Geometry-Guided Representations for Coherent Lane and Traffic Topology Reasoning in Driving Scenes

Yueru Luo
222010057@link.cuhk.edu.cn
Shenzhen Future Network of Intelligence Institute
Shenzhen, Guangdong, China
School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong, China

Changqing Zhou
czhou149@connect.hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China

Yiming Yang
224010097@link.cuhk.edu.cn
Shenzhen Future Network of Intelligence Institute
Shenzhen, Guangdong, China
School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong, China

Erlong Li
Chao Zheng
erlongli@tencent.com
chrisczheng@tencent.com
Tencent Map, Tencent AI Lab
Beijing, China

Shuguang Cui
shuguangcui@cuhk.edu.cn
School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong, China
Shenzhen Future Network of Intelligence Institute
Shenzhen, Guangdong, China

Zhen Li ✉
lizhen@cuhk.edu.cn
School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong, China
Shenzhen Future Network of Intelligence Institute
Shenzhen, Guangdong, China

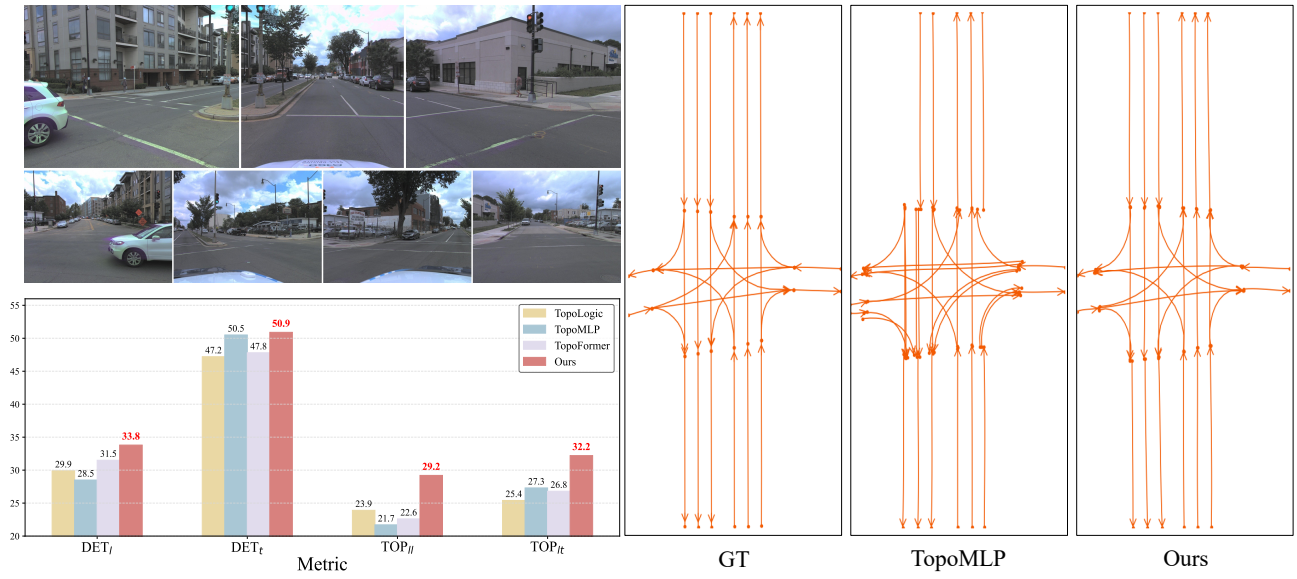

Figure 1: Performance and visualization comparison between previous methods and ours. The top-left shows multi-view input images. Our approach, CoPo , integrates relational modeling across multiple levels, strengthening both perception and topology reasoning. As shown in the bottom-left quantitative results on OpenLane-V2, CoPo significantly outperforms prior methods across all metrics. On the right, qualitative comparisons demonstrate that CoPo produces more accurate lane geometries and well-aligned connectivity than prior approaches.

Abstract

Road topology reasoning is fundamental for autonomous driving, requiring both accurate perception of road elements and understanding of their complex connectivity, including lane connectivity (Lane-to-Lane, L2L) and traffic regulation (Lane-to-Traffic signs, L2T). However, existing methods typically treat perception and topology reasoning as fragmented tasks, ignoring their potential for mutual enhancement. Crucially, while topology is inherently relational, prior works often overlook geometric relationships during feature extraction, relying instead on brittle post-processing or coordinate-based heuristics applied only at inference time. To bridge this gap, we propose **Copo** (**Co**herent Perception and To**po**logy), a unified framework that integrates geometry-guided relational modeling across three levels: **1)** Perception-level: We introduce a relation-aware lane detector that utilizes geometry-biased self-attention and curve-guided cross-attention to enrich lane representations with structural priors; **2)** Reasoning-level: We design relation-enhanced topology heads, including a geometry-enhanced L2L head and a cross-view L2T head, which effectively align features to infer connectivity; and **3)** Supervision-level: We implement a contrastive InfoNCE strategy to regularize relational embeddings, pulling connected pairs closer in the latent space. This coherent multi-level design enables end-to-end joint optimization of perception and reasoning. Extensive experiments on OpenLane-V2 demonstrate that CoPo significantly outperforms existing methods, achieving gains of **+3.1** in DET_l, **+5.3** in TOP_{ll}, **+4.9** in TOP_{lt}, and **+4.4** overall in OLS, setting a new state-of-the-art.

CCS Concepts

• **Computing methodologies** → **Hierarchical representations**; **Scene understanding**.

Keywords

Relational Reasoning, Traffic topology Inference, Geometry-Guided Representations

ACM Reference Format:

Yueru Luo, Changqing Zhou, Yiming Yang, Erlong Li, Chao Zheng, Shuguang Cui, and Zhen Li ✉. 2026. 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 (KDD 2026), August 9–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3817889

KDD Availability Link:

The source code and data of this paper has been made publicly available at https://doi.org/10.5281/zenodo.20455469.

1 Introduction

Understanding road topology [22] is fundamental for autonomous driving, as it provides structured knowledge of lanes and traffic elements for navigation and compliance with traffic rules. A complete topology model requires not only perceive road elements but also reason about how they connect: how lanes interconnect (Lane-to-Lane, L2L), and how traffic elements (*e.g.*, lights or signs) regulate specific lanes (Lane-to-Traffic-element, L2T). We follow the OpenLane-V2 benchmark [43], where **perception** involves detecting lanes in 3D space from multi-view (MV) images and recognizing traffic elements in the front-view (FV), while **reasoning** infers L2L connectivity among lanes and L2T associations between lanes and traffic elements within only the FV.

Current approaches [10, 23, 37, 46] typically organize topology reasoning into a sequential perception-then-reasoning pipeline. Despite recent advances, two major gaps remain. **First, fragmented task optimization:** methods often prioritize either perception or L2L reasoning, leaving L2T reasoning underexplored and seldom optimizing these components jointly. This fragmented treatment overlooks the opportunity for mutual enhancement between perception and reasoning, as the lack of end-to-end relational coupling prevents effective across-stage synergy, thereby constraining holistic optimization. Concretely, in **perception**, many works [23, 37, 46] adapt generic object detectors [31, 54] by converting bounding boxes into lane points or curves, but overlook intrinsic geometric relationships [19, 52] (*e.g.*, parallelism, continuity) that humans naturally exploit. TopoLogic [10] makes a partial attempt by encoding end-to-start distances as a topology prior via a GNN, yet this design is narrowly tailored to connectivity and relies on a separate graph module. A key question remains: *how can we inject structural cues directly into perception so that lane features become inherently relation-aware?* In **reasoning**, existing works rely heavily on coordinate encodings that are brittle to small endpoint shifts, making L2L prediction fragile. TopoLogic [10] partly mitigates this issue via geometric distance in post-processing, but our study (**??** *and Sec. C.1*) shows that performance degrades sharply when removing that post-processing, implying that the model fails to internalize relational modeling. L2T reasoning is even less explored: most methods naïvely fuse BEV lane features and FV traffic elements, ignoring cross-view misalignment. Some approaches [20] mitigates this issue by leveraging 2D lane features from an additional decoder, but adds computational overhead.

Second, weak relational modeling: Although topology reasoning inherently involves relation prediction, most works directly force the model relation prediction via supervision [20, 23, 34, 46] or post-processing (*e.g.*, distance-based scoring in TopoLogic [10]), rather than embedding relational cues into feature learning and connectivity prediction. As a result, structural priors (such as parallelism or proximity) between lanes and traffic elements are underutilized, leaving models brittle to geometric variation.

To address these gaps, we propose CoPo , a **multi-level relational modeling** approach that that systematically integrates relational modeling across three levels. This enables perception and reasoning to be optimized jointly and coherently, grounded by structural relationships inherent in road scenes: **1) Relational Perception:** we develop a relation-aware lane detector based on a compact Bézier-curve representation. A *Geometry-Biased Self-Attention* module encodes inter-lane affinities (distance, orientation) as attention biases, while a *Curve-Guided Cross-Attention* module

KDD 2026, Jeju Island, Republic of Korea.

