CAUSALDRIVE: COUNTERFACTUAL CAUSAL DECOMPOSITION FOR END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

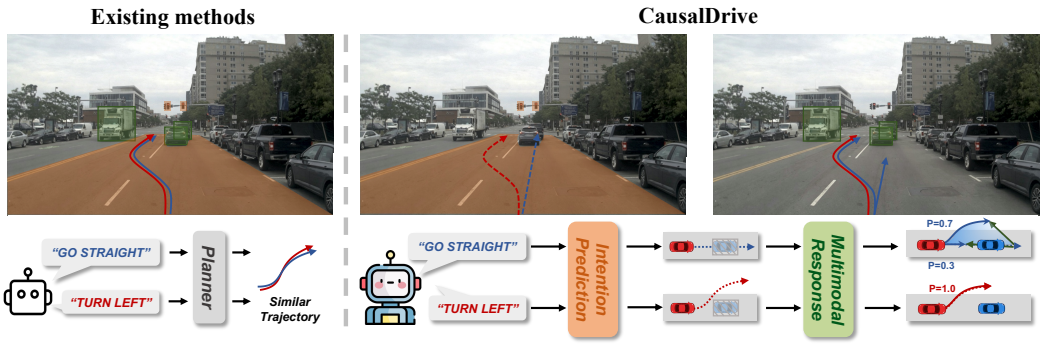


Figure 1: Comparison of existing methods and **CausalDrive**. An intention to turn left and an intention to proceed straight aggressively exhibit the same lane-changing trajectory, making it difficult to identify the factors that truly influence the driving decision. CausalDrive assigns the two sources of driving behavior to separate learning objectives instead of requiring a single expert trajectory to supervise both implicitly.

ABSTRACT

End-to-end autonomous driving frameworks typically learn driving policies by imitating expert trajectories. However, the expert trajectories entangle two factors with distinct roles: a nominal intention that specifies where the vehicle should go, and multimodal driving styles that determine how the intention is executed. Different combinations of the two factors may produce similar trajectories over a short planning horizon, so supervising only the final trajectory provides no explicit signal for attributing the observed behavior to either factor. To resolve this ambiguity, we propose **CausalDrive**, a causal decomposition framework that factorizes driving policy learning into two sequential stages: **deterministic nominal intention prediction** and **probabilistic driving style modeling**. Specifically, we apply an explicit causal intervention that removes style components, constructing counterfactual nominal trajectories from navigation and map context alone; these serve as independent supervision for the proposed **Nominal Intention Planner**. We then develop a diffusion-based **Style Execution Policy** to capture the multimodal residual between nominal and expert trajectories, modeling the conditional probability distribution of driving styles under a given nominal intention. CausalDrive achieves 91.2 PDMS on NAVSIM v1 and 89.5 EPDMS on NAVSIM v2, establishing state-of-the-art performance. Extensive experiments further demonstrate improved overall trajectory quality and clear advantages in challenging driving scenarios.

1 INTRODUCTION

End-to-end autonomous driving aims to learn a direct mapping from scene observations and driving context to future ego trajectories. Early studies established imitation learning as an effective paradigm for driving policy learning (Codevilla et al., 2018; Bansal et al., 2018; Chen et al., 2019;

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.

Prakash et al., 2021; Wu et al., 2022), while recent approaches have incorporated structured scene representations and planning-oriented architectures into unified frameworks (Hu et al., 2022; 2023; Jiang et al., 2023; Sun et al., 2025). Despite their strong performance, most existing methods treat the expert trajectory as a monolithic supervision target, even though it is jointly determined by heterogeneous factors that play fundamentally different roles in driving decision making.

