Yield on Green" sign is visible near the intersection.

Q: Is there any other vehicle around you?

A: Yes, there are several other vehicles around. Directly ahead is a white box truck, and a gray minivan is seen in the adjacent lane to the left. Across the intersection, a white pickup truck is making a turn. Additional vehicles are visible further down the road and parked along the sides of the street.

Q: What type of road markings can you see on the road ahead?

A: The road ahead features double yellow lines separating opposing lanes, a crosswalk with red and gray stripes at the intersection, and white dashed lane markings. There are also painted road markings such as "ONLY" indicating lane restrictions. The pavement shows some wear and minor cracks.

Q: Are there any pedestrians close to the vehicle's path?

A: No pedestrians are visible in close proximity to the vehicle's path. The sidewalks are clear, and although there is a bus stop shelter with a bench, no one is present near the road or crosswalks in the immediate vicinity.

Q: What is the direction of the ego lane?

A: The ego lane is oriented straight ahead, as indicated by the forward-facing view and the alignment of the road and traffic signals. The vehicle is traveling in a straight path through the intersection without any immediate turns.

Q: What is the direction of the downstream topology connected to the ego lane?

A: The downstream topology connected to the ego lane includes two left-turn lanes, both designated for turning left. These lanes diverge from the main straight path ahead, allowing vehicles to make left turns at the upcoming intersection.

Q: What structures or features are visible on the left side of the road?

A: On the left side of the road, there is a continuous stone wall running parallel to the roadway, with a sidewalk behind it. Beyond the wall, there are trees, a gated area, and a multi-story building with a distinctive tower-like structure. The wall appears to be made of large, light-gray stone blocks.

Q: What is the weather condition, and how does it affect visibility?

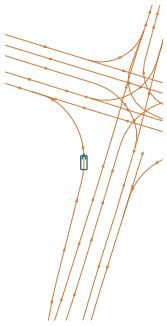
A: The weather is sunny with clear blue skies, providing excellent visibility. The bright sunlight casts distinct shadows on the road and surrounding structures, enhancing contrast and making road markings, vehicles, and traffic signals clearly visible. There is no glare or obstruction from rain, fog, or haze.

Q: What is the likely purpose of the covered bench on the right side of the road?

A: The covered bench on the right side of the road is likely a public seating area or a bus stop shelter. Its placement on the sidewalk near a building and adjacent to a street with light traffic suggests it is intended for pedestrians to rest or wait for transportation. The canopy provides shade and protection from weather, and its proximity to a sidewalk and crosswalk supports its use as a pedestrian-friendly feature.

Figure 8: **Qualitative VQA supervision.** Subset-A: a signalized urban junction with roadside structures and a covered bench. Green phrases indicate statements grounded either in observable scene content or in structured lane annotations used to construct topology-oriented supervision. The BEV depicts ground-truth annotations for qualitative inspection rather than model predictions.





Description

The current driving scenario depicts an urban environment with moderate traffic and ongoing construction. The vehicle is positioned at an intersection with a pedestrian crossing, where a man is actively walking across the zebra crossing. There are multiple pedestrians gathered on the sidewalk, and construction equipment, including an excavator and barriers, is visible on the side of the road. The surroundings include modern buildings, trees, and overcast skies, suggesting a typical city street during daytime.

Q: Are there any visible traffic signs or signals?

A: Yes, there are several visible traffic signs and signals. A red traffic light is clearly visible at the intersection, indicating a stop for vehicles in the current lane. Additionally, a "GIVE WAY" sign is present near the construction zone, and a street sign reading "Fusionopolis Way" is visible in another view, indicating the name of the road. Other traffic lights are also visible further down the road, some of which are red.

Q: Is there any other vehicle around you?

A: Yes, there are multiple vehicles around. A dark-colored sedan is seen turning left ahead of the vehicle, and a white SUV and a maroon SUV are visible on the adjacent lane or at the intersection. Further in the distance, a green truck and other cars can be seen near the construction area, indicating active traffic flow in the vicinity.

Q: What type of road markings can you see on the road ahead?

A: The road ahead features white lane markings, including a directional arrow painted on the road surface indicating a left turn. There is also a zebra crossing with white stripes for pedestrian use, and lane dividers are visible along the road. Orange and white construction barriers line parts of the road, marking off the construction zone.

Q: Are there any pedestrians close to the vehicle's path?

A: Yes, there are pedestrians close to the vehicle's path. One man is actively crossing the zebra crossing directly in front of the vehicle, while a group of pedestrians is standing on the sidewalk to the right, waiting to cross. Additional pedestrians can be seen further ahead near the construction site, indicating a busy pedestrian environment.

Q: What is the direction of the ego lane?

A: The direction of the ego lane is a left turn. The vehicle is positioned in a lane designated for turning left, as confirmed by the painted left-turn arrow on the road surface and the presence of a zebra crossing where pedestrians are crossing in the direction of the turn.