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817889

aggregates contextual cues along curves, ensuring relational awareness even with sparse control points. **2) Relational Reasoning:** we design relation-enhanced topology heads: a *Geometry-Enhanced L2L head* that embeds inter-lane distances to produce robust connectivity predictions, and a *Cross-View L2T head* that aligns BEV lanes with FV traffic elements to overcome spatial discrepancy. **3) Relational Supervision:** we introduce a InfoNCE [14, 30, 48]-based contrastive objective that pulls connected pairs closer and separates non-connected ones in the embedding space, regularizing relational learning.

Overall, our contributions can be summarized as follows:

- We identify two core limitations in existing topology reasoning, namely fragmented task optimization and weak relational modeling, and highlight the need for a unified relational perspective.
- We propose CoPo , a multi-level relational modeling approach that integrates relational modeling through perception, reasoning, and supervision levels.
- Extensive experiments on the OpenLaneV2 dataset validate the effectiveness of our approach, demonstrating multi-level relation benefit mutually. With sufficient improvements, our method surpasses previous methods across both perception and reasoning, setting a new state-of-the-art.

2 Related Work

2.1 3D Lane Detection

3D lane detection provides the lane geometry for topology reasoning. Existing methods are typically designed for *monocular* front-view images and fall into BEV-based and front-view-based paradigms. BEV-based methods employ inverse perspective mapping (IPM) to transform FV images into BEV space for lane prediction [6, 9, 12, 13, 19, 29]. However, IPM-based methods suffer from distortions on non-flat roads due to their planar assumption, limiting its reliability in real-world settings. Front-view-based approaches instead predict 3D lanes directly from FV features, avoiding IPM distortions. Recent works [1, 17, 35] adopt query-based detectors [5, 54] to model 3D lanes information end-to-end, achieving stronger performance. Our work extends lane perception further by incorporating multi-view input and embedding relational cues directly into the perception process, enabling lane features to become inherently relation-aware.

2.2 Online HD Map Construction

Online HD map construction aims to dynamically generate structured map elements. Early methods [21] use dense segmentation with heuristic vectorization. VectorMapNet [32] advances this direction with an end-to-end detection-and-serialization pipeline. More recent works [7, 16, 26, 27, 39, 42, 49, 52, 53] adopt Transformer-based architectures with curve- or point-based representations. MapTR [26] introduces hierarchical query embeddings for polyline generation, BeMapNet [39] and PivotNet [7] employ piecewise Bézier curves and dynamic points, and InstaGraM [42] formulates map element generation as a graph problem with GNNs. GeMap [52] models structural and relational properties of map elements, but focuses mainly on individual geometries without explicitly capturing topology relationships among lanes, and it relies on polyline representations composed of equidistant points, which lack flexibility and precision [7, 51] for nuanced lane description. To improve computational efficiency, various decoupled self-attention mechanisms [16, 27, 52] have been proposed for integrating intra-/inter-instance information. However, topology reasoning among elements are limited in this area.

2.3 Driving Scene Topology Reasoning

Topology reasoning aims to capture the connectivity among road elements. Early works like STSU [3] and TPLR [4] construct lane graphs directly in BEV space, but focus only on L2L reasoning and ignore interactions with traffic elements. Recent methods [20, 23, 24, 37, 46] begin to address both L2L and L2T reasoning. TopoNet [23] employs a GNN-based framework that enhances topology prediction through message passing between element embeddings. TopoMLP [46] adopts MLP heads for more efficiency, and LaneSegNet [24] proposes lane segment attention to capture intra-lane dependencies [50]. However, these methods do not explicitly model relationships among lanes or between lanes and traffic elements, limiting their ability to capture structural dependencies. Topo2D [20] incorporates additional 2D lane detection to support topology reasoning. TopoLogic [10] combines geometry (end-to-start) distance-based topology (GDT) estimation with query similarity-based relational modeling via GNNs, but applies geometric cues primarily as post-processing for L2L inference. Subsequently, works like[11, 36] move GDT into feature encoding, yet they still measures lane geometry relationships via only end-to-start patterns and rely on extra GNNs.

In contrast, our work takes a unified view: we embed geometric relational modeling coherently across perception, reasoning, and supervision, encoding comprehensive geometric priors (distance, angular relationships, spatial proximity) directly into feature learning for coherent joint optimization of perception and topology.

3 Method

3.1 Overview

CoPo weaves relational modeling through perception, reasoning, and supervision. Given multi-view images, the model comprises two branches: one decodes BEV lane features, and another detects front-view traffic elements. These outputs feed into relational topology heads to infer L2L and L2T connections. We inject relation cues early (in the decoder, Sec. 3.2), maintain them in reasoning (in the heads, Sec. 3.3), and sharpen them through contrastive loss in supervision (Sec. 3.4.2). This multi-level integration enables coherent perception and topology reasoning, where geometry guides both feature learning and relational inference in a unified framework. The following subsections detail each module.

3.2 Relational Perception

Lanes in real driving scenes exhibit strong geometric relationships, such as parallelism and convergence, that naturally guide perception but are often neglected in prior works. We embed these geometric relational cues directly into the decoder so that lane features become structurally informed. Concretely, we introduce a **relation-aware decoder** that comprises two complementary mechanisms: 1) **geometry-biased self-attention** (Sec. 3.2.1) adds inter-lane geometric biases into the self-attention logits, steering each lane

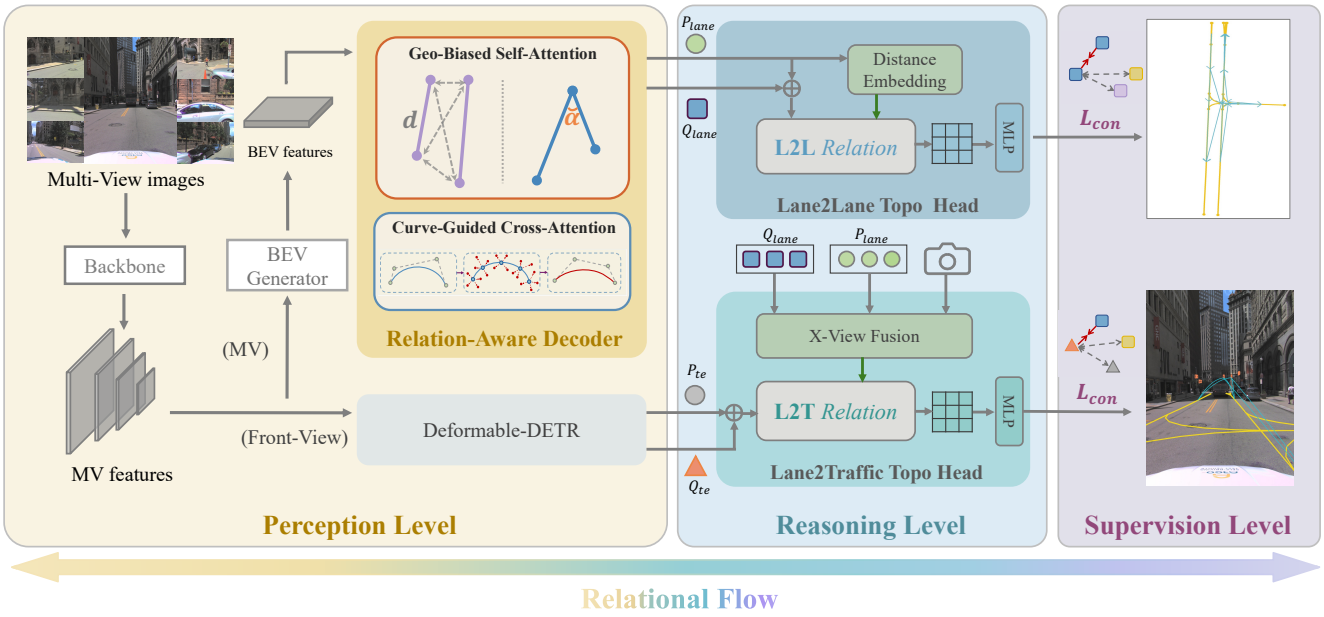

Figure 2: The overall framework of CoPo , Our framework coherently integrates geometry-guided relational modeling across three levels: Perception Level: A relation-aware lane decoder enrich lane features with geometric cues. Reasoning Level: Relation-enhanced topology heads infer topology between lanes (L2L) and between lanes and traffic elements (L2T) by embedding structural relations into relation embeddings, Supervision Level: Contrastive loss $\mathcal{L}_{con}$ regularizes relational embeddings, pulling connected pairs closer and pushing apart non-connected pairs in the feature space. ■ (Q_{lane}), ▲ (Q_{te}) denote lane and traffic element queries, and ● (P_{lane}) and ● (P_{te}) denotes their predictions.

to attend to its peers in a relation-aware fashion; 2) **curve-guided cross-attention** (Sec. 3.2.2) samples points along the Bézier curve representation of each lane and performs deformable attention over those points. This enables context aggregation along the lane's shape rather than just at control points. These two design choices ensure that lane queries evolve under structural guidance, improving robustness in downstream topology reasoning.