In particular, driving behavior consists of two factors: a nominal trajectory that specifies where the vehicle should go, and multimodal driving styles that determine how this intention is executed. Driving styles refer to ego interaction responses to surrounding agents, such as following, yielding, or overtaking, and are inherently multimodal. Given the navigation command, map topology, and ego state, the vehicle is expected to adhere to an intention-consistent nominal trajectory (Paden et al., 2016; Schwarting et al., 2018). In contrast, how this intention is executed varies with driving style and is inherently multimodal, such as cautiously yielding or assertively overtaking in response to surrounding agents while preserving the same nominal intention. Conventional trajectory-supervised methods cannot determine whether the observed trajectory is primarily driven by the nominal intention or by the driving style. As shown in Figure 1, an intention to turn left and an intention to proceed straight aggressively exhibit the same lane-changing trajectory. Consequently, supervising only the final trajectory provides limited guidance regarding the respective contributions of the two factors. It may also encourage the policy to fit spurious statistical correlations rather than learn the underlying causal decision structure.

Existing approaches do not explicitly resolve this ambiguity. Conventional end-to-end planners directly regress the expert trajectory, thereby coupling nominal intention and multimodal driving styles within a single prediction target. These approaches (Zheng et al., 2024; Liao et al., 2025; Zou et al., 2025) leverage the modeling capacity of generative models to formulate trajectory as a holistic multimodal distribution, yet they typically fall short of further disentangling the underlying sources of such multimodality. Unlike causality-based methods (Zheng et al., 2026; Tang et al., 2026; Moon et al., 2026) that mitigate spurious correlations at the representation or training level, we explicitly factorize the supervision target itself into a deterministic nominal intention and probabilistic driving styles, assigning each a dedicated learning objective.

To this end, we introduce **CausalDrive**, a causal decomposition framework that models nominal intention deterministically and driving style probabilistically. First, we construct a counterfactual nominal trajectory generator from navigation commands, map topology, and ego state. The generated trajectory represents how the ego vehicle would proceed under the specified nominal intention. These trajectories provide supervision independent of ground-truth expert trajectory for a **Nominal Intention Planner (NIP)**, which predicts an intention-consistent nominal trajectory. Second, conditioned on the nominal trajectory and surrounding agents, a diffusion-based **Style Execution Policy (SEP)** captures the multimodal residual between nominal and expert trajectories, modeling the conditional probability distribution of driving styles under a given nominal intention. This two-stage design assigns the two sources of driving behavior to separate learning objectives instead of requiring a single expert trajectory to supervise both implicitly. This allows for more interpretable and effective learning of driving policies. Extensive experiments validate the effectiveness of CausalDrive. On NAVSIM v1, CausalDrive achieves a PDMS of 91.2 and an EPDMS of 89.5 on the more challenging NAVSIM v2. Further analyses show that CausalDrive generates consistently high-quality trajectory candidates and provides clear advantages in complex interactive traffic scenarios.

Our contributions are summarized as follows:

- We identify the causal ambiguity in expert trajectory-only supervision: different combinations of nominal intention and driving styles may produce similar trajectories over a short horizon, leaving the observed behavior unattributable to either factor.
- We construct deterministic counterfactual nominal trajectories from navigation commands, map topology, and the ego state, without style components. These trajectories provide supervision independent of ground-truth expert trajectory for learning nominal intentions.
- We propose **CausalDrive**, which combines a **Nominal Intention Planner** with a diffusion-based **Style Execution Policy** to separately model deterministic nominal behavior and multimodal driving styles.
- CausalDrive achieves 91.2 PDMS on NAVSIM v1 and 89.5 EPDMS on NAVSIM v2, with clear advantages in challenging interactive driving scenarios.

