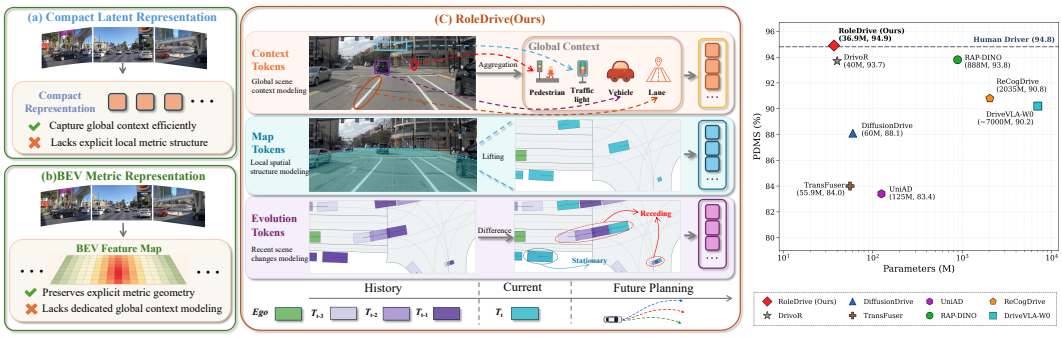
ROLEDRIVE: SCENE REPRESENTATIONS WITH DISTINCT PLANNING ROLES FOR END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end autonomous driving offers a promising route to safe and efficient navigation by learning to predict driving trajectories directly from sensor observations. However, many existing end-to-end planners mix different types of planning-relevant information within a unified representation, without explicitly differentiating their roles in trajectory planning, potentially limiting planning performance. To address this problem, we propose RoleDrive, an intermediate scene representation framework that explicitly organizes scene information into three complementary representations with distinct planning roles. Specifically, RoleDrive constructs Context Tokens for global scene context modeling and Map Tokens for fine-grained local spatial reasoning from shared visual features, while Evolution Tokens capture recent scene changes through residuals between aligned historical and current Map representations. These complementary representations are then progressively integrated through hierarchical trajectory reasoning. Without using any additional training data and with only 36.9M parameters, RoleDrive achieves a PDMS of 94.9 on NAVSIM-v1, surpassing the reported human-driver baseline, while also attaining state-of-the-art performance on NAVSIM-v2 and the photo-realistic closed-loop HUGSIM benchmark. Code will be publicly released upon acceptance at `https://github.com/K1ngAO/RoleDrive`.

1 INTRODUCTION



(a) Scene Representation Paradigms (b) Model Size vs. Planning Performance

Figure 1: **Overview of RoleDrive.** (a) Distinct scene representations are assigned to different planning roles. (b) RoleDrive achieves 94.9 PDMS on NAVSIM-v1 with only 36.9M parameters, surpassing the human-driver baseline.

End-to-end autonomous driving (E2E-AD) has made rapid progress by learning to predict driving trajectories directly from sensor observations and ego states, reducing the dependence on manually designed perception, prediction, and planning pipelines (Chitta et al., 2023; Hu et al., 2023; Jiang et al., 2023; Liao et al., 2025). Recent studies have increasingly explored intermediate scene representations as an interface between environmental perception and trajectory planning (Jiang et al.,

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.

2023; Sun et al., 2025; Li & Cui, 2025; Kirby et al., 2026). However, existing efforts mainly focus on the representation form itself, such as how scene information is vectorized, sparsified, or compressed, while paying less attention to whether different types of planning-relevant information should be expressed in different forms according to their roles in trajectory planning. This motivates us to revisit intermediate scene representation design from the perspective of planning roles.

For the current driving scene, trajectory planning requires both holistic scene understanding and detailed local spatial reasoning. As illustrated in Fig. 1(a), compact latent representations efficiently summarize visual observations into a small set of features, making them well suited for capturing global scene context (Li & Cui, 2025; Sun et al., 2025; Kirby et al., 2026; Kim et al., 2026; Jia et al., 2025). However, these compact representations may not capture local spatial details precisely enough to determine whether a planned trajectory can safely pass nearby obstacles. In contrast, BEV representations preserve explicit geometry in the ego coordinate frame and spatial correspondence, providing stronger support for detailed local spatial reasoning (Li et al., 2025c; 2023; Huang et al., 2021; Harley et al., 2023; Jiang et al., 2023; Liao et al., 2025; Zhang et al., 2025a). Yet, these representations focus more on local geometry, while the overall scene context is less explicitly represented. These differences suggest that global context and local geometry play complementary roles in trajectory planning and should be represented in different forms that preserve the information needed for each role.

Another challenge arises when historical observations are incorporated into trajectory planning. These observations provide additional cues about how the scene changes over time. One common approach combines features from historical observations with those of the current frame to build a scene representation that incorporates past information (Wang et al., 2023; Sun et al., 2025; Yuan et al., 2024). This allows previous observations to support current planning, but can also retain substantial unchanged scene content, leaving recent changes implicit within the combined features. Another approach uses historical observations to predict future scene states, providing planning with information about how the environment may evolve (Li et al., 2025b;a; Zhang et al., 2026; Li et al., 2026a; Huang et al., 2026). However, this introduces additional prediction objectives and modeling complexity. While both approaches exploit temporal information, they do not necessarily provide a separate representation of changes that have already occurred. This motivates us to consider whether historical information should be represented in terms of recent scene changes, rather than complete past scene content, to better support trajectory planning.

To address these challenges, we propose **RoleDrive**, an intermediate scene representation framework that organizes scene information according to its distinct roles in trajectory planning, as illustrated in Fig. 1(a). Specifically, **Context Tokens** aggregate visual observations across camera views into a small set of features, capturing surrounding traffic and road conditions for an overall understanding of the driving scene. **Map Tokens** preserve the local environment as a spatial grid in the ego coordinate frame, retaining the positions and spatial relationships of nearby vehicles and road structures for reasoning about trajectory feasibility. **Evolution Tokens** capture recent scene changes by aligning historical Maps to the current ego frame and encoding their differences from the current Map. This allows temporal changes to be represented separately while reducing repeated information from unchanged scene content. These representations support planning progressively: global context guides initial trajectory reasoning, while local geometry and recent changes provide further refinement. Extensive experiments demonstrate state-of-the-art performance on NAVSIM-v1, NAVSIM-v2, and the photorealistic closed-loop HUGSIM benchmark. As shown in Fig. 1(b), RoleDrive achieves **94.9 PDMS on NAVSIM-v1 with only 36.9M parameters**.

- **Role-Specific Scene Representation:** Compact Context Tokens capture global scene context, while metric Map Tokens preserve detailed local geometry, matching different representation forms to their planning roles.
- **Residual Evolution Modeling:** Residuals between aligned historical Maps and the current Map capture recent scene changes while reducing redundancy from unchanged scene content.
- **Coarse-to-Fine Trajectory Reasoning:** Global context first guides the formation of trajectory hypotheses, which are then refined using local geometry and recent scene changes.

2 Related Work

End-to-End Autonomous Driving. End-to-end (E2E) autonomous driving aims to learn driving policies directly from sensor observations and ego states, reducing dependence on manually designed modules. Representative methods explore trajectory prediction, control learning, and sensor fusion. TCP combines trajectory prediction with control prediction, while TransFuser uses transformers to fuse information from multiple sensor modalities. ST-P3 incorporates spatial and temporal information to support perception, prediction, and planning (Wu et al., 2022; Chitta et al., 2023; Hu et al., 2022). Other approaches explicitly organize the driving system around the planning task. UniAD integrates perception, prediction, and planning within a unified framework, while VAD uses vectorized scene representations for efficient trajectory planning (Hu et al., 2023; Jiang et al., 2023). Recent studies further improve how planners use temporal information and scene understanding. BridgeAD incorporates historical prediction and planning information into current reasoning, while MomAD uses historical guidance to improve trajectory consistency (Zhang et al., 2025a; Song et al., 2025). ReCogDrive combines a vision language model with a diffusion planner and reinforcement learning to connect scene understanding with trajectory generation (Li et al., 2026b). Related work on learned planning, such as Flow Planner, uses flow matching and explicit modeling of agent interactions to improve planning in interactive driving scenarios (Tan et al., 2026).