Q: What is the direction of the downstream topology connected to the ego lane?

A: The downstream topology connected to the ego lane is also a left-turning lane. This lane continues straight ahead after the turn and is flanked by construction barriers and modern high-rise buildings. The road ahead is wide and appears to connect to a larger urban thoroughfare, with an overpass visible in the distance.

Q: What type of construction equipment is present at the site to the right?

A: There is an orange digger present at the construction site to the right, along with barriers indicating ongoing construction work.

Q: Are there any notable buildings or structures in the vicinity?

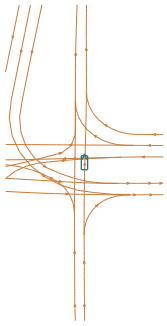
A: There are several buildings in the vicinity, including commercial buildings and one with an architecture that could indicate a governmental or institutional function. There's also a pedestrian bridge overhead.

Q: What is the weather condition as observed in the scene?

A: The weather appears to be overcast with no direct sunlight, which suggests it is either a cloudy day or possibly the light conditions are subdued due to the time of day.

Figure 9: **Qualitative VQA supervision.** Subset-B: a pedestrian crossing beside construction along a left-turning connection. Green phrases indicate statements grounded either in observable scene content or in structured lane annotations used to construct topology-oriented supervision. The BEV depicts ground-truth annotations for qualitative inspection rather than model predictions.





Description

The driving scenario takes place on a wet urban street during overcast or rainy weather, with reflective road surfaces and reduced visibility. The vehicle is navigating through city blocks lined with modern buildings, trees, and commercial storefronts, including one named "Casa de Pedro." Traffic flow appears moderate, with vehicles ahead and at intersections, and construction activity is visible in the background with cranes and barriers.

Q: Are there any visible traffic signs or signals?

A: Yes, multiple traffic signals and signs are visible. In several views, traffic lights are shown, some displaying green for straight movement and others red for stopping. A "NO TURN ON RED" sign is visible at an intersection, and a street sign indicating "D STREET" is also present. Additional signage includes a "P" parking sign and directional arrows at intersections.

Q: Is there any other vehicle around you?

A: Yes, there are several other vehicles in the vicinity. Ahead, a dark sedan is visible traveling in the same direction. At intersections, multiple cars are seen waiting or moving, including a white SUV, a silver sedan, and a black pickup truck. A white van is also visible in the distance on a side street, and a red scooter is parked near a curb in one view.

Q: What type of road markings can you see on the road ahead?

A: The road ahead features standard lane markings, including solid and dashed white lines separating lanes. There are also crosswalk markings at intersections, and yellow lane dividers are visible near construction zones. The wet pavement reflects the road markings, making them appear more pronounced despite the rain.

Q: Are there any pedestrians close to the vehicle's path?

A: No pedestrians are clearly visible in close proximity to the vehicle's immediate path. However, in some images, figures can be seen in the distance near building entrances or sidewalks, but they are not crossing the road or approaching the vehicle's trajectory.

Q: What is the direction of the ego lane?

A: The ego lane is oriented straight ahead, as indicated by the forward-facing perspective and the alignment of the road and traffic signals. The vehicle is proceeding straight through intersections, with no indication of turning or changing direction in the current sequence.

Q: What is the direction of the downstream topology connected to the ego lane?

A: The downstream topology connected to the ego lane continues straight ahead, merging into a broader urban thoroughfare. The road ahead maintains a straight alignment with minimal curvature, leading toward a distant intersection or city block, and is connected to lanes that allow for continued forward movement without immediate branching or turning.

Q: What weather conditions are apparent in this scene?

A: The weather is overcast with wet road conditions, suggesting recent rainfall.

Q: What type of area does this driving scenario take place in?

A: This is a mixed-use urban area with commercial buildings, indicative of a city environment.

Q: What driving behavior is recommended in this situation?

A: The driver should maintain a moderate speed due to the wet conditions, keep the current lane as the road ahead is clear, and proceed straight following the traffic signals and lane markings.

Figure 10: **Qualitative VQA supervision.** Subset-B: a rainy urban intersection with a straight downstream connection. Green phrases indicate statements grounded either in observable scene content or in structured lane annotations used to construct topology-oriented supervision. The BEV depicts ground-truth annotations for qualitative inspection rather than model predictions.

G LIMITATIONS AND FUTURE WORK

The current formulation of TOPOMIND operates on single-frame observations. While this setting isolates the interaction between scene understanding and lane topology perception, it does not exploit temporal cues for estimating agent motion. Incorporating multi-frame observations or temporal BEV memory could strengthen motion-aware scene understanding and support more reliable kinematic reasoning. In addition, our experiments focus primarily on lane-to-lane topology under offline perception metrics. Extending TOPOMIND to jointly model lane-to-traffic-control topology, including relations to traffic signals and signs, and integrating these structured predictions into a closed-loop autonomous-driving system would enable direct evaluation of how topology reasoning benefits downstream planning and driving behavior.