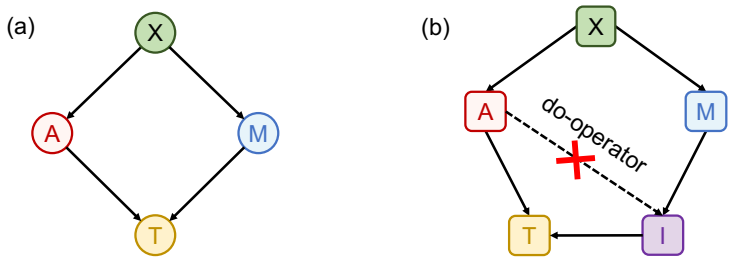


Figure 2: Comparison of causal graphs for trajectory prediction. (a) In conventional trajectory-only learning, the scene representation X is encoded into agent and map representations A and M, which jointly determine the executed trajectory T. (b) CausalDrive introduces a nominal trajectory I and excludes the direct influence of A on I; T is determined by I and the driving style induced by A.

2 Related Work

2.1 End-to-End Autonomous Driving

End-to-end autonomous driving directly maps sensor observations and driving context to future ego trajectories, allowing perception, prediction, and planning to be optimized toward a unified objective. Early approaches demonstrated the effectiveness of imitation learning of driving policy and improved robustness through synthesized perturbations or privileged supervision (Bansal et al., 2018; Chen et al., 2019). More recent systems increasingly incorporate structured scene representations into the end-to-end pipeline. ST-P3 (Hu et al., 2022) learns spatial-temporal features for vision-based planning, UniAD (Hu et al., 2023) unifies perception and prediction tasks around a planning objective, VAD (Jiang et al., 2023) represents map elements and agent motion as vectors, and SparseDrive (Sun et al., 2025) further develops a sparse, object-centric driving framework. Despite these architectural advances, most methods supervise the expert trajectory as a whole, providing limited guidance on the respective contributions of heterogeneous decision factors to the final trajectory. In contrast, CausalDrive separately supervises intention-consistent nominal trajectory and multimodal driving styles, allowing for more interpretable and effective learning of driving policies.

2.2 Generative and Diffusion-Based Planning

Generative planners capture the inherent uncertainty of driving by producing diverse trajectory hypotheses rather than a single deterministic prediction. GenAD (Zheng et al., 2024) jointly generates agent motion and ego plans through latent generative modeling, whereas DiffusionDrive (Liao et al., 2025) employs truncated diffusion for efficient multimodal planning. DiffusionDriveV2 (Zou et al., 2025) further introduces reinforcement learning to improve plan quality while preserving behavioral diversity. Rather than generating absolute trajectories, ResAD (Zheng et al., 2026) models normalized residuals from inertial reference, simplifying the trajectory-generation objective. While these methods capture diverse driving trajectories, they do not explicitly identify driving style as the source of this multimodality; instead, they directly predict the complete trajectory. CausalDrive predict the deterministic nominal intention with a regression model and predict the multimodal driving styles with a diffusion model.

3 Method

3.1 Preliminary

Problem formulation. Each planning step is characterized by a navigation command $\mathcal{N}$, surrounding agents $\mathcal{A}$, map topology $\mathcal{M}$, and ego state $\mathcal{E}$. Specifically, $\mathcal{N}$ denotes a route-level driving command; $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^{N_a}$ represents the states of nearby agents; $\mathcal{M}$ describes the local road topology and traffic-rule elements; and $\mathcal{E}$ contains the current ego state and its motion history. Figure 2 summarizes the corresponding causal structures. For clarity, X denotes the scene representation, while A and M are the extracted agent and map representations, respectively. In conventional trajectory-only

learning (Figure 2 (a)), A and M jointly determine the executed trajectory T, without an explicit attribution mechanism. CausalDrive (Figure 2 (b)) introduces the nominal trajectory I. The crossed dashed arrow indicates that the direct influence of A on I is excluded. Unlike Figure 2 (a), where A and M jointly determine T, T is here determined by I and the driving style induced by A. The ground-truth ego trajectory is denoted by $\boldsymbol{\tau}^{\mathrm{gt}} = \{\mathbf{p}_t^{\mathrm{gt}}\}_{t=1}^{H}$, where H is the planning horizon and $\mathbf{p}_t^{\mathrm{gt}}$ is the future ego pose at step t. Conventional behavior cloning trains a driving policy π by minimizing

$$\mathcal{L}_{\mathrm{BC}} = \mathbb{E}\left[\ell\big(\pi(\mathcal{N}, \mathcal{A}, \mathcal{M}, \mathcal{E}), \boldsymbol{\tau}^{\mathrm{gt}}\big)\right]. \tag{1}$$