Intermediate Scene Representation. Intermediate scene representations serve as an interface between visual perception and trajectory planning by organizing sensor features into information useful for planning. Latent features and queries provide flexible ways to extract and organize visual information, as demonstrated by general representation learning and visual perception methods (Jaegle et al., 2021; 2022; Carion et al., 2020; Liu et al., 2022). A major line of research uses bird's-eye-view (BEV) representations to organize features from multiple camera views in a shared ego coordinate frame, preserving explicit spatial relationships and metric geometry for downstream reasoning (Philion & Fidler, 2020; Li et al., 2025c; 2023; Huang et al., 2021; Xue et al., 2024; 2025; Harley et al., 2023). To reduce the computational cost and spatial redundancy of dense representations, other studies explore vectorized and sparse representations that describe relevant scene elements using structured tokens or queries (Jiang et al., 2023; Sun et al., 2025; Jia et al., 2025). Compact latent methods instead summarize scene information into a small number of informative tokens, showing that effective planning does not always require a dense scene representation (Li & Cui, 2025; Kirby et al., 2026; Kim et al., 2026). Beyond representing the current scene, streaming methods incorporate information from historical observations (Wang et al., 2023; Yuan et al., 2024), while predictive representations and world models capture scene evolution and possible future states (Xue et al., 2026; Li et al., 2025a; Min et al., 2024; Zheng et al., 2024; Li et al., 2025b; Zhang et al., 2026). Overall, existing studies have extensively explored how scene information can be represented more explicitly, sparsely, or compactly. However, comparatively less attention has been paid to whether different types of planning information should adopt representation forms suited to their distinct roles in trajectory planning.

3 Method

3.1 RoleDrive Framework Overview

Given images from multiple cameras and the ego state, RoleDrive predicts future ego trajectories through the framework illustrated in Fig. 2. Inspired by the visual token compression design of DrivoR (Kirby et al., 2026), we introduce learnable Context Tokens alongside image patch tokens to form a compact representation of global scene context. We additionally introduce trajectory queries conditioned on the ego state into the same ViT sequence, enabling planning cues to guide visual feature learning during scene encoding. After joint encoding, the Context Tokens are retained as the global scene representation, while the encoded patch features are geometrically lifted into the ego coordinate frame to construct Map Tokens for local metric reasoning. Historical Maps are stored in a temporal memory and aligned to the current ego frame. Residuals between the aligned historical Maps and the current Map are then aggregated into Evolution Tokens to capture recent scene changes. Finally, the encoded trajectory queries first interact with Context Tokens and are then refined using Map and Evolution Tokens. The refined queries are decoded into candidate trajectories, which are evaluated by a scoring module to select the final trajectory.

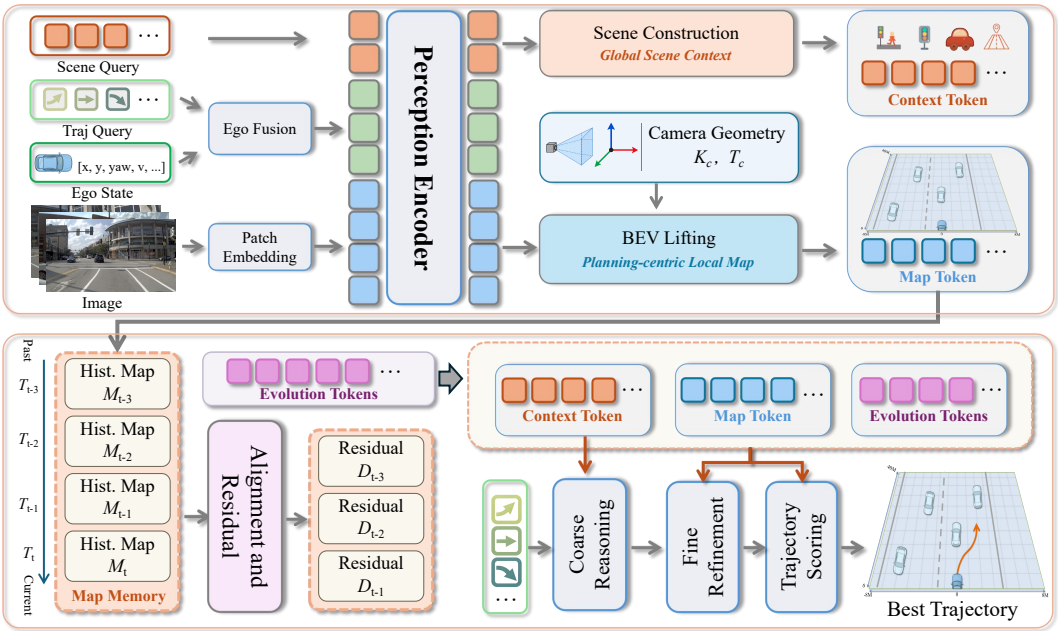


Figure 2: Overview of **RoleDrive**. Context, Map, and Evolution Tokens represent global scene context, local geometry, and recent scene changes, respectively. They support coarse-to-fine trajectory reasoning, followed by scoring to select the final trajectory.

3.2 Current Scene Modeling

For the current driving scene, RoleDrive constructs two complementary representations: compact **Context Tokens** summarize the overall scene, while **Map Tokens** preserve detailed local geometry in the ego coordinate frame. Both are built from the same visual observations, but organize the information differently to support global scene understanding and local spatial reasoning, respectively.

Context Tokens. Following the visual encoding design of DrivoR (Kirby et al., 2026), we use a pretrained DINOv2 ViT-S (Oquab et al., 2024) and add learnable Context Tokens to the image token sequence of each camera to summarize global scene information. We use 16 Context Tokens for each of the front, front-left, and front-right cameras, giving 48 Context Tokens in total. We additionally introduce trajectory queries conditioned on the ego state and feed them into the ViT together with the Context Tokens and image patch tokens. Through self-attention, these trajectory queries interact with visual features during encoding, allowing planning information to guide scene representation learning. After encoding, we concatenate the Context Tokens from the three views to form the global context representation for initial trajectory reasoning. The encoded patch features preserve spatial information and are used to construct Map Tokens in the ego coordinate frame.

Map Tokens. While Context Tokens summarize the overall scene, detailed trajectory reasoning requires explicit information about local positions and distances. We therefore construct Map Tokens from the spatial patch features $\widetilde{\mathbf{P}}_t^c$ produced by the shared visual encoder. We define a 3D grid in the ego coordinate frame, covering 0–40 m longitudinally, −8–8 m laterally, and −5–5 m vertically. Grid locations are projected into the three camera views using the camera intrinsics K_c and extrinsics T_c to sample the corresponding visual features. Features from valid views are combined across cameras and compressed along the height dimension. We then add positional embeddings that encode metric coordinates, obtaining the Map representation $\mathbf{M}_t \in \mathbb{R}^{2560\times256}$. The representation is arranged on a 32×80 grid with a resolution of 0.5 m in both horizontal directions. Each Map Token corresponds to a fixed spatial cell in the current ego frame, preserving local geometry for trajectory reasoning.

To help Map Tokens identify regions important for safe planning, we introduce auxiliary risk supervision inspired by occupancy and safety modeling (Hu et al., 2022; 2023). The idea is to assign high scores to regions occupied by objects and lower scores to nearby regions, with the latter decreasing

as distance increases. We first project annotated object boxes onto the current Map grid. For each cell i, $c_i \in [0, 1]$ denotes the fraction covered by an object, and $h_i \in [0, 1]$ assigns a larger value to locations closer to dynamic agents. The target score is defined as

$$r_i = \begin{cases} 0.5 + 0.5c_i, & c_i > 0, \\ 0.5h_i, & c_i = 0. \end{cases} \tag{1}$$

A lightweight head predicts these scores from the current Map features. We train it jointly with the planner using weighted binary cross-entropy and Dice losses, placing greater emphasis on cells near the ego vehicle and those with high target scores. This supervision helps Map Tokens retain information about obstacles and nearby regions that matter for planning, despite the large proportion of empty cells. The predicted scores are used only for auxiliary supervision and are not fed into the trajectory planner.

3.3 Scene Evolution Modeling

While Map Tokens describe the current local environment, historical observations reveal how it has changed. As illustrated in Fig. 3, RoleDrive stores historical Maps, aligns them to the current ego frame, and computes their differences from the current Map. These differences are aggregated into **Evolution Tokens**, allowing recent changes to be represented separately from the current scene.