3.2.1 Geometry-Biased Self-Attention. Training queries in DETR-like decoders often suffer from slow convergence due to lack of structural inductive bias [15]. To address this, we encode pairwise lane geometry as an additional bias in self-attention. Given two lanes $\boldsymbol{l}_i$ and $\boldsymbol{l}_j$ with endpoints $\mathbf{p}_*^s, \mathbf{p}_*^e$, we define a symmetric *minimum endpoint distance*:

$$\mathrm{Dist}(\boldsymbol{l}_i, \boldsymbol{l}_j) = \min_{x\in\{s,e\},\, y\in\{s,e\}} \left\| \mathbf{p}_i^x - \mathbf{p}_j^y \right\|_2, \tag{1}$$

which measures the closest endpoint-to-endpoint proximity and is robust to endpoint ordering. We further compute an *orientation difference* $\mathrm{Angle}(\boldsymbol{l}_i, \boldsymbol{l}_j)$. We concatenate these cues, and embed to biases via a linear projection $\mathtt{GE}$ (after sinusoidal encoding):

$$\mathbf{Q} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{\text{model}}}} + \mathbf{Geometry}(\boldsymbol{l}, \boldsymbol{l})\right)\mathbf{V} \tag{2}$$

$$\mathbf{Geometry}(\boldsymbol{l}, \boldsymbol{l})_{(i,j)} = \mathtt{GE}\big(\mathrm{Dist}(\boldsymbol{l}_i, \boldsymbol{l}_j) \| \mathrm{Angle}(\boldsymbol{l}_i, \boldsymbol{l}_j)\big). \tag{3}$$

We illustrate this mechanism in □ of Fig. 3 (a), where the ***Geometry Bias*** represents the relation bias term $\mathbf{Geometry}(\boldsymbol{l}, \boldsymbol{l})_{(i,j)}$ between i-th and j-th lanes. This mechanism enhances query learning by introducing structural bias [15], while simultaneously capturing inherent geometric relationships between lanes. Unlike Topologic [10], which relies on GNNs to encode connectivity, our approach is simpler yet more effective: we directly encode geometric relationships as attention biases within self-attention, eliminating the need for separate graph propagation. Additionally, we incorporate angular information for comprehensive geometry relation modeling.

By enhancing geometrically proximal lanes, the model allocates greater focus to spatially relevant lanes through the interleaved attention process, leading to a more robust understanding of lane topology (detailed in Sec. 3.3.1). Our method differs from [15], which encodes progressive cross-layer box positional relationships for individual objects.

3.2.2 Curve-Guided Cross-Attention. Polyline representations with fixed spacing are inflexible [7, 51]. We instead adopt a compact cubic Bézier representation with four control points. However, two challenges arise: 1) the sparsity of control points limits feature aggregation; 2) intermediate control points may be misaligned with the underlying curve Fig. 3(b).

To address these issues, instead of relying solely on sparse control points as reference points in deformable attention as in TopoDBA [18], we sample K points along the Bézier curve, which serves as reference points for feature aggregation (Fig. 3 (b)). Furthermore, to capture long-range intra-lane dependencies, we employ a shared

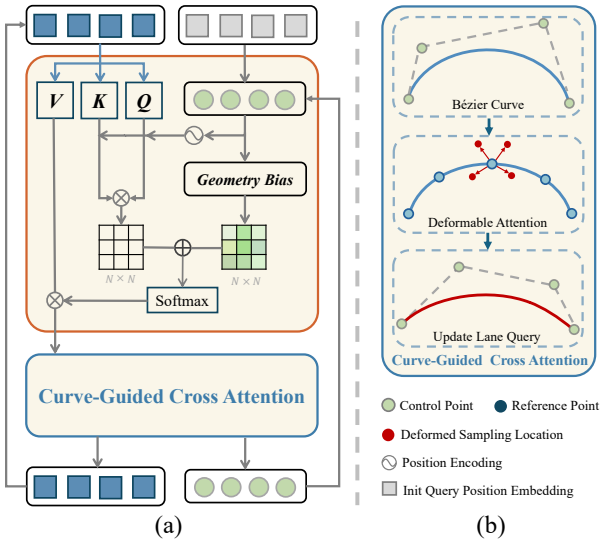

Figure 3: Illustration of our Bézier lane decoder layer, featuring our geometry-biased SA and curve-guided CA.

lane query to integrally generate offsets and weights for all sampled K points. This design enables contextual information flow holistically within each lane.

Given the feature map $\boldsymbol{x}$, the l^{th} lane query $\boldsymbol{q}_l$ and its reference point $\boldsymbol{p}_l$, we adopt deformable attention mechanism [54] to update query features, formulated as:

$$\text{DeformAttn}(\boldsymbol{q}_l, \boldsymbol{p}_l, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{i=1}^{N} \sum_{j=1}^{K} A_{mlij} \cdot \boldsymbol{W}_m' \boldsymbol{x}(\boldsymbol{p}_l + \Delta \boldsymbol{p}_{mlij}) \right], \tag{4}$$

where M is the number of attention heads, N denotes the number of offset locations per sampled point, and K is the number of sampled points along each curve. A_{mlij} and $\Delta \boldsymbol{p}_{mlij}$ denote the attention weights and sampling offset, respectively, for the i^{th} offset of the j^{th} sampled points along the curve in the m^{th} attention head.

Unlike BézierFormer [8], which projects sampled points onto image feature maps and performs `grid_sample`[1] to generate N_{ref} point queries, each undergoing separate deformable cross-attention, our approach is simpler and more effective: it avoids the inefficiency of per-point queries and leverages global lane structure to guide local sampling, enhancing the modeling of long-range intra-lane dependencies. Besides, unlike BeMapNet [39], which represents each map element with multiple Bézier curves, our single Bézier curve per lane balances efficiency and accuracy, as validated in Tab. 4.

3.3 Relational Reasoning

With relation-aware lane features in place, we infer connectivity in a unified relational space for both L2L and L2T. Our design departs from pipelines that either rely on brittle coordinate encodings [23, 46] or defer geometry to post-processing heuristics [10]. Instead, we embed relational priors *inside* the pairwise embeddings and align cross-view features end-to-end, thereby reducing sensitivity to endpoint shifts and eliminating reliance on external fixes.

3.3.1 Geometry-Enhanced L2L Reasoning. Existing methods [20, 37, 46] enhance lane features using coordinate-based encoding to predict L2L connectivity, which are, however, highly sensitive to endpoint shifts. Although [10] attempt to remedy this by incorporating endpoint geometric distance as extra topology features (GDT), but the post-hoc design of their GDT fails to internalize relational modeling in model learning (as we discussed in Sec. 1 and analysis in Sec. C.1). This leaves relational structure underutilized. Motivated by these limitations, we propose a geometry-enhanced L2L reasoning head that embeds geometric priors directly into pairwise embeddings. Concretely, given refined lane features $\mathbf{Q}_{\text{lane}} \in \mathbb{R}^{N \times C}$ and endpoints $\mathbf{P}_{\text{lane}} \in \mathbb{R}^{N \times 2 \times 2}$, we first project $\mathbf{Q}_{\text{lane}}$ into predecessor and successor embeddings with two MLPs (as L2L is directional, this choice is for capturing the asymmetric nature of L2L topology). Besides, we employ a sinusoidal positional encoding to produce lane positional embeddings based on $\mathbf{P}_{\text{lane}}$, yielding $\mathbf{PE}_{\text{lane}}$.

$$\mathbf{G}_{\text{L2L}} = \left(\text{MLP}_1(\mathbf{Q}_{\text{lane}}) + \mathbf{PE}_{\text{pre}}\right) \copyright \left(\text{MLP}_2(\mathbf{Q}_{\text{lane}}) + \mathbf{PE}_{\text{suc}}\right), \tag{5}$$

where © denotes broadcast concatenation and $\mathbf{G}_{\text{L2L}} \in \mathbb{R}^{N \times N \times C}$. To reinforce connectivity relationships and stabilize against endpoint noise, we introduce a distance embedding:

$$\text{DistEmbed}_{\text{L2L}}^{i,j} = \text{MLP}(\texttt{distance}(\mathbf{P}_i^e - \mathbf{P}_j^s)), \quad \text{DistEmbed} \in \mathbb{R}^{N \times N \times C} \tag{6}$$

where $\mathbf{P}_i^e$ and $\mathbf{P}_j^s$ denote the endpoint of lane i and the starting point of lane j, respectively. and then predict L2L topology via:

$$\mathbf{T}_{\text{L2L}} = \text{MLP}(\mathbf{G}_{\text{L2L}} + \text{DistEmbed}_{\text{L2L}}), \quad \mathbf{T}_{\text{L2L}} \in \mathbb{R}^{N \times N \times 1} \tag{7}$$

By knitting geometry into the trainable relational embedding, our design internalizes robustness to small endpoint shifts and supports directional reasoning, unlike external post-processing hacks or pure coordinate encodings. This strengthens connected-pair discrimination and reduces false negatives in challenging scenarios. More details are illustrated in *Appendix* algorithm 1.