However, these inputs play fundamentally different roles in trajectory prediction. We assume that nominal intention is primarily determined by navigation command $\mathcal{N}$, map topology $\mathcal{M}$, and ego state $\mathcal{E}$, while surrounding agents $\mathcal{A}$ primarily affect how this intention is executed. Under this assumption, trajectory generation can be conceptually represented as:

$$(\mathcal{N}, \mathcal{M}, \mathcal{E}) \rightarrow \boldsymbol{\tau}^N, \qquad (\boldsymbol{\tau}^N, \mathcal{A}) \rightarrow \boldsymbol{\tau}, \tag{2}$$

where $\boldsymbol{\tau}^N$ denotes nominal trajectory, and $\boldsymbol{\tau}$ denotes the executed trajectory. This formulation motivates a counterfactual view of the nominal trajectory: if surrounding agents were removed, the resulting trajectory would approximate the nominal trajectory, namely:

$$\boldsymbol{\tau}^N \approx \boldsymbol{\tau} \mid do(\mathcal{A} = \emptyset). \tag{3}$$

Framework overview. CausalDrive reflects this asymmetry by modeling the nominal trajectory $\hat{\boldsymbol{\tau}}^N$ deterministically and the driving style $\boldsymbol{\delta}^{\mathrm{S}}$ probabilistically:

$$\hat{\boldsymbol{\tau}}^N = f_\theta(\mathcal{N}, \mathcal{M}, \mathcal{E}), \qquad \boldsymbol{\delta}^{\mathrm{S}} \sim p_\phi\big(\boldsymbol{\delta} \mid \mathcal{A}, \hat{\boldsymbol{\tau}}^N\big), \tag{4}$$

where f_θ is the Nominal Intention Planner (NIP) and p_ϕ is the Style Execution Policy (SEP). The final trajectory is composed as

$$\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}}^N + \boldsymbol{\delta}^{\mathrm{S}}. \tag{5}$$

As shown in Figure 3, the NIP is supervised by nominal trajectories constructed in section 3.2, and learns a deterministic and intention-consistent trajectory. Conditioned on this intention trajectory and the surrounding agents, the SEP models how the intention is executed, capturing multimodal driving styles. This design assigns the two sources of driving behavior to separate learning objectives instead of requiring a single expert trajectory to supervise both implicitly.

3.2 Counterfactual Nominal Trajectory Construction

Since the expert trajectory is inherently entangled with driving style, we do not use it to define the nominal intention target. Instead, we construct a nominal trajectory generator $\mathcal{R}$ from the navigation command, map topology and ego state through the following three steps.

Minimum-deviation path planning. Given the current ego pose and the navigation command $\mathcal{N}$, we select a command-consistent, on-route reference lane from the global map topology. We then construct a feasible path γ_n that starts from the current pose and minimizes lateral deviation from this reference centerline, while ensuring destination reachability with minimal driving maneuvers.

Geometric construction. We express γ_n in the ego-centric coordinate system and parameterize it by arc length s. The resulting continuous path $\tilde{\gamma}_n(s)$ fixes where the vehicle can proceed, independent of surrounding agents and driving style; its temporal discretization is determined by the longitudinal profile constructed in the next step.