Historical Map Alignment. We maintain Maps from K historical observations sampled at $0.5\,\mathrm{s}$ intervals. Each observation is processed by the shared visual encoder and geometric lifting module to produce a Map with the same 32×80 layout as the current Map. As illustrated on the left of Fig. 3, the planning region moves and rotates with the ego vehicle: the purple dashed boxes show historical planning regions, while the red dashed box shows the current region. The same grid index can therefore refer to different physical locations across frames. We align each historical Map to the current ego frame before comparing their features:

$$\widetilde{\mathbf{M}}_{t-k\to t} = \mathcal{W}(\mathbf{M}_{t-k}, \mathbf{T}_{t-k\to t}), \tag{2}$$

where $\mathbf{T}_{t-k\to t}$ denotes the pose transformation from the historical ego frame to the current ego frame, and $\mathcal{W}$ denotes backward warping with bilinear sampling. For each current grid cell, we use the inverse transformation to find the corresponding position in the historical Map and sample its features. The aligned Maps shown in the middle of Fig. 3 therefore share the current grid, allowing features at the same physical locations to be compared.

Alignment also identifies where historical information is available. In the aligned Maps, pink cells indicate valid overlap with historical coverage, while gray cells indicate current locations without historical correspondence. We record this distinction using a validity mask $\mathbf{V}_{t,k}$, which marks a cell as valid only when its sampling position lies within the historical Map region $\Omega = [0, 40) \times [-8, 8)$. Only valid cells contribute to temporal comparison, preventing missing history from being interpreted as scene changes.

Evolution Token Construction. As illustrated in the lower middle part of Fig. 3, we subtract each aligned historical Map from the current Map and retain only valid overlapping cells:

$$\mathbf{D}_{t,k} = \mathbf{V}_{t,k} \odot \left(\mathbf{M}_t - \widetilde{\mathbf{M}}_{t-k\to t}\right), \qquad k = 1, \ldots, K, \tag{3}$$

where $\odot$ denotes elementwise multiplication, with the validity mask applied across feature channels. Comparing aligned features reduces differences caused by ego motion. Features that remain similar over time largely cancel, while changes in object positions or appearance can produce larger residuals. The purple regions in Fig. 3 illustrate these residual responses within the valid overlap.

The temporal residual aggregation module then combines information from multiple historical observations at each grid location. We add temporal and relative pose information to the residual features and aggregate them using temporal attention, excluding historical entries without valid coverage. This produces one Evolution Token per cell, forming $\mathbf{E}_t \in \mathbb{R}^{2560\times 256}$ with the same spatial layout as the current Map. As shown on the right of Fig. 3, the current Map is retained as Map Tokens, while the aggregated residuals form Evolution Tokens. We concatenate the two token sets as $\mathbf{G}_t = [\mathbf{M}_t; \mathbf{E}_t]$, allowing trajectory reasoning to use both the current local environment and its recent changes. Further details of temporal aggregation are provided in Appendix A.

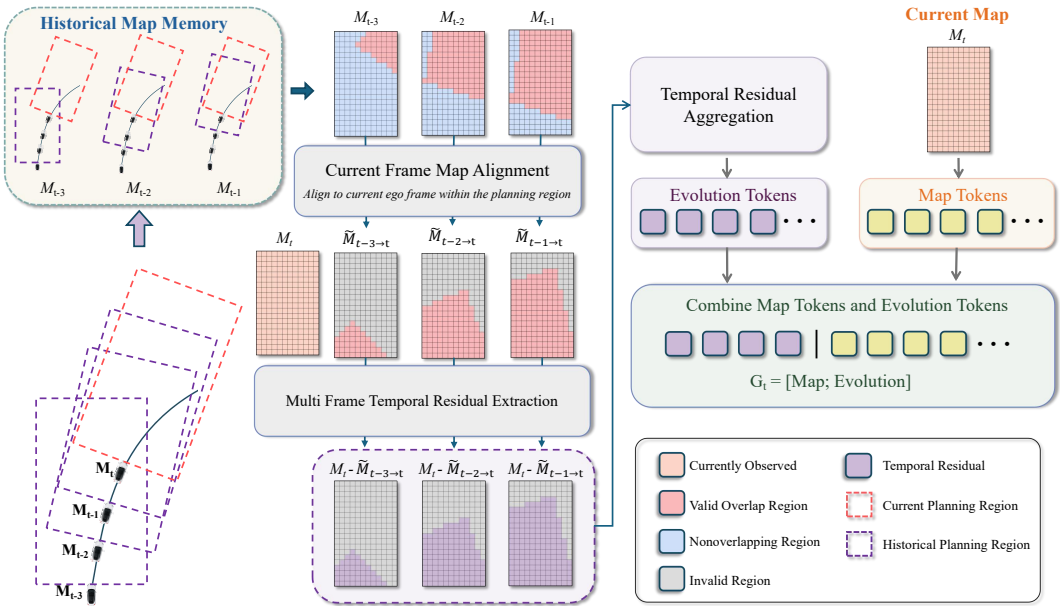


Figure 3: Scene evolution modeling in **RoleDrive**. Historical Maps are aligned to the current ego frame and compared with the current Map within valid overlapping regions. The resulting residuals are aggregated into Evolution Tokens, which complement Map Tokens with information about recent scene changes.

3.4 Coarse-to-Fine Trajectory Reasoning

After constructing Context, Map, and Evolution Tokens, RoleDrive exploits their complementary planning roles in a coarse-to-fine manner. The ego-conditioned trajectory queries, which have already interacted with visual observations in the perception encoder, are retained as the initial planning queries. In the coarse stage, these queries interact with **Context Tokens** to establish trajectory hypotheses from the global driving context. The context-conditioned queries are then further refined using **Map Tokens** and **Evolution Tokens**, incorporating fine-grained local geometry and recent scene changes into trajectory reasoning. The refined queries are finally decoded into $M = 64$ candidate trajectories, which are evaluated by a trajectory scoring head to select the final trajectory. In this way, RoleDrive progressively moves from global scene understanding to detailed spatial and temporal reasoning rather than exposing trajectory queries to all scene information at once.

Training Objective. RoleDrive is trained in a single stage by jointly optimizing trajectory generation, proposal scoring, and auxiliary Map risk prediction. We follow DrivoR (Kirby et al., 2026) for the trajectory and scoring losses and add the risk objective described in Sec. 3.2:

$$\mathcal{L} = \mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{score}} + \lambda_{\text{risk}}\mathcal{L}_{\text{risk}}, \tag{4}$$

where $\mathcal{L}_{\text{traj}}$, $\mathcal{L}_{\text{score}}$, and $\mathcal{L}_{\text{risk}}$ supervise trajectory generation, proposal scoring, and Map risk prediction, respectively. The risk weight λ_{risk} decreases from 2.0 to 0.2 over the first 15 epochs and remains at 0.2 thereafter.

4 Experiments

4.1 Experimental Details

We evaluate **RoleDrive** on NAVSIM-v1 and NAVSIM-v2 (Dauner et al., 2024), as well as the photorealistic closed-loop HUGSIM benchmark (Zhou et al., 2026). We use a pretrained DINOv2 ViT-S (Oquab et al., 2024) as the visual backbone. For NAVSIM, the final models are trained on `navtrain+navval` and evaluated on `navtest`, while ablation models are trained on `navtrain` and evaluated on `navval`. We report PDMS on NAVSIM-v1 and EPDMS on NAVSIM-v2. For HUGSIM, we evaluate the model trained on NAVSIM-v1 without additional

Table 1: **NAVSIM-v1.** Comparison with representative methods on the NAVSIM-v1 `navtest` split. Higher is better for all metrics.