Algorithm 1 Geometry-Enhanced L2L Reasoning

Require: Lane query features $\mathbf{Q}_{\text{lane}} \in \mathbb{R}^{N \times C}$, lane endpoints $\mathbf{P}_{\text{lane}} \in \mathbb{R}^{N \times 2 \times 2}$
Ensure: Connectivity logits $\mathbf{T}_{\text{L2L}} \in \mathbb{R}^{N \times N \times 1}$

$\mathbf{E}_{\text{pre}} \leftarrow \text{MLP}_1(\mathbf{Q}_{\text{lane}})$ ▷ $\in \mathbb{R}^{N \times \frac{C}{2}}$
$\mathbf{E}_{\text{suc}} \leftarrow \text{MLP}_2(\mathbf{Q}_{\text{lane}})$ ▷ $\in \mathbb{R}^{N \times \frac{C}{2}}$
$(\mathbf{PE}_{\text{pre}}, \mathbf{PE}_{\text{suc}}) \leftarrow \text{PosEnc}(P_{\text{lane}})$ ▷ sinusoidal positional encoding
$\tilde{\mathbf{E}}_{\text{pre}} \leftarrow \mathbf{E}_{\text{pre}} + \mathbf{PE}_{\text{pre}}$
$\tilde{\mathbf{E}}_{\text{suc}} \leftarrow \mathbf{E}_{\text{suc}} + \mathbf{PE}_{\text{suc}}$
$\mathbf{G}_{\text{L2L}} \leftarrow \text{BroadcastConcat}(\tilde{\mathbf{E}}_{\text{pre}}, \tilde{\mathbf{E}}_{\text{suc}})$ ▷ pairwise feature, $\in \mathbb{R}^{N \times N \times C}$
$\text{Dist}_{ij} \leftarrow \texttt{distance}(P_i^e - P_j^s)$ ▷ distance: lane i end point → lane j start point, $\in \mathbb{R}^{N \times N \times 1}$
$\text{DistEmbed} \leftarrow \text{MLP}(\text{Dist}_{ij})$ ▷ $\in \mathbb{R}^{N \times N \times C}$
$\mathbf{T}_{\text{L2L}} \leftarrow \text{MLP}(\mathbf{G}_{\text{L2L}} + \text{DistEmbed})$ ▷ $\in \mathbb{R}^{N \times N \times 1}$
return $\mathbf{T}_{\text{L2L}}$

[1]This denotes the grid sampling operation defined in PyTorch.

3.3.2 X-View L2T Reasoning. L2T reasoning requires bridging two inherently different representations: BEV features for lanes and FV features for traffic elements. The disparity between these representations poses a challenge for learning spatially consistent relationships. Many prior efforts treat these features as if they lie in the same space and naïvely fuse or concatenate them [37, 46]. Topo2D [20] mitigates this by injecting 2D priors via auxiliary decoders.

In contrast, we introduce a X-view fusion module that aligns BEV and FV features more directly and cohesively. First, we project predicted 3D lane coordinates $\mathbf{P}^{3D}_{\text{lane}} \in \mathbb{R}^{N\times K\times 3}$ into the front-view plane to obtain $\mathbf{P}^{2D}_{\text{lane}} \in \mathbb{R}^{N\times K\times 2}$. We then sample spatial features $\mathbf{F}^{2D}_{\text{lane}}$ at those locations via `grid_sample`. These sampled features serve as a bridge between BEV and FV spaces, aligned by 3D geometry. Next, we fuse $\mathbf{F}^{2D}_{\text{lane}}$ into lane queries along with their 2D positional embeddings (using similar sinusoidal encoding strategy as in L2L module), producing enhanced queries:

$$\tilde{\mathbf{Q}}_{\text{lane}} = \text{MLP}_1(\mathbf{Q}_{\text{lane}}) + \mathbf{F}^{2D}_{\text{lane}} + \mathbf{PE}^{2D}_{\text{lane}},\ \tilde{\mathbf{Q}}_{\text{te}} = \text{MLP}_2(\mathbf{Q}_{\text{te}}) + \mathbf{PE}^{2D}_{\text{te}}. \quad (8)$$

Here the positional embeddings encode 2D spatial layout in the front-view, ensuring the fused features remain spatially coherent. We then form the relational embedding $\mathbf{G}_{\text{L2T}} \in \mathbb{R}^{N\times M\times C}$ by broadcast concatenation of $\tilde{\mathbf{Q}}_{\text{lane}}$ and $\tilde{\mathbf{Q}}_{\text{te}}$, and feed it into an MLP to predict the L2T topology:

$$\mathbf{T}_{\text{L2T}} = \text{MLP}(\mathbf{G}_{\text{L2T}}), \quad \mathbf{T}_{\text{L2T}} \in \mathbb{R}^{N\times M\times 1}. \quad (9)$$

By aligning BEV lane features to FV contexts via projected geometry and position embeddings, we reduce misalignment between two space representation, enabling more robust lane–traffic associations. Empirically, this design lowers erroneous associations in setups where naïve fusion would fail, especially at larger distances or oblique angles (shown in the qualitative comparison in Sec. B).

3.4 Loss Functions

3.4.1 Perception Loss. For lane detection, we employ Focal Loss [28] for classification and a combination of point-wise L1 loss and Chamfer distance loss over K sampling points to guide accurate geometric regression. For traffic element detection, we again use Focal Loss [28] for classification, along with L1 loss and GIoU loss [41] to supervise bounding box regression over (x, y, w, h).

3.4.2 Relational Supervision. Most existing methods rely solely on Focal Loss to determine connectivity over pairs. However, this approach primarily emphasizes binary classification without explicitly distinguishing the relative importance of connected and non-connected pairs. To better capture topological relationships, we augment this with a **contrastive** InfoNCE loss [38] over pairwise relation embeddings.

Concretely, consider the L2T case (the same logic applies to L2L). Let N be the number of lane queries and M the number of traffic element queries. The relational head produces an embedding tensor $\mathbf{G}_{\text{L2T}} \in \mathbb{R}^{N\times M\times C}$ and outputs a logit tensor $\mathbf{T}_{\text{L2T}}$. We treat pairs labeled as "connected" as positives, and all others as negatives. To sharpen discrimination, we introduce a **hard negative mining** strategy: for each positive pair, we select the top-n hardest negative pairs according to predicted logits, focusing the contrastive signal where confusion is highest. We then apply a symmetric InfoNCE loss (in the spirit of CLIP [40]-style contrastive formulations) over these logits. Let $\mathbf{v}^+$ and $\{\mathbf{v}^-\}$ denote the positive and selected negative logits, then:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\mathbf{v}^+)}{\exp(\mathbf{v}^+) + \sum_{\mathbf{v}^-} \exp(\mathbf{v}^-)} = \log\left[1 + \sum_{\mathbf{v}^-} \exp(\mathbf{v}^- - \mathbf{v}^+)\right], \quad (10)$$

To handle multiple positives for a given anchor, we extend Eq. (10), following previous works [44, 47], to:

$$\mathcal{L}_{\text{con}} = \log\left[1 + \sum_{\mathbf{v}^+}\sum_{\mathbf{v}^-} \exp(\mathbf{v}^- - \mathbf{v}^+)\right]. \quad (11)$$

4 Experimental Results

4.1 Dataset and Metrics

Dataset. We evaluate our method on OpenLane-V2 [43], a large-scale dataset specifically designed for topology reasoning in autonomous driving comprising two subsets: `subsetA` (from Argoverse-V2 [45]) and `subsetB` (from nuScenes [2]).

Evaluation Metrics. Following the official evaluation of OpenLane-V2 [43], we utilize DET_l and DET_t to measure detection accuracy for lanes and traffic elements, respectively. For topology reasoning, we employ TOP_{ll} and TOP_{lt} to assess L2L and L2T relationship prediction. The overall performance is quantified using the OpenLane-V2 Score (OLS):

$$\text{OLS} = \frac{1}{4}\left[\text{DET}_l + \text{DET}_t + f(\text{TOP}_{ll}) + f(\text{TOP}_{lt})\right], \quad (12)$$

where f denotes the square root function. Our evaluations follow the latest version (`V2.1.0`) of the metrics, as updated in the official OpenLane-V2 GitHub repository[2].

4.2 Implementation Details

Model Details. We use a ResNet-50 backbone to extract features, coupled with a pyramid network, FPN, for multi-scale feature learning. Following prior work [10, 23], a BEVFormer encoder [25] with 3 layers is employed to generate a BEV feature map of size 100×200. We employ six decoder layers, using 300 queries for the lane decoder and 100 queries for the traffic element decoder following [46].