Kinematic and rule refinement. Along $\tilde{\gamma}_n$, we interpolate a speed profile that starts from the current ego speed and remains consistent with the speed limit, traffic rules, and local curvature. This profile determines the longitudinal progress $\{s_t\}_{t=1}^{H}$ over the planning horizon. The nominal target is then obtained by sampling $\tilde{\gamma}_n$ at the corresponding progress values:

$$\bar{\boldsymbol{\tau}}^N = \mathcal{R}(\mathcal{N}, \mathcal{M}, \mathcal{E}) = \{\tilde{\gamma}_n(s_t)\}_{t=1}^{H}. \tag{6}$$

The construction uses neither the surrounding agents $\mathcal{A}$ nor the expert trajectory $\boldsymbol{\tau}^{\mathrm{gt}}$. The resulting target therefore depends solely on $(\mathcal{N}, \mathcal{M}, \mathcal{E})$: it is a deterministic, style-free counterfactual, rather than a target inherited from the expert demonstration.

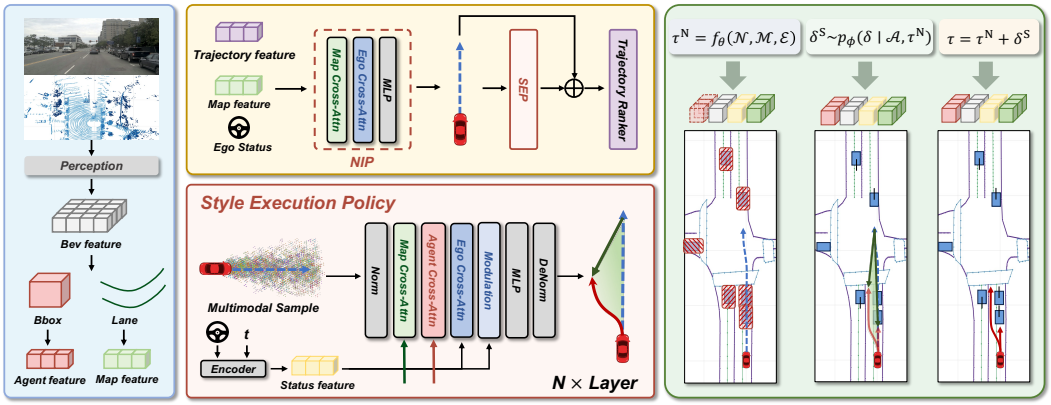


Figure 3: Overview of CausalDrive. The NIP predicts a counterfactual nominal trajectory from the map feature and ego state. Conditioned on this trajectory and the complete scene context, the SEP samples multimodal style residuals. Adding each residual to the nominal trajectory yields a set of diverse candidates, from which the trajectory ranker selects the executable trajectory.

3.3 Architecture

As shown in Figure 3, CausalDrive takes multi-view images and LiDAR point clouds as input, which are fused by a Transfuser-style encoder (Chitta et al., 2023). We first use a decoder with cross-attention in the NIP to fuse the encoded map representation and the ego condition, and predict a nominal trajectory $\hat{\boldsymbol{\tau}}^N$. Diffusion decoders in the SEP then additionally fuse the encoded agent features via cross-attention and use $\hat{\boldsymbol{\tau}}^N$ as a condition to guide residual denoising. Finally, each residual is added to the shared $\hat{\boldsymbol{\tau}}^N$, and the ranker selects the final plan.

Nominal Intention Planner. The NIP is designed for deterministic nominal intention: it predicts how the ego vehicle would proceed under the nominal intention alone, rather than imitating $\boldsymbol{\tau}^{\mathrm{gt}}$, which couples intention with driving style. A decoder fuses the encoded map features and the ego condition (driving command and current state) via cross-attention, predicting a counterfactual nominal trajectory $\hat{\boldsymbol{\tau}}^N$. The NIP is supervised by the counterfactual target $\bar{\boldsymbol{\tau}}^N$ in Eq. 6. Crucially, the NIP branch is kept completely free of agent features, ensuring that no interaction relevant driving styles leaks into the nominal trajectory.