Method	Venue	NC	DAC	TTC	Comf.	EP	PDMS
Human driver (Dauner et al., 2024)	NeurIPS'24	100	100	100	99.9	87.5	94.8
RAP-DINO† (Feng et al., 2026)	ICLR'26	99.1	98.9	96.7	100	90.3	93.8
UniAD (Hu et al., 2023)	CVPR'23	97.8	91.9	92.9	100	78.8	83.4
World4Drive (Zheng et al., 2025)	ICCV'25	97.4	94.3	92.8	100	79.9	85.1
VAD-v2 (Jiang et al., 2026)	ICLR'26	98.1	94.8	94.3	100	80.6	86.2
DiffusionDrive (Liao et al., 2025)	CVPR'25	98.2	96.2	94.7	100	82.2	88.1
DriveVLA-W0 (Li et al., 2026a)	ICLR'26	98.7	99.1	95.3	99.3	83.3	90.2
SeerDrive (Zhang et al., 2025b)	NeurIPS'25	98.8	98.6	95.8	100	84.2	90.7
ReCogDrive (Li et al., 2026b)	ICLR'26	97.9	97.3	94.9	100	87.3	90.8
SafeDrive (Kim et al., 2026)	CVPR'26	99.5	99.0	97.2	100	84.3	91.6
DriveSuprim (Yao et al., 2026)	AAAI'26	98.6	98.6	95.5	100	91.3	93.5
DrivoR (Kirby et al., 2026)	CVPR'26	99.0	98.9	96.7	100	90.0	93.7
RoleDrive	Ours	99.1	99.2	96.5	100	92.7	**94.9**

† RAP-DINO (Feng et al., 2026) uses approximately $10\times$ the training data of `navtrain`. Bold indicates the highest PDMS in the table.

Table 2: **NAVSIM-v2.** Planning performance on the `navtest` benchmark. "C" and "L" denote camera and LiDAR inputs. Pre-fix and Corrected distinguish evaluator versions for the same model predictions. Higher is better for all metrics.

Method	Venue	Input	NC	DAC	DDC	TLC	EP	TTC	LK	HC	EC	EPDMS
DiffusionDrive (Liao et al., 2025)	CVPR'25	C&L	98.2	95.9	99.4	99.8	87.5	97.3	96.8	98.3	87.7	84.5
DriveSuprim (Yao et al., 2026)	AAAI'26	C	97.8	97.9	99.5	99.9	90.6	97.1	96.6	98.3	77.9	86.0
ReCogDrive (Li et al., 2026b)	ICLR'26	C	98.3	95.2	99.5	99.8	87.1	97.5	96.6	98.3	86.5	83.6
GaussianFusion (Liu et al., 2025)	NeurIPS'25	C&L	98.3	97.3	99.0	97.4	87.5	97.4	97.4	98.3	–	85.0
DiffRefiner (Yin et al., 2026)	AAAI'26	C	98.5	97.4	99.6	99.8	87.6	97.7	97.7	98.3	86.2	86.2
DrivoR† (Kirby et al., 2026) (Pre-fix)	CVPR'26	C	99.0	98.7	98.6	99.5	92.8	98.6	95.0	97.2	68.4	86.1
DrivoR† (Corrected)	CVPR'26	C	99.0	98.7	98.6	99.5	92.8	98.6	95.0	97.2	68.4	89.6
RoleDrive† (Pre-fix)	Ours	C	99.5	99.3	99.8	99.6	90.4	99.0	96.6	97.9	72.8	**87.8**
RoleDrive† (Corrected)	Ours	C	99.5	99.3	99.8	99.6	90.4	99.0	96.6	97.9	72.8	**91.5**

† Pre-fix and Corrected denote rescoring the same predictions before and after the evaluator fix on all 12,146 scenarios using one-stage averaging. Reported component metrics remain unchanged. Bold marks the higher EPDMS between DrivoR and RoleDrive under each version. Literature results follow their reported protocols, which may differ. "–" denotes unavailable results.

Table 3: **Photorealistic closed-loop evaluation on HUGSIM.** RoleDrive is evaluated without additional training using the model trained on NAVSIM-v1. Higher is better for all metrics.

	RC					HD-Score				
Method	E	M	H	X	Avg.	E	M	H	X	Avg.
VAD (Jiang et al., 2023)	51.3	31.1	25.3	26.5	31.4	36.3	9.5	8.0	11.5	13.4
LTF (Chitta et al., 2023)	67.8	35.1	26.2	40.5	38.9	58.9	18.0	9.8	25.9	23.7
UniAD (Hu et al., 2023)	78.4	60.5	33.6	17.8	45.9	64.9	45.8	20.6	6.6	32.7
DrivoR‡ (Kirby et al., 2026)	80.6	44.0	32.2	44.6	46.4	72.1	26.6	15.9	29.1	31.0
RoleDrive	82.9	50.2	31.9	40.8	**48.0**	75.6	35.3	16.3	26.8	**34.2**

E: Easy, M: Medium, H: Hard, X: Extreme. RC denotes Route Completion. Bold highlights the highest reported values in the Avg. columns.
‡ DrivoR is reproduced by us using PyTorch 2.1.0, CUDA 12.1, PyTorch Lightning 2.4.0, and NumPy 1.24.4.

training and report Route Completion (RC) and HD-Score. Unless otherwise specified, we follow the training protocol of DrivoR (Kirby et al., 2026). We train our models on four NVIDIA GeForce RTX 4090 GPUs. Further implementation details are provided in Appendix A.

4.2 Benchmarks

NAVSIM. Tables 1 and 2 report planning performance on NAVSIM-v1 and NAVSIM-v2, respectively. Using only **36.9M parameters** and **no additional training data**, RoleDrive achieves **94.9 PDMS** on NAVSIM-v1, outperforming DrivoR by 1.2 points and slightly exceeding the reported human-driver reference of 94.8. The improvement is mainly reflected in ego progress: EP increases from 90.0 to **92.7**, while safety scores remain comparable. NC and DAC improve from 99.0 and 98.9

Table 4: **Ablation studies of RoleDrive.** We evaluate role-specialized representations, current-scene modeling, temporal evolution modeling, and representation integration.

(a) Role-Specialized Representations

Context	Map	Evol.	NC	DAC	TTC	EP	PDMS
✓			98.8	96.9	96.0	88.1	91.5
	✓		98.6	97.8	95.4	89.7	92.3
✓	✓		98.7	97.8	95.7	90.0	92.5
✓	✓	✓	99.0	97.7	96.3	90.1	**92.9**

(b) Current Scene Modeling

Traj. in ViT	Risk Sup.	NC	DAC	TTC	EP	PDMS
		98.4	97.3	95.2	89.3	92.0
✓		98.6	97.6	95.6	89.7	92.3
	✓	98.5	97.7	95.5	89.6	92.2
✓	✓	98.7	97.8	95.7	90.0	**92.5**

(c) Evolution Token Construction

Variant	NC	DAC	TTC	EP	PDMS
Current	98.7	97.8	95.7	90.0	92.5
Fusion	98.6	97.8	95.4	90.0	92.4
Unaligned	98.9	98.0	96.1	89.5	92.7
Evolution	99.0	97.7	96.3	90.1	**92.9**

(d) Role Integration Strategy

Strategy	NC	DAC	TTC	EP	PDMS
Joint	98.7	97.9	95.4	90.1	92.5
Progressive	99.0	97.7	96.3	90.1	**92.9**

Context, Map, and Evolution indicate the representations explicitly provided to the trajectory planner. Traj. in ViT denotes the introduction of ego-conditioned trajectory queries into the visual encoder, while Risk Sup. denotes risk-aware auxiliary supervision for Map representation learning. Fusion directly aggregates aligned historical representations, whereas Unaligned constructs Evolution Tokens without ego-motion alignment. Joint and Progressive denote joint and progressive representation integration, respectively.

to 99.1 and 99.2, respectively, TTC decreases slightly from 96.7 to 96.5, and comfort remains at 100. These results indicate greater progress with broadly maintained safety performance. RoleDrive also surpasses RAP-DINO despite the latter using approximately $10\times$ more training data. On NAVSIM-v2, RoleDrive achieves **91.5 EPDMS** under the corrected scoring protocol, exceeding DrivoR by 1.9 points. NC, DAC, and TTC increase from 99.0, 98.7, and 98.6 to 99.5, 99.3, and 99.0, respectively, while EP decreases from 92.8 to 90.4. The higher aggregate score therefore accompanies improvements in safety and compliance despite lower ego progress. Together, these results demonstrate strong planning performance across both benchmarks with a compact model.

HUGSIM. We evaluate zero-shot closed-loop generalization on HUGSIM using the model trained on NAVSIM-v1 without additional training. As reported in Table 3, RoleDrive achieves an average RC of **48.0** and an average HD-Score of **34.2**, compared with 46.4 and 31.0 for our DrivoR reproduction. Improvements in both metrics suggest promising transfer to photorealistic closed-loop driving. The comparison includes differences in controller settings, as detailed in Appendix A, and therefore reflects overall system performance rather than an isolated comparison of the planning architectures.