Training Details. We utilize the AdamW optimizer [33] for model training, with a weight decay of 0.01 and an initial learning rate of 2.0×10^{-4}, which decays following a cosine annealing schedule. Training is conducted for 24 epochs using a total batch size of 8 on 8 NVIDIA 4090 GPUs. Input images are resized to 1024×800, following [46]. Due to space limitations, the training loss is provided in the *Appendix* Sec. A.1.

4.3 Main Results

We compare our model against SOTA methods on OpenLane-V2, with results shown in Tab. 1. On `subsetA`, CoPo achieves an OSL of 48.9, surpassing all previous methods by a significant margin (+4.4). The gains are consistent gains across metrics, with particularly strong improvements in lane detection (+3.1 DET_l) and topology (+5.3 TOP_{ll}, +4.4 TOP_{lt}). On `subsetB`, CoPo further improves to

[2] https://github.com/OpenDriveLab/OpenLane-V2/issues/76

Table 1: Performance comparison with existing methods on the OpenLane-V2 `subsetA` and `subsetB` dataset under the latest V2.1.0 evaluation. Results for RoadPainter‡ derive from their paper which uses old metrics (due to the absence of open-source code or model). TopoMLP† results were obtained with their provided model, and TopoFormer§ denotes our reimplemented [36] results, due to their ×2 input size, another fine-tuning stage and lack of runnable code. Metrics include detection accuracy (DET_l and DET_t) and topology reasoning accuracy (TOP_{ll} and TOP_{lt}), with overall score (OLS) indicating aggregated performance.

Subset	Method	Venue	Backbone	Epoch	DET_l ↑	DET_t ↑	TOP_{ll} ↑	TOP_{lt} ↑	OLS ↑
`setA`	VectorMapNet	ICML2023	ResNet-50	24	11.1	41.7	2.7	9.2	24.9
	MapTR	ICLR2023	ResNet-50	24	17.7	43.5	5.9	15.1	31.0
	TopoNet	Arxiv2023	ResNet-50	24	28.6	48.6	10.9	23.8	39.8
	TopoMLP†	ICLR2024	ResNet-50	24	28.5	50.5	21.7	27.3	44.5
	RoadPainter‡	ECCV2024	ResNet-50	24	30.7	47.7	7.9	24.3	38.9
	TopoLogic	NeurIPS2024	ResNet-50	24	29.9	47.2	23.9	25.4	44.1
	TopoFormer§	CVPR2025	ResNet-50	24	31.5	47.8	22.6	26.8	44.7
	Ours	-	ResNet-50	24	**33.8**	**50.9**	**29.2**	**32.2**	**48.9**
	Improvement	-	-	-	3.1 ↑	0.4 ↑	5.3 ↑	4.9 ↑	4.4 ↑
`setB`	TopoNet	Arxiv2023	ResNet-50	24	24.3	55.0	6.7	16.7	36.8
	TopoMLP†	ICLR2024	ResNet-50	24	26.0	58.2	21.0	19.8	43.6
	RoadPainter‡	ECCV2024	ResNet-50	24	28.7	54.8	8.5	17.2	38.5
	TopoLogic	NeurIPS2024	ResNet-50	24	25.9	54.7	21.6	17.9	42.3
	TopoFormer§	CVPR2025	ResNet-50	24	29.2	55.1	20.9	19.7	43.6
	Ours	-	ResNet-50	24	**32.6**	**58.8**	**31.8**	**25.8**	**49.7**
	Improvement	-	-	-	3.4 ↑	0.6 ↑	10.2 ↑	6.0 ↑	6.1 ↑

Table 2: Ablation study on our key components: Relational Perception (geometry-biased SA and curve-guided CA), Relational Reasoning, and Relational Supervision. Colored rows mark the milestones upon equipping the respective modules.

	Rel. Perception		Rel. Reasoning		Rel. Sup.	Metrics					Efficiency & Performance			
#L	SA	CA	L2L	L2T	NCE	DET_l	DET_t	TOP_{ll}	TOP_{lt}	OLS	Size (MB)	Lat. (ms)	Δ Cost	Δ OLS
#1						27.7	50.1	21.8	28.4	44.5	218.7	159.7	–	–
#2	✓					28.1	48.8	24.7	28.3	44.9	218.7	160.6	+0.6%	+0.4
#3	✓	✓				32.7	50.1	26.8	29.9	47.3	223.2	161.4	+1.1%	+2.8
#4	✓	✓	✓			33.7	48.0	28.5	29.7	47.4	224.2	161.8	+1.3%	+2.9
#5	✓	✓		✓		33.5	48.4	26.4	31.4	47.4	224.2	161.7	+1.3%	+2.9
#6	✓	✓	✓	✓		33.9	50.3	28.6	30.6	48.2	224.8	162.2	+1.6%	+3.7
#7	✓	✓	✓	✓	✓	**33.8**	**50.9**	**29.2**	**32.2**	**48.9**	224.8	162.2	+1.6%	+4.4

49.7 OLS (+6.1), driven by large improvements in topology reasoning (+10.2 TOP_{ll} and +6.0 TOP_{lt}), while maintaining consistent improvements in detection metrics. These results confirm that our relational modeling at perception, reasoning, and supervision levels systematically strengthens both perception and topology reasoning. Although our model adopts the same traffic decoder as prior work [46], it still achieves +0.4/+0.6 gains in DET_t on `setA`/`setB`, suggesting that relational cues learned for L2T reasoning also benefit traffic element perception.

Overall, these results highlight the superiority of CoPo , which sets a new state-of-the-art on both OpenLane-V2 `SubsetA` and `SubsetB`. To further illustrate its effectiveness, we present **qualitative results** in Figs. 1 and 4, where our model produces more accurate lane predictions, better-aligned connection points and more accurate topology estimation in complex driving scenes. More detailed visualizations are provided in *Appendix* Sec. B.

4.4 Ablation Studies

To thoroughly evaluate our core designs, we conduct ablation studies on OpenLane-V2 `subsetA`. Our baseline model (#1) adopts a deformable-DETR decoder following [46] with lightweight MLP-based topology heads. Results are shown in Tab. 2.

Additional analyses are provided in our *Appendix*, covering: comparison of our geometry-biased `SA` and L2L relational modeling against previous methods [10, 11, 36] (Sec. C.1). We also discuss the limitations of our method and potential directions for future work.