Style Execution Policy. The SEP models uncertainty over the multimodal residual between the nominal and expert trajectories. A scene typically admits multiple valid executions of the same nominal intention, of which the expert trajectory is only one realization. If the fully coupled expert trajectory supervises the entire plan, these modes remain entangled with the nominal intention. They collapse into a compromise path that matches none of them and often approaches surrounding agents too closely. Once driving style is decoupled as a residual from $\hat{\boldsymbol{\tau}}^N$, the nominal trajectory remains intention-consistent, while multimodal driving styles, such as following, yielding, and overtaking, are carried by the residual. For each sample, this residual target is

$$\boldsymbol{\delta}_0 = \boldsymbol{\tau}^{\mathrm{gt}} - \mathrm{sg}[\hat{\boldsymbol{\tau}}^N], \tag{7}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator, which prevents residual learning from updating the NIP. Since a scene may admit multiple valid executions, we model its conditional distribution with diffusion rather than regressing a single deterministic correction. Crucially, the SEP branch is the sole pathway through which agent information is introduced into the final trajectory.

For diffusion step $s \in \{1, \ldots, S\}$, the forward process gradually adds Gaussian noise to $\boldsymbol{\delta}_0$:

$$q(\boldsymbol{\delta}_s \mid \boldsymbol{\delta}_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_s}\,\boldsymbol{\delta}_0,\, (1-\bar{\alpha}_s)\mathbf{I}\big), \tag{8}$$

where $\bar{\alpha}_s$ is determined by the noise schedule. Equivalently, $\boldsymbol{\delta}_s = \sqrt{\bar{\alpha}_s}\boldsymbol{\delta}_0 + \sqrt{1-\bar{\alpha}_s}\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. The diffusion decoder ϵ_ϕ fuses the encoded map, surrounding agent, and ego state via cross-attention, and is conditioned on $\mathrm{sg}[\hat{\boldsymbol{\tau}}^N]$. We train with the noise-prediction objective

$$\mathcal{L}_{\mathrm{SEP}} = \mathbb{E}_{\boldsymbol{\delta}_0, s, \boldsymbol{\epsilon}}\Big[\big\|\boldsymbol{\epsilon} - \epsilon_\phi(\boldsymbol{\delta}_s, s, \mathbf{c})\big\|_2^2\Big], \tag{9}$$

where $\mathbf{c}$ collects these inputs. At inference, reverse diffusion yields K residuals $\{\hat{\boldsymbol{\delta}}_k^{\mathrm{S}}\}_{k=1}^{K}$. Adding each to the shared nominal trajectory gives

$$\hat{\boldsymbol{\tau}}_k = \hat{\boldsymbol{\tau}}^N + \hat{\boldsymbol{\delta}}_k^{\mathrm{S}}, \qquad k = 1, \ldots, K. \tag{10}$$

All candidates therefore share $\hat{\boldsymbol{\tau}}^N$ and differ only in the style residual.

Training and Inference. Training is conducted in two stages. In Stage I, the backbone and the NIP are jointly trained with an ℓ_1 loss on $\bar{\boldsymbol{\tau}}^N$, the ground-truth counterfactual nominal trajectory constructed by $\mathcal{R}$. In Stage II, we jointly train the NIP and the SEP. The SEP is trained with a DDPM (Ho et al., 2020) of S=1000 diffusion steps. The ground-truth nominal trajectory $\bar{\boldsymbol{\tau}}^N$ replaces the NIP prediction $\hat{\boldsymbol{\tau}}^N$ as the conditioning input to the SEP with a probability of 10% to mitigate the exposure bias. At inference, two-step DDIM (Song et al., 2021) draws $K = 60$ candidate trajectories. Coarse ranking reduces them to 32, and fine ranking scores the rest using benchmark-specific metric predictions aggregated into PDMS/EPDMS.