4.3 Ablation Studies

4.3.1 Quantitative Analysis

Table 4(a) evaluates the overall contribution of the three role-specialized representations. Context and Map Tokens individually achieve 91.5 and 92.3 PDMS, respectively, highlighting the value of explicit local geometry for trajectory planning. Combining them improves PDMS to 92.5, suggesting that global context provides complementary cues beyond local geometry. Adding Evolution Tokens further raises PDMS to 92.9, with improvements in NC and TTC, supporting the benefit of incorporating recent scene changes alongside current scene information. Together, these results support representing global context, local geometry, and recent changes as complementary sources of information for planning.

Table 4(b–d) examines how these representations are learned, constructed, and integrated. For current scene modeling, introducing trajectory queries into the visual encoder and applying risk supervision individually improve PDMS from 92.0 to 92.3 and 92.2, respectively. Combining both achieves 92.5, supporting the use of early interaction between planning queries and visual features together with spatial auxiliary supervision. For temporal modeling, directly fusing aligned historical Maps yields 92.4 PDMS, slightly below the baseline without history at 92.5, indicating that direct historical feature aggregation provides no improvement in this comparison. Temporal residuals achieve 92.7 without alignment and 92.9 with alignment, while alignment also increases EP from 89.5 to 90.1. These results favor extracting temporal differences over directly aggregating historical features and show the benefit of comparing corresponding spatial locations. For representation in-

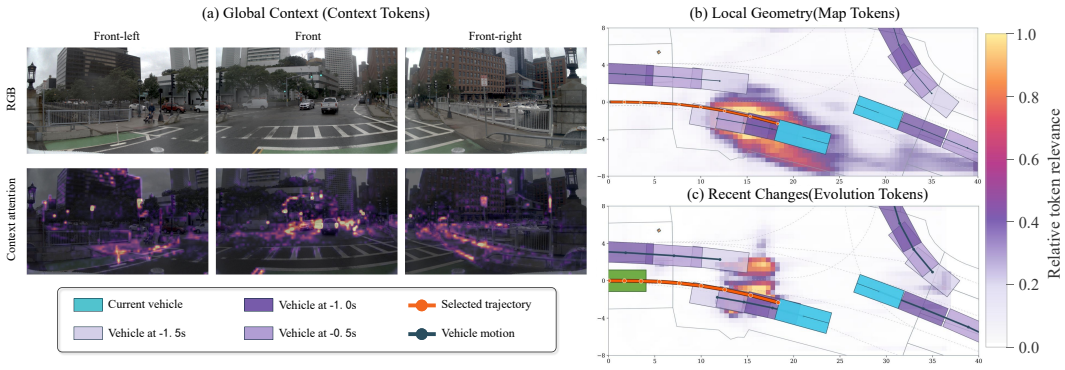


Figure 4: Role-specialized representations in **RoleDrive**. (a) Multi-view images and Context Token attention. (b) Planning attention over Map Tokens. (c) Evolution Token relevance weighted by temporal residual evidence within valid historical overlap. Heatmaps are normalized separately. Historical observations at -1.5, -1.0, and -0.5 s are used to construct Evolution Tokens.

tegration, progressive reasoning improves PDMS from 92.5 to 92.9, with higher NC and TTC and unchanged EP at the reported precision. This supports first using global context to guide trajectory reasoning and then incorporating local geometry and recent changes for refinement.

4.3.2 Qualitative Analysis

Fig. 4 visualizes the three representations in the same driving scene. In Fig. 4(a), Context Token attention highlights vehicles, traffic lights, road signs, and road structure across the three camera views. These responses span multiple scene elements, illustrating how compact Context Tokens gather visual cues about surrounding traffic and road conditions for global scene understanding. In Fig. 4(b), Map Token relevance extends along the selected trajectory and nearby vehicles. Its explicit metric layout makes the positions of surrounding vehicles relative to the planned path directly visible, providing a spatial reference for interpreting the current driving situation. In Fig. 4(c), Evolution Token relevance is more localized around regions associated with recent vehicle movement. Historical positions and motion traces help relate these responses to changes over time, illustrating the temporal information represented alongside the current scene geometry.

Map and Evolution Tokens highlight overlapping regions because moving vehicles contribute to both current geometry and temporal changes. However, Map relevance covers a broader region around the trajectory and surrounding vehicles, whereas Evolution relevance concentrates on a more localized portion of the region associated with recent motion. The shared interaction region therefore provides a useful reference for comparing their different spatial emphases: Map Tokens describe the current spatial configuration, while Evolution Tokens emphasize changes extracted from aligned historical observations. Their complementary roles are thus consistent with overlapping responses, without requiring attention to entirely separate regions. Together with the Context Token visualization, these observations provide a qualitative interpretation of the representation design and complement the planning improvements observed in the quantitative ablations.

5 Conclusion

We presented RoleDrive, a framework that uses different representation forms for distinct planning functions: compact Context Tokens capture global context, spatially structured Map Tokens preserve local geometry, and residual-based Evolution Tokens represent recent changes. These complementary representations are progressively integrated for trajectory planning. Experiments on NAVSIM-v1 and NAVSIM-v2 demonstrate strong performance with a compact model, while HUGSIM evaluation shows promising transfer to closed-loop driving. These findings support matching representation forms to their intended planning roles, allowing each to preserve relevant information and provide complementary planning cues.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, drafting portions of the manuscript, literature retrieval, and related-work discovery. They were not used for research ideation or execution, synthetic data generation, or proving mathematical claims. All AI-assisted content was carefully reviewed, verified, and revised by the authors to ensure technical accuracy and consistency with the presented methods and experimental results. The authors take full responsibility for the final content of this work, including all text, claims, and conclusions.

REPRODUCIBILITY STATEMENT

Implementation details, model configurations, training procedures, and evaluation settings are provided in Appendix A. Additional experiments and representation analyses are presented in Appendix B. We will publicly release the source code and configuration files upon acceptance.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12878–12895, 2023. doi: 10.1109/TPAMI.2022.3200245.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28706–28719. Curran Associates, Inc., 2024. doi: 10.52202/079017-0902. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/32768f7faf1995026ef9821c696f3404-Paper-Datasets_and_Benchmarks_Track.pdf.

Lan Feng, Yang Gao, Eloi Zablocki, Quanyi Li, Wuyang Li, Sichao Liu, Matthieu Cord, and Alexandre Alahi. Rap: 3d rasterization augmented end-to-end planning. In *International Conference on Learning Representations*, volume 2026, pp. 37722–37737, 2026.

Adam W. Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2759–2765, 2023. doi: 10.1109/ICRA48891.2023.10160831.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pp. 533–549. Springer, 2022.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862, 2023. doi: 10.1109/CVPR52729.2023.01712.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Zhao Huang, Bowen Zhang, Zhongzhu Li, and Di Lin. Occdriver: Future occupancy guided dual-branch trajectory planner in autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 141940–141962, 2026.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *International Conference on Learning Representations*, 2022. URL https://mlanthology.org/iclr/2022/jaegle2022iclr-perceiver/.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 67227–67243, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a7afc9957f1190223763b6ea93218f98-Paper-Conference.pdf.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8306–8316, 2023. doi: 10.1109/ICCV51070.2023.00766.

Bo Jiang, Shaoyu Chen, Hao Gao, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. In *International Conference on Learning Representations*, volume 2026, pp. 68886–68900, 2026.

Jungho Kim, Jiyong Oh, Seunghoon Yu, Hongjae Shin, Donghyuk Kwak, and Jun Won Choi. Safedrive: Fine-grained safety reasoning for end-to-end driving in a sparse world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24854–24864, June 2026.

Ellington Kirby, Alexandre Boulch, Yihong Xu, Yuan Yin, Gilles Puy, Éloi Zablocki, Andrei Bursuc, Spyros Gidaris, Renaud Marlet, Florent Bartoccioni, et al. Driving on registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32058–32069, 2026.

Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 29118–29134, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/47cf6de08432fb54ffd6b44fc1260413-Paper-Conference.pdf.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *International Conference on Learning Representations*, volume 2025, pp. 42942–42959, 2025a.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27137–27146, 2025b. doi: 10.1109/ICCV51701.2025.02519.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 7890–7911, 2026a.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1477–1485, 2023.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 157518–157556, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ff7bf6014f7826da531aa50f4538ee19-Paper-Conference.pdf.