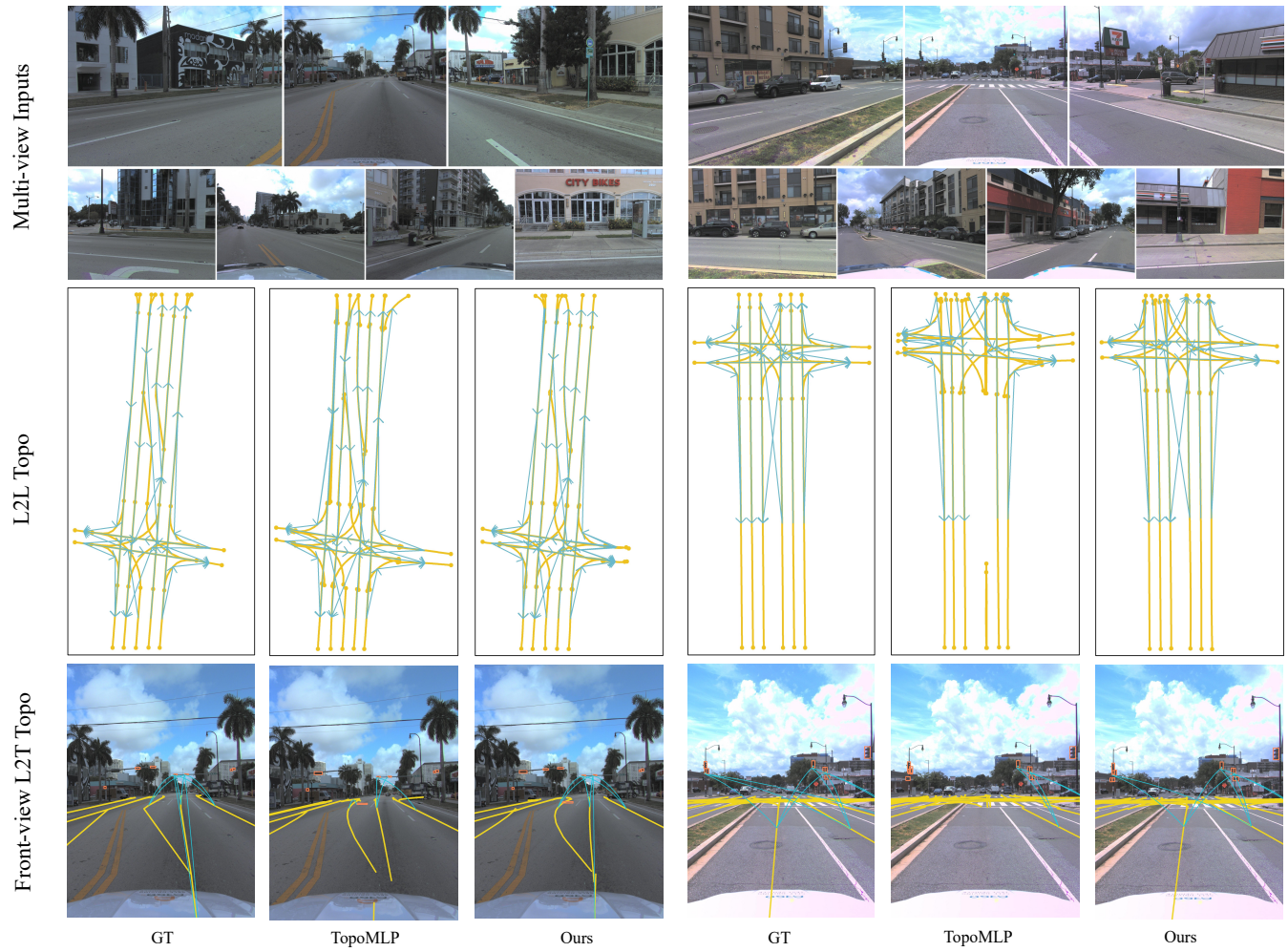

Figure 4: Qualitative results comparison. The 1st row presents the multi-view input images, the 2rd row illustrates the predicted L2L topology in BEV, where light blue arrowlines indicate directed L2L connectivity estimation (detailed lane centerline predictions are provided in the supplementary material). The final row depicts the front-view L2T topology predictions, where orange boxes highlight detected traffic elements (*e.g.*, traffic lights and signs), yellow lines denote projected lane centerlines, and blue lines represent lane-to-traffic element (L2T) associations.

Effect of Relational Perception. Our relation-aware lane decoder includes Geometry-Biased Self-Attention (SA) and Curve-Guided Cross-Attention (CA). Replacing standard self-attention with SA (#2) improves TOP_{ll} by +2.9, showing the benefit of encoding inter-lane geometry. Besides, we compare our geometry encoding with the distance topology method from Topologic in *Appendix* Tab. 3, further validating the advantages of our method. Introducing CA (#3) further boosts DET_l by +4.6, as curve-guided sampling better aggregates long-range features. Together, SA and CA (#3) deliver substantial gains over baseline (#1): +5.0 DET_l, +5.0 TOP_{ll}, +1.5 TOP_{lt}, and +2.8 OLS. These results confirm the combined benefits within our relational perception.

Effect of Relational Reasoning. We next evaluate our relational topology heads. Adding our geometry-enhanced L2L head (#3→#4) improves TOP_{ll} by +1.7 and also raises DET_t by +1.0, validating its effectiveness in capturing L2L relationships and showing that stronger lane connectivity reasoning indirectly benefits perception. Replacing the baseline L2T head with our cross-view design (#3→#5) improves TOP_{lt} by +1.5, demonstrating better lane–traffic alignment. Integrating both heads (#6) yields complementary gains (+1.8 TOP_{ll}, +0.7 TOP_{lt}, +0.9 OLS), confirming that relational reasoning at both L2L and L2T levels jointly strengthens performance.
Effect of Relational Supervision: Finally, we incorporate our contrastive loss for additional supervision in topology learning.

Compared to #6, adding our contrastive regularization (#7) provides further gains (+0.6 TOP_{ll}, +1.6 TOP_{lt}), validating that relational supervision helps structure the embedding space for more discriminative topology learning.

Efficiency and Performance. To quantify the overhead of our relational modules, we report model size and inference latency for all ablation variants in Tab. 2 on a single RTX 4090 GPU, as shown in the "Efficiency & Performance" part.Equipping our relational perception and relational reasoning increases latency only slightly, from 159.7 ms to 162.2 ms (about ≈1.6% over the baseline), while yielding a substantial OLS gain of +4.4 points (#1 → #7). Note that our relational supervision is used only during training (*i.e.*, #6 → #7 introduces no extra inference overhead). Thanks to the lightweight and effective design of our multi-level relational modeling, these gains come at negligible computational cost.

Table 3: Comparison with Geometric Distance Topology (GDT). "– GDT_{L2L}": removes GDT from L2L inference. "†": inference using the hyperparameters learned from training. "+ GDT_{lane}": replaces our geometry encoding with GDT for lane representation. "+ GDT_{L2L}": ensembles our L2L predictions with GDT at inference.

Model	DET_l	DET_t	TOP_{ll}	TOP_{lt}	OLS
Topologic	29.9	47.2	23.9	25.4	44.1
Topologic – GDT_{L2L}	29.9	47.2	11.6	25.4	40.4
Topologic†	29.9	47.2	23.2	25.4	43.9
Ours	**33.8**	**50.9**	**29.2**	**32.2**	**48.9**
Ours + GDT_{lane}	32.8	50.0	28.4	30.8	47.9
Ours + GDT_{L2L}	33.8	50.9	28.7	32.2	48.7

Comparison with Geometric Distance Topology (GDT). TopoLogic [10] proposes a geometric distance topology (GDT)approach that shares certain similarities with our method, but differs fundamentally in design and effectiveness. **First**, for lane representation learning, GDT-based methods [10, 11, 36] only captures end-to-start point distances between lanes, while our geometry-biased self-attention encodes richer spatial cues including inter-lane distances and angular relations (*e.g.*, parallelism, perpendicularity), which are common in real-world road layouts. **Second**, for L2L topology reasoning, TopoLogic applies GDT as a post-hoc refinement **only** at inference time with manually tuned hyperparameters, whereas our method integrates geometric cues directly into L2L relation features during end-to-end training. As shown in Table 3, when TopoLogic uses the GDT hyperparameters learned during training, its performance drops. Further replacing our geometry encoding with GDT leads to performance drops (DET_l: 33.8→32.8, TOP_{ll}: 29.2→28.4), and applying GDT to our L2L inference also degrades TOP_{ll} (29.2→28.7), highlighting the superiority of our learnable geometric modeling approach. More detailed analysis can be found in our appendix material.

4.5 Comparsion of Bézier representation.

We compare our Bézier-based lane representation with other alternative formulations in Tab. 4, BézierFormer [8] uses `grid_sample` to sample features and separately predicts offsets and attention weights for each point within deformable attention, which increases computational overhead and compromises instance-level consistency in lane prediction. In contrast, our method predicts offsets and attention weights jointly for the entire lane instance using a single lane query, resulting in improved efficiency and coherence. BeMapNet [39] employs a piecewise Bézier representation by dynamically predicting the number of segments, which increases model complexity and uncertainty. In the task of lane topology reasoning, such dynamic segmentation does not yield performance gains and lacks the structural regularity that our fixed Bézier formulation provides.

Table 4: Comparsion with different bézier representation.

Lane Rep.	DET_l	DET_t	TOP_{ll}	TOP_{lt}	OLS
BeMapNet	29.9	50.0	25.9	28.9	46.2
BézierFormer	31.0	49.7	26.1	29.6	46.5
Ours	**33.8**	**50.9**	**29.2**	**32.2**	**48.9**

5 Conclusion

We present CoPo , a unified framework that coherently integrates geometric relational modeling across perception, reasoning, and supervision for driving scene topology reasoning. By embedding comprehensive geometric priors into relation-aware lane features and topology heads, CoPo enables end-to-end joint optimization where perception and reasoning mutually reinforce each other. Extensive experiments on OpenLane-V2 demonstrate significant improvements, establishing new state-of-the-art performance. We believe CoPo demonstrates the effectiveness of coherent relational modeling for driving scene perception.

Acknowledgments

This work was supported in part by Guangdong S&T Programme with Grant No. 2024B0101030002, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, the Shenzhen Outstanding Talents Training Fund 202002, the NSFC with Grant No. 62293482, by NSFC with Grant No. 62573371, by the Guangdong Province Radio Science Data Center with grant No. 2025B1212070001, by the Shenzhen General Program No. JCYJ20220530143600001, by the Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of BigData Computing CHUK-Shenzhen, by the NSFC 61931024 & 12326610, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. SYSPG20241211173853027), by National Key Research and Development Program of China: 2025YFF0515300 and 2025YFF0515304, the Open Project Program (Grant No. QHSFCS-2606) of Key Laboratory of Tibetan Information Processing, Ministry of Education, by the Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, and by Tencent & Huawei Open Fund 2024E0009&202301030019.

References

[1] Yifeng Bai, Zhirong Chen, Zhangjie Fu, Lang Peng, Pengpeng Liang, and Erkang Cheng. 2023. Curveformer: 3d lane detection by curve propagation with curve queries and attention. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7062–7068.

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

[3] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. 2021. Structured bird's-eye-view traffic scene understanding from onboard images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15661–15670.