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, 2025c. doi: 10.1109/TPAMI.2024.3515454.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12037–12047, 2025. doi: 10.1109/CVPR52734.2025.01124.

Shuai Liu, Quanmin Liang, Zefeng Li, Boyang Li, and Kai Huang. Gaussianfusion: Gaussian-based multi-sensor fusion for end-to-end autonomous driving. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48062–48084. Curran Associates, Inc., 2025. doi: 10.52202/085713-1603. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/44a76997ae5874529e49f2c2bd47b41c-Paper-Conference.pdf.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*, pp. 531–548. Springer, 2022.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15522–15533. IEEE, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research*, 2024. URL https://mlanthology.org/tmlr/2024/oquab2024tmlr-dinov2/.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pp. 194–210. Springer, 2020.

Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei Yang, and Yadan Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22432–22441, 2025. doi: 10.1109/CVPR52734.2025.02089.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8795–8801, 2025. doi: 10.1109/ICRA55743.2025.11128800.

Tianyi Tan, Yinan Zheng, Ruiming Liang, Zexu Wang, Kexin Zheng, Jinliang Zheng, Jianxiong Li, Xianyuan Zhan, and Jingjing Liu. Flow matching-based autonomous driving planning with advanced interactive behavior modeling. *Advances in Neural Information Processing Systems*, 38:38310–38335, 2026.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3598–3608, 2023. doi: 10.1109/ICCV51070.2023.00335.

Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6119–6132. Curran Associates, Inc., 2022. doi: 10.52202/068431-0443. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/286a371d8a0a559281f682f8fbf89834-Paper-Conference.pdf.

Yujie Xue, Ruihui Li, Fan Wu, Zhuo Tang, Kenli Li, and Mingxing Duan. Bi-ssc: Geometric-semantic bidirectional fusion for camera-based 3d semantic scene completion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20124–20134. IEEE, 2024.

Yujie Xue, Huilong Pi, Jiapeng Zhang, Yunchuan Qin, Zhuo Tang, Kenli Li, and Ruihui Li. Sd-former: Vision-based 3d semantic scene completion via sam-assisted dual-channel voxel transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26837–26847. IEEE, 2025.

Yujie Xue, Meng Wang, Ruihui Li, Fan Wu, Zhizhong Liu, Zhuo Tang, and Kenli Li. Learning spatial-temporal consistency for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28567–28577, June 2026.

Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M Alvarez, and Zuxuan Wu. Drivesuprim: Towards precise trajectory selection for end-to-end planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11910–11918, 2026.

Liuhan Yin, Runkun Ju, Guodong Guo, and Erkang Cheng. Diffrefiner: Coarse to fine trajectory planning via diffusion refinement with semantic interaction for end to end autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12009–12017, 2026.

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, 2024. doi: 10.1109/WACV57701.2024.00719.

Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6854–6863, 2025a. doi: 10.1109/CVPR52734.2025.00643.

Bozhou Zhang, Nan Song, jingyu li, Xiatian Zhu, Jiankang Deng, and Li Zhang. Future-aware end-to-end driving: Bidirectional modeling of trajectory planning and scene evolution. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 10204–10229. Curran Associates, Inc., 2025b. doi: 10.52202/085713-0345. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0ee633a6ade45eab4276352b3ee79c7a-Paper-Conference.pdf.

Jinqing Zhang, Zehua Fu, zelinxu, wenying.dai, Qingjie Liu, and Yunhong Wang. Resworld: Temporal residual world model for end-to-end autonomous driving. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 6471–6483, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0b6df1a973b82b3cf7fadca6c387ae5a-Paper-Conference.pdf.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28632–28642. IEEE, 2025.

Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4): 4673–4691, 2026. doi: 10.1109/TPAMI.2025.3647952.

A Implementation and Evaluation Details

Model Configuration. RoleDrive uses images from the front-left, front, and front-right cameras, with an input resolution of 1148×672 pixels (width $\times$ height). The visual encoder is initialized from a pretrained DINOv2 ViT-S and adapted using LoRA with rank 32. Each camera uses 16 learnable Context Tokens, yielding 48 tokens across the three views. The geometric lifting module constructs an ego-centric representation covering $[0, 40)$ m longitudinally, $[-8, 8)$ m laterally, and $[-5, 5)$ m vertically, using eight vertical bins. Height compression produces a 32×80 Map with a horizontal resolution of 0.5 m. Map and Evolution representations each contain 2,560 tokens with a feature dimension of 256. The planner generates 64 candidate trajectories. Historical observations are sampled at 0.5 s intervals, and the effect of the history length K is evaluated in Appendix B. Table 5 summarizes the model configuration and training settings.

Table 5: Model configuration and training settings of RoleDrive.

Model Configuration		**Training Settings**	
Configuration	Setting	Configuration	Setting
Camera views	Front-left, front, front-right	Framework	PyTorch
Input resolution	1148×672 (W $\times$ H)	Training epochs	25
Visual backbone	DINOv2 ViT-S (pretrained)	Optimizer	AdamW
Backbone adaptation	LoRA, rank 32	Base learning rate	2×10^{-4}
Context Tokens	16 per view, 48 total	LR schedule	Linear warmup, then cosine decay
Longitudinal range	$[0, 40)$ m	Warmup fraction	10% of configured schedule
Lateral range	$[-8, 8)$ m	Batch size per GPU	16
Vertical range	$[-5, 5)$ m, 8 bins	Global batch size	64
Map grid	32×80	Gradient accumulation	None
Grid resolution	0.5 m	Precision	Mixed precision
Map / Evolution tokens	2,560 each	Random seed	2
Map / Evolution dim.	256	Risk losses	Weighted BCE and Dice
Temporal attention	4 heads	Risk weight schedule	Cosine decay, $2.0 \rightarrow 0.2$
Default history length	$K = 3$	Risk decay duration	First 15 epochs
History interval	0.5 s	Final risk weight	0.2
Candidate trajectories	64	Training GPUs	$4\times$ RTX 4090[†]
Prediction horizon	4 s, 8 waypoints	Memory per GPU	48 GB

[†] The GPUs are modified RTX 4090 cards with 48 GB memory each. Training uses distributed data parallelism.

Temporal Residual Aggregation. Historical Maps are aligned to the current ego frame through bilinear sampling, with validity masks identifying cells with historical correspondence. At each current Map location, residual features are augmented with learned embeddings of spatial coordinates, relative ego pose, temporal offset, and frame identity. A learned query, conditioned on the cell coordinates and the fraction of historical frames with valid coverage, aggregates these inputs through four-head attention. Temporal aggregation is performed independently at each spatial location, and invalid historical entries are excluded through an attention mask. An additional current-frame reference entry has a zero residual component and remains valid, preventing an entirely masked attention sequence when historical correspondence is unavailable. The aggregated output is projected and combined with the query and a learned Evolution-type embedding, followed by layer normalization. Current Map features receive a separate type embedding and normalization before concatenation with Evolution Tokens.

Training Details. We implement RoleDrive in PyTorch and train the model for 25 epochs using distributed data parallelism on four NVIDIA GeForce RTX 4090 GPUs modified to provide 48 GB of memory per GPU. The batch size is 16 per GPU, yielding a global batch size of 64 without

gradient accumulation. We use AdamW with a base learning rate of 2×10^{-4}. Linear warmup occupies the first 10% of the configured optimization schedule, followed by cosine annealing. Training uses mixed precision with random seed 2. Trajectory regression, proposal scoring, and risk-aware Map supervision are optimized jointly. The risk objective combines weighted binary cross-entropy and Dice losses to address foreground–background imbalance and emphasize planning-relevant regions. Its weight follows a cosine decay from 2.0 to 0.2 over the first 15 epochs and remains at 0.2 thereafter. Risk prediction serves as auxiliary supervision and is not supplied as an additional representation to the trajectory planner.