[4] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. 2022. Topology preserving local road network estimation from single onboard camera image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17263–17272.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 213–229.

[6] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. 2022. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision*. Springer, 550–567.

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

[8] Zhiwei Dong, Xi Zhu, Xiya Cao, Ran Ding, Caifa Zhou, Wei Li, Yongliang Wang, and Qiangbo Liu. 2024. BézierFormer: A Unified Architecture for 2D and 3D Lane Detection. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[9] Netalee Efrat, Max Bluvstein, Shaul Oron, Dan Levi, Noa Garnett, and Bat El Shlomo. 2020. 3d-lanenet+: Anchor free lane detection using a semi-local representation. *arXiv preprint arXiv:2011.01535* (2020).

[10] Yanping Fu, Wenbin Liao, Xinyuan Liu, Yike Ma, Feng Dai, Yucheng Zhang, et al. 2024. TopoLogic: An Interpretable Pipeline for Lane Topology Reasoning on Driving Scenes. *arXiv preprint arXiv:2405.14747* (2024).

[11] Yanping Fu, Xinyuan Liu, Tianyu Li, Yike Ma, Yucheng Zhang, and Feng Dai. 2025. TopoPoint: Enhance Topology Reasoning via Endpoint Detection in Autonomous Driving. *arXiv preprint arXiv:2505.17771* (2025).

[12] Noa Garnett, Rafi Cohen, Tomer Pe'er, Roee Lahav, and Dan Levi. 2019. 3d-lanenet: end-to-end 3d multiple lane detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2921–2930.

[13] Yuliang Guo, Guang Chen, Peitao Zhao, Weide Zhang, Jinghao Miao, Jingao Wang, and Tae Eun Choe. 2020. Gen-lanenet: A generalized and scalable approach for 3d lane detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 666–681.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[15] Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, Badong Chen, and Xuguang Lan. 2025. Relation detr: Exploring explicit position relation prior for object detection. In *European Conference on Computer Vision*. Springer, 89–105.

[16] Haotian Hu, Fanyi Wang, Yaonong Wang, Laifeng Hu, Jingwei Xu, and Zhiwang Zhang. 2024. ADMap: Anti-disturbance framework for reconstructing online vectorized HD map. *arXiv preprint arXiv:2401.13172* (2024).

[17] Shaofei Huang, Zhenwei Shen, Zehao Huang, Zi-han Ding, Jiao Dai, Jizhong Han, Naiyan Wang, and Si Liu. 2023. Anchor3dlane: Learning to regress 3d anchors for monocular 3d lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17451–17460.

[18] Muhammet Esat Kalfaoglu, Halil Ibrahim Ozturk, Ozsel Kilinc, and Alptekin Temizel. 2024. TopoBDA: Towards Bezier Deformable Attention for Road Topology Understanding. *arXiv preprint arXiv:2412.18951* (2024).

[19] Chenguang Li, Jia Shi, Ya Wang, and Guangliang Cheng. 2022. Reconstruct from top view: A 3d lane detection approach based on geometry structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4370–4379.

[20] Han Li, Zehao Huang, Zitian Wang, Wenge Rong, Naiyan Wang, and Si Liu. 2024. Enhancing 3D Lane Detection and Topology Reasoning with 2D Lane Priors. *arXiv preprint arXiv:2406.03105* (2024).

[21] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. 2022. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 4628–4634.

[22] Shuhao Li, Weidong Yang, Yue Cui, Xiaoxing Liu, Lingkai Meng, Lipeng Ma, and Fan Zhang. 2025. Fine-Grained Traffic Inference from Road to Lane via Spatio-Temporal Graph Node Generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 1529–1540.

[23] Tianyu Li, Li Chen, Huijie Wang, Yang Li, Jiazhi Yang, Xiangwei Geng, Shengyin Jiang, Yuting Wang, Hang Xu, Chunjing Xu, et al. 2023. Graph-based topology reasoning for driving scenes. *arXiv preprint arXiv:2304.05277* (2023).

[24] Tianyu Li, Peijin Jia, Bangjun Wang, Li Chen, Kun Jiang, Junchi Yan, and Hongyang Li. 2023. Lanesegnet: Map learning with lane segment perception for autonomous driving. *arXiv preprint arXiv:2312.16108* (2023).

[25] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*. Springer, 1–18.

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

[27] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. 2023. Maptrv2: An end-to-end framework for online vectorized hd map construction. *arXiv preprint arXiv:2308.05736* (2023).

[28] T Lin. 2017. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:1708.02002* (2017).

[29] Ruijin Liu, Dapeng Chen, Tie Liu, Zhiliang Xiong, and Zejian Yuan. 2022. Learning to predict 3d lane shape and camera pose from a single image via geometry constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1765–1772.

[30] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering* 35, 1 (2021), 857–876.

[31] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. 2022. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*. Springer, 531–548.

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

[33] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[34] Jiachen Lu, Renyuan Peng, Xinyue Cai, Hang Xu, Hongyang Li, Feng Wen, Wei Zhang, and Li Zhang. 2023. Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23–33.

[35] Yueru Luo, Chaoda Zheng, Xu Yan, Tang Kun, Chao Zheng, Shuguang Cui, and Zhen Li. 2023. Latr: 3d lane detection from monocular images with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7941–7952.

[36] Changsheng Lv, Mengshi Qi, Liang Liu, and Huadong Ma. 2025. T2sg: Traffic topology scene graph for topology reasoning in autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 17197–17206.

[37] Zhongxing Ma, Shuang Liang, Yongkun Wen, Weixin Lu, and Guowei Wan. 2024. RoadPainter: Points Are Ideal Navigators for Topology transformER. *arXiv preprint arXiv:2407.15349* (2024).

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[39] Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. 2023. End-to-end vectorized hd-map construction with piecewise bezier curve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13218–13228.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[41] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 658–666.

[42] Juyeb Shin, Francois Rameau, Hyeonjun Jeong, and Dongsuk Kum. 2023. Instagram: Instance-level graph modeling for vectorized hd map learning. *arXiv preprint arXiv:2301.04470* (2023).

[43] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, et al. 2024. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. *Advances in Neural Information Processing Systems* 36 (2024).

[44] Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144* (2023).

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

[46] Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. 2023. TopoMLP: An Simple yet Strong Pipeline for Driving Topology Reasoning. *arXiv preprint arXiv:2310.06753* (2023).

[47] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. 2022. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*. Springer, 588–605.
[48] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3733–3742.
[49] Jingyi Yu, Zizhao Zhang, Shengfu Xia, and Jizhang Sang. 2023. ScalableMap: Scalable Map Learning for Online Long-Range Vectorized HD Map Construction. In *Conference on Robot Learning*. PMLR, 2429–2443.
[50] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. 2024. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 7356–7365.
[51] Gongjie Zhang, Jiahao Lin, Shuang Wu, Yilin Song, Zhipeng Luo, Yang Xue, Shijian Lu, and Zuoguan Wang. 2023. Online Map Vectorization for Autonomous Driving: A Rasterization Perspective. *Advances in Neural Information Processing Systems*.
[52] Zhixin Zhang, Yiyuan Zhang, Xiaohan Ding, Fusheng Jin, and Xiangyu Yue. 2023. Online vectorized hd map construction using geometry. *arXiv preprint arXiv:2312.03341* (2023).
[53] Yi Zhou, Hui Zhang, Jiaqian Yu, Yifan Yang, Sangil Jung, Seung-In Park, and ByungIn Yoo. 2024. HIMap: HybrId Representation Learning for End-to-end Vectorized HD Map Construction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15396–15406.
[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).

A Implementation Details

A.1 Overall Optimization Objective

As discussed in the main paper, our training objective comprises detection losses for individual elements and reasoning losses for their relationships. Following prior methods [23, 46, 54], traffic element detection is supervised using a combination of classification loss and regression loss. Lane detection is similarly optimized with classification and regression losses. The regression loss includes L1 loss on Bézier control points and a chamfer distance loss $L_{\text{BézierCD}}$ computed on sampled on-curve points (introduced in our curve-guided cross-attention). We sample $K = 11$ points along each Bézier curve. Detection loss weights follow previous works [46]. For topology reasoning, we adopt focal loss for relation classification and introduce a contrastive loss $\mathcal{L}_{\text{con}}$ to enhance relational learning in the embedding space. Each contrastive pair includes one positive and three negative samples. The focal and contrastive losses are weighted by $\lambda_1 = 5$ and $\lambda_2 = 0.1$, respectively. The complete loss function is defined as:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{relation}}, \tag{13}$$

$$\mathcal{L}_{\text{relation}} = \lambda_1 \mathcal{L}^{\text{topo}}_{\text{class}} + \lambda_2 \mathcal{L}_{\text{con}}. \tag{14}$$

B More Qualitative Results

We present additional qualitative results in Fig. 5 (regarding the lane centerline predictions) and **??** (covering predictions: lane centerlines, L2L and L2T topology relationship estimation). Leveraging our proposed components, CoPo demonstrates superior perception of lane centerlines compared to previous methods, as shown in Fig. 5 and the second row of the figures. Moreover, CoPo achieves more accurate lane-to-lane (L2L) and lane-to-traffic-element (L2T) topology reasoning, as illustrated in the last two rows. Notably, CoPo exhibits significantly improved lane perception accuracy and topology reasoning performance in complex scenarios involving intricate L2T relationships.