HUGSIM Evaluation Settings. We evaluate both systems on 345 HUGSIM scenarios using the same decision and control intervals of 0.25 s, trajectory and iLQR time steps of 0.5 s, and a controller solver budget of 50 ms. Our DrivoR reproduction drives from the first step, whereas RoleDrive takes over after 1.5 s of initialization with the V0 policy. DrivoR uses four current-frame camera views, while RoleDrive uses three forward-facing views with three historical observations. The systems also differ in camera extrinsic construction and lateral acceleration input: DrivoR uses the original `l2c` transformation and zero lateral acceleration, whereas RoleDrive uses extrinsics aligned with the rendering pipeline and lateral acceleration computed from kinematics. These differences mean that the reported results compare overall system performance rather than isolating the effect of the planning architecture.

B Additional Experimental Results

Representation Similarity and Variance Analysis. Fig. 5 examines the statistical properties of the three representations using the $K = 3$ model on all 18,179 NAVSIM validation scenes. We average the tokens within each representation to obtain one feature vector per scene, enabling comparisons between representations with different token counts. Fig. 5(a) reports pairwise linear CKA, which measures the similarity of feature structures across scenes. Context–Map and Context–Evolution similarities are 0.442 and 0.443, respectively, whereas Map–Evolution similarity is lower at 0.305. These results suggest that the representations share scene information while retaining distinct feature structures. In particular, the lower Map–Evolution similarity is consistent with the distinction between modeling the current spatial configuration and extracting recent changes from aligned observations. The similar Context–Map and Context–Evolution scores indicate comparable scene-level alignment with Context, rather than identical spatial attention or semantic content in the other two branches. To assess the stability of these observations, Fig. 5(c) reports 95% confidence intervals obtained from 2,000 bootstrap resamples of the 214 validation logs. Resampling entire logs preserves the grouping of scenes within each driving sequence. The intervals retain the lower Map–Evolution similarity, indicating that this ordering is stable under log-level resampling for the evaluated model.

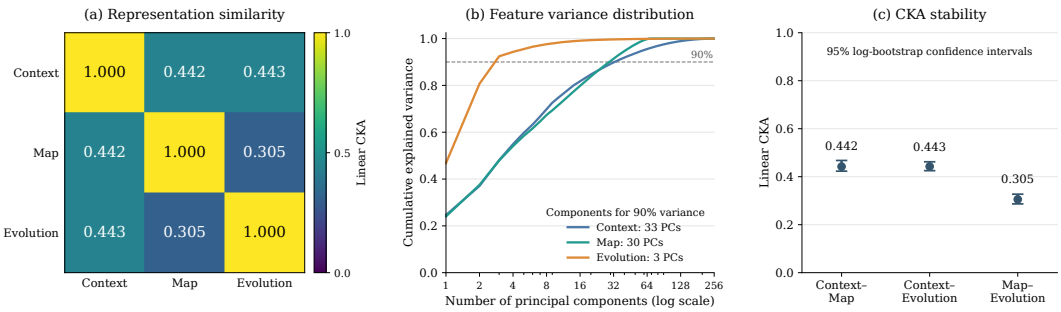


Figure 5: Representation analysis on the full NAVSIM validation set ($K = 3$). (a) Linear CKA between scene-level mean-pooled representations. (b) Cumulative explained variance. (c) CKA with 95% confidence intervals from 2,000 bootstrap resamples of 214 logs.

Fig. 5(b) further analyzes the cumulative explained variance of the centered scene-level features through principal component analysis. Context and Map require 33 and 30 principal components, respectively, to explain 90% of their variance, whereas Evolution reaches the same threshold with

only three components. Variation across scenes is therefore distributed over more feature directions in Context and Map, while Evolution exhibits a substantially more concentrated variance distribution. This difference is consistent with their respective construction: Context aggregates visual information across camera views, Map preserves the current metric scene structure, and Evolution aggregates temporal residuals within regions with valid historical correspondence. However, variance concentration alone does not determine the semantic content or planning importance of a representation. We therefore interpret these curves as evidence of different statistical structures, rather than a ranking of representation quality. Together, the CKA estimates, bootstrap intervals, and variance distributions complement the qualitative visualizations and performance ablations, supporting the intended specialization into global context, local geometry, and recent scene changes.

Risk Region Visualization. Fig. 6 compares current-frame ground-truth occupancy with the risk regions predicted by the auxiliary risk head. Both panels use the same ego-centric coordinates and cover 40 m longitudinally and 16 m laterally. The left panel shows annotated vehicle footprints, while the right panel displays the sigmoid output of the risk head applied to current Map features. Ground-truth outlines are overlaid on the prediction to facilitate spatial comparison. The displayed values are risk scores learned from the constructed supervision, rather than planning attention weights or calibrated collision probabilities. The supervision combines object occupancy with spatially decaying regions around vehicles, pedestrians, and bicycles; consequently, its expected support extends beyond the occupied footprints.

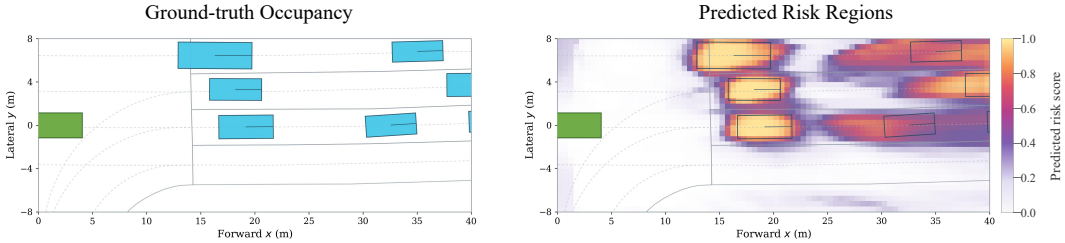
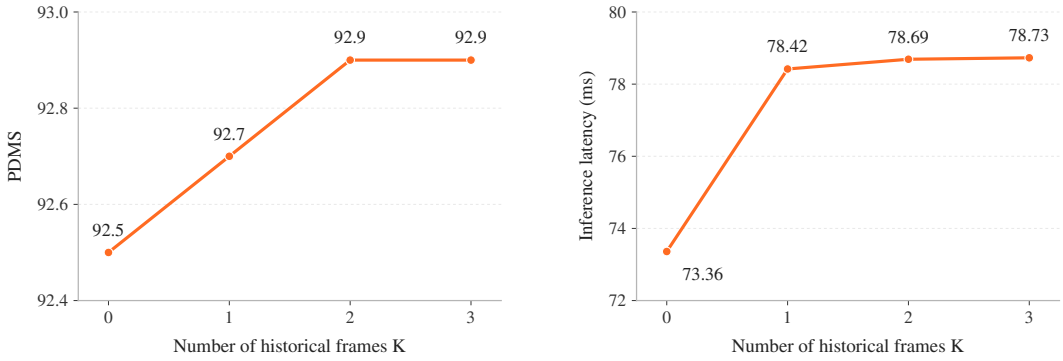


Figure 6: Visualization of risk regions in **RoleDrive**. Left: ground-truth vehicle footprints in the current ego frame. Right: predicted risk scores from current Map features, with ground-truth outlines overlaid for reference. Green denotes the ego vehicle. Scores are displayed on a fixed $[0, 1]$ scale without per-scene normalization. The auxiliary supervision includes decaying regions around vehicles, so risk responses can extend beyond their occupied footprints.

In this scene, several nearby vehicles occupy different lateral positions, and the predicted map exhibits strong responses around their annotated locations. In particular, the three vehicles approximately 15–20 m ahead correspond to distinct high-response regions, illustrating that the Map features retain information about their spatial arrangement. Responses also extend around more distant vehicles, although their boundaries are less localized and some elevated regions connect neighboring objects. Such extensions are partly expected from the spatially decaying supervision, while positional offsets and excess responses remain prediction errors. This example illustrates the spatial information accessible through the auxiliary risk head and complements the risk-supervision ablation in the main paper. The risk map itself is not supplied to the trajectory planner; supervision instead encourages the shared Map features to preserve information about objects and their surrounding regions.

Effect of History Length. Fig. 7(a) evaluates the effect of historical observations on planning performance, with historical frames sampled at 0.5 s intervals. Without history ($K = 0$), RoleDrive achieves 92.5 PDMS. Incorporating one historical frame improves PDMS to 92.7, while two and three historical frames both achieve 92.9. The improvement over the current-only baseline indicates that short-term temporal context provides useful planning cues beyond a single observation. The validation improvement saturates at $K = 2$ at the reported precision: extending the temporal window from 1.0 s to 1.5 s does not further increase the rounded aggregate score. However, matching aggregate scores do not imply identical predictions or equivalent performance across individual scenes and evaluation splits. The test-set results favor $K = 3$, which we retain as the default configuration for the reported test-set evaluation. These results support the benefit of short-term history

while showing that the incremental gain from a longer temporal window depends on the evaluated scenes.