C Extra Experiments

C.1 More Analysis on Geometry Topology.

As discussed in the main paper, TopoLogic [10] proposes a geometric distance topology (GDT), which shares certain similarities with our Geometry-Biased Self-Attention and Geometry-Enhanced L2L Topology, but *differs fundamentally in design and effectiveness*. Below we provided more detailed analysis.

Table 5: Experiments on Geometric Distance Topology (GDT). "Topologic – GDT_{L2L}": Inference without GDT_{L2L}. TopoLogic†: Inference using the hyperparameters learned during their lane representation training. "Ours + GDT_{lane}": Replaces our geometry encoding with TopoLogic's GDT for lane representation learning. "Ours + GDT_{L2L}": Ensembles our L2L predictions with GDT, same as TopoLogic.

Model	DET_l	DET_t	TOP_{ll}	TOP_{lt}	OLS
Topologic	29.9	47.2	23.9	25.4	44.1
Topologic – GDT_{L2L}	29.9	47.2	11.6	25.4	40.4
Topologic†	29.9	47.2	23.2	25.4	43.9
Ours	**33.8**	**50.9**	**29.2**	**32.2**	**48.9**
Ours + GDT_{lane}	32.8	50.0	28.4	30.8	47.9
Ours + GDT_{L2L}	33.8	50.9	28.7	32.2	48.7

Lane Representation Learning. *GDT [10, 36] focuses exclusively on connectivity estimation by computing the end-to-start point distance between lanes and updating representations using an additional GCN block. In contrast, our method captures richer spatial cues beyond connectivity, such as inter-lane distances and angular relations. These geometric priors—*e.g.*, parallelism, perpendicularity, and merging—are common in real-world road layouts and intuitively leveraged by humans. We explicitly encode the shortest endpoint distance and angular difference between lanes into high-dimensional relational features, which are further projected as biases in self-attention to enhance both perception and downstream topology reasoning.*

L2L Topology Reasoning. *Our method leverages the observation that annotated connected lanes share overlapping endpoints. We encode end-to-start distances as high-dimensional embeddings and integrate them directly into L2L relation features during training, allowing the model to learn geometric cues end-to-end. In contrast, TopoLogic [10] computes a scalar topology score for each lane pair based on GDT and applies it to adjust L2L predictions (GDT_{L2L})* ***only*** *at inference time. This post-hoc refinement is excluded from learning and relies on manually assigned hyperparameters, which may not generalize across scenarios. As shown in Tab. 5, this design limits performance and robustness, further highlighting the advantage of our approach.* To better understand these differences, we conduct three comparative experiments in Tab. 5 and analyze them below.

Exp. 1: GDT *vs.* Geometry-Biased Self-Attention. We replace our geometry encoding with TopoLogic's end-to-start distance to construct the self-attention bias for lane representation learning. As shown in Tab. 5, this modification degrades DET_l by 1.0, TOP_{ll} by 0.8, and TOP_{lt} by 1.4, suggesting that GDT is insufficient for capturing rich inter-lane spatial relationships and reaffirming the advantage of our relation embedding.

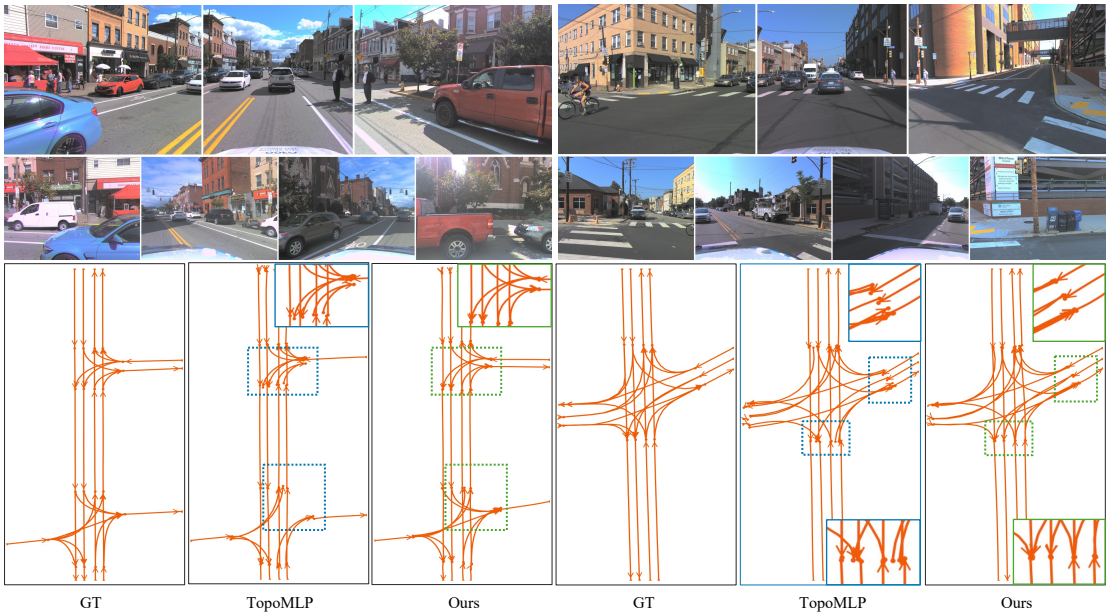

Figure 5: Comparative visual results on OpenLane-V2 subsetA. The top row shows multi-view input images, the bottom row shows lane predictions. We show comparison between ground truth, TopoMLP [46] and ours. The blue box highlights misaligned connection point predictions from TopoMLP, and the green box shows the corresponding aligned predictions from our CoPo . For clarity, zoomed-in views of selected regions are displayed at the top-right or bottom-right corners.

Exp. 2: Evaluating GDT in L2L Topology Reasoning. Unlike our geometry-enhanced L2L topology head, which integrates geometric cues into relation embeddings during training, TopoLogic applies GDT only as an inference-time adjustment, using manually assigned hyperparameters to balance GDT and model-based predictions rather than the weights learned during lane representation training. We evaluate TopoLogic's model under two variants: TopoLogic†, which uses the learned hyperparameters, and TopoLogic – GDT_{L2L}, which removes GDT from L2L inference. As shown in Tab. 5, both variants degrade performance, with a TOP_{ll} drop when GDT is removed, suggesting that TopoLogic's lane features alone do not sufficiently encode relational information and that post-hoc/manual adjustment limits robustness and generalizability.

Exp 3: Applying GDT to Our L2L Inference. We further investigate the generalizability of GDT by incorporating it into our model's L2L inference, following TopoLogic's ensembling strategy. Specifically, we fuse our L2L predictions with GDT outputs using their predefined hyperparameters, without modifying our trained model. As shown in Tab. 5, this ensembling leads to a decrease in TOP_{ll} by 0.5 (Ours + GDT_{L2L} *vs.* Ours). This outcome highlights the limited transferability of TopoLogic's formulation. In contrast, our model achieves strong L2L performance without requiring additional tuning or post-processing, demonstrating that our relation embedding offers a more effective and learnable approach to modeling geometric topology.

D Limitations and Future Work

Through qualitative analysis, we identify recurring challenging cases for topology reasoning, including small traffic elements, distant lanes, and occluded or partially visible lanes. Since OpenLane-V2 does not provide explicit scenario annotations such as occlusion labels, we further conduct attribute-based quantitative breakdowns for lane distance and traffic-element size.

Table 6: Lane detection by distance. %GT denotes the percentage of total lane instances.

Distance	DET_l	%GT
Near (< 25m)	34.4	45.3%
Far (≥ 25m)	30.2	54.7%

Table 7: TE detection by size. Area ratio denotes element area divided by image area.

Area ratio	DET_t	%GT
< 0.01%	13.2	19.2%
< 0.05%	41.5	68.7%
≥ 0.05%	59.9	31.3%

As shown in Tabs. 6 and 7, far-range lanes and small traffic elements are the dominant challenges. Far lanes account for 54.7% of lane instances and obtain lower DET_l than near lanes. Similarly, small traffic elements are much harder to detect: elements with area ratio below 0.01% only achieve 13.2 DET_t, while larger elements (≥ 0.05%) reach 59.9 DET_t. These detection failures can further affect downstream lane-to-lane and lane-to-traffic topology reasoning.

These challenging cases point toward future directions: incorporating temporal continuity across frames, leveraging motion cues or sequential context, and fusing richer spatiotemporal relational information may help improve robustness in far-range, small-object, and occluded-scene cases.