(a) Planning performance

(b) Inference overhead

Figure 7: Effect of history length on planning performance and inference latency in **RoleDrive**. (a) Validation PDMS improves with additional history and plateaus at $K = 2$ at the reported precision. (b) Inference latency increases modestly with additional history. Historical observations are sampled at 0.5 s intervals. Latency is measured with precomputed historical Maps, excluding historical Map preparation, data loading, and input transfer.

Fig. 8 illustrates how temporal context affects planning in slow traffic, turning, and straight driving, shown from top to bottom. All trajectories share the same 4 s prediction horizon, allowing their predicted travel distances to be compared directly. In the slow-traffic scenario, the lead vehicle gradually moves forward. Historical observations provide evidence of this slow movement, and the history-enabled models predict greater progress while maintaining short travel distances: the predicted distance increases from approximately 0.9 m at $K = 0$ to 2.0 m at $K = 3$. In the turning scenario, successive observations reveal the forward movement of a relevant pedestrian, providing additional information about the evolving interaction near the planned path. The history-enabled trajectories extend further into the turn, increasing the predicted travel distance from 4.6 m without history to 15.6 m with three historical frames. In the straight-driving scenario, the nearby vehicle remains nearly stationary after historical observations are aligned to the current ego frame. This consistency provides evidence of its stationary state that is less directly available from a single observation. The history-enabled predictions proceed further along the road, with travel distance increasing from 15.6 m at $K = 0$ to 27.5 m at $K = 3$. Across these examples, temporal observations provide cues about different motion states and are associated with corresponding adjustments in predicted progress. These selected cases complement the aggregate comparison; longer predicted trajectories alone do not establish improved safety or overall planning quality.

Computational Efficiency. Table 6 reports the computational cost and planning performance for different history lengths, while Fig. 7(b) visualizes the corresponding latency trend. We measure inference on a single NVIDIA GeForce RTX 4090 with 48 GB memory, using FP32 and a batch size of 1. Latency is averaged over 250 forward passes on 50 validation scenes after warmup. Historical Maps are precomputed to simulate an available Map memory, so timing includes current-frame encoding, historical Map alignment, temporal residual aggregation, and trajectory planning, but excludes historical Map preparation, data loading, and input transfer. Peak allocated GPU memory includes the resident historical Maps.

Without history, RoleDrive contains 36.25M parameters and requires 73.36 ms per forward pass. Introducing one historical frame increases the parameter count to 36.85M and latency to 78.42 ms. Extending the history to three frames leaves the parameter count almost unchanged at the reported precision and increases latency only slightly to 78.73 ms. As illustrated in Fig. 7(b), most of the measured latency increase occurs when temporal processing is introduced, while extending the history from one to three frames adds only 0.31 ms under the precomputed-history setting. Compared with the baseline without history, the $K = 3$ configuration adds 5.37 ms of latency, approximately

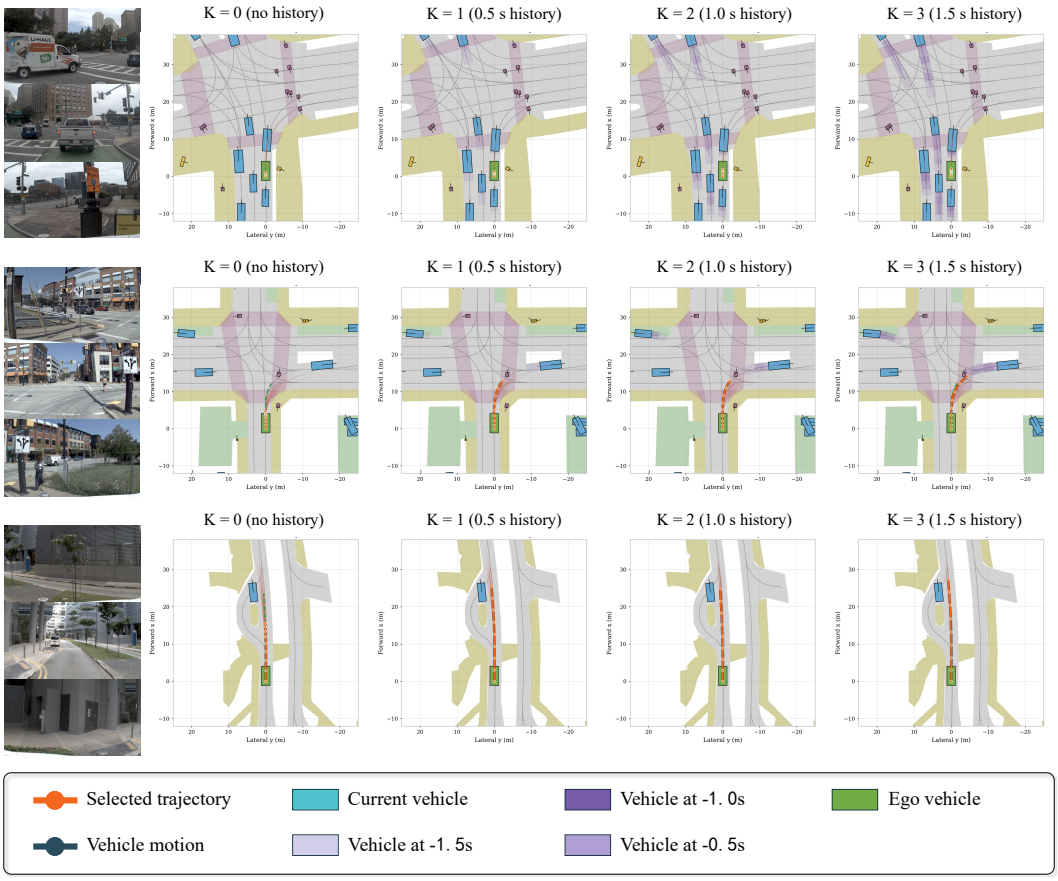


Figure 8: Qualitative comparisons with different history lengths in **RoleDrive**. Rows show slow traffic, turning, and straight driving, from top to bottom. Camera images provide scene context, and columns compare $K = 0, 1, 2, 3$ historical frames sampled at 0.5 s intervals. Orange curves show selected trajectories over the same 4 s prediction horizon. Cyan and translucent purple boxes denote current and historical vehicle positions, respectively; green denotes the ego vehicle.

Table 6: Computational efficiency and planning performance with different history lengths. Latency is measured with precomputed historical Maps.

K	Params (M)	Latency (ms) ↓	Peak memory (GB) ↓	PDMS ↑
0	36.25	73.36	0.406	92.5
1	36.85	78.42	0.412	92.7
2	36.85	78.69	0.414	**92.9**
3	36.85	78.73	0.417	**92.9**

Efficiency is measured on a single RTX 4090 with 48 GB memory using FP32 and batch size 1. Latency is averaged over 250 forward passes on 50 validation scenes after warmup, excluding historical Map preparation, data loading, and input transfer. Peak allocated GPU memory includes the historical Map cache. PDMS values are taken from the history-length experiments.

7.3%, while peak allocated memory increases from 0.406 to 0.417 GB. Together with the PDMS improvement from 92.5 to 92.9, these results show that temporal information can improve planning with modest additional inference overhead when historical Maps are already available.

C LIMITATIONS

Under the available hardware and GPU memory budget, our implementation uses three forward-facing cameras and a bounded local Map region. This configuration prioritizes the area most directly relevant to forward trajectory planning, but provides limited coverage of the rear environment. Extending RoleDrive to more camera views and a larger spatial region would increase the cost of visual encoding and temporal feature processing, requiring further improvements in computational and memory efficiency.

D FUTURE WORK

Future work will investigate broader spatial coverage and more robust temporal comparisons under viewpoint and localization changes. We also plan to evaluate the representations using task-specific probes and controlled interventions to better understand their contributions to planning. Additional training seeds and evaluations with matched controller settings would strengthen the assessment of performance gains. Finally, implementing and evaluating an online Map memory would allow us to measure the full computational cost of maintaining and using historical representations.