Multimodal Trajectory Ranker. The K residuals yield diverse candidates that are not equally suitable for execution. We therefore apply a coarse-to-fine ranker after Eq. 10. A lightweight coarse head scores all candidates and retains a subset $\mathcal{C}_{\mathrm{fine}}$ at inference. A fine head then predicts the benchmark subscores of each candidate and aggregates them into the official driving score r_ω (PDMS or EPDMS). We supervise r_ω on all K candidates with the simulator-computed target s_k:

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{K} \sum_{k=1}^{K} \left\| r_\omega(\hat{\boldsymbol{\tau}}_k) - s_k \right\|_1. \tag{11}$$

At inference, the final trajectory is the candidate in $\mathcal{C}_{\mathrm{fine}}$ with the highest r_ω.

4 Experiments

4.1 Benchmark

We evaluate on the NAVSIM v1 (Dauner et al., 2024) and NAVSIM v2 (Cao et al., 2025) benchmarks using their official metrics, PDMS and EPDMS, respectively. NAVSIM provides a suite of real-world, planning-centric driving scenarios built upon OpenScene (Contributors, 2023), which is a compact version of the extensive nuPlan (Caesar et al., 2021) dataset. The dataset is split into *navtrain* (1,192 training scenes) and *navtest* (136 evaluation scenes).

4.2 Implementation Details

For a fair comparison with prior NAVSIM methods, we follow TransFuser (Chitta et al., 2023) and use a ResNet-34 backbone (He et al., 2016) with camera–LiDAR fusion. We train from scratch on *navtrain* and evaluate on *navtest*, using the same architecture for NAVSIM v1 and v2. The planner outputs 8 waypoints at $0.5\,\mathrm{s}$ intervals over a $4\,\mathrm{s}$ planning horizon. The model is trained with AdamW for 100 epochs, using a learning rate of 6×10^{-4} and a global batch size of 512 on 8 NVIDIA H20 GPUs.

4.3 Main Results

Results on NAVSIM v1. As shown in Table 1, CausalDrive achieves 91.2 PDMS on the NAVSIM v1 *navtest* split, outperforming all prior methods, including ResAD and Hydra-MDP with a V2-99 backbone (90.6 and 90.3). CausalDrive also achieves the best DAC and TTC at 98.1 and 95.9, respectively. These gains follow from the decomposition. The NIP predicts where the vehicle should go and is supervised by a style-free nominal trajectory. It therefore favors trajectories inside the drivable area, which mainly accounts for the gain in DAC. Conditioned on that intention and surrounding agents, the SEP models following, yielding, and overtaking as multimodal driving styles, which accounts for the gain in TTC.

Results on NAVSIM v2. As shown in Table 2, CausalDrive achieves an EPDMS of 89.5 on NAVSIM v2 with a ResNet-34 backbone, ranking first among all methods. CausalDrive also achieves the best NC (98.6), TTC (98.1), DDC (99.5), and TL (99.9) across all compared methods, together with the best DAC (97.9) among ResNet-34 methods. These subscores indicate stronger collision avoidance and rule compliance.

Table 1: Comparison on NAVSIM v1 *navtest* with closed-loop metrics. Best results are in **bold** and second best are underlined.

Method	Img. Backbone	NC ↑	DAC ↑	TTC ↑	Comf. ↑	EP ↑	PDMS ↑
VADv2 (Chen et al., 2024)	ResNet-34	97.2	89.1	91.6	**100**	76.0	80.9
UniAD (Hu et al., 2023)	ResNet-34	97.8	91.9	92.9	**100**	78.8	83.4
Transfuser (Chitta et al., 2023)	ResNet-34	97.7	92.8	92.8	**100**	79.2	84.0
PARA-Drive (Weng et al., 2024)	ResNet-34	97.9	92.4	93.0	99.8	79.3	84.0
DRAMA (Yuan et al., 2024)	ResNet-34	98.0	93.1	94.8	**100**	80.1	85.5
GoalFlow (Xing et al., 2025)	ResNet-34	98.3	93.8	94.3	**100**	79.8	85.7
Hydra-MDP (Li et al., 2024)	ResNet-34	98.3	96.0	94.6	**100**	78.7	86.5
Hydra-MDP++ (Li et al., 2025a)	ResNet-34	97.6	96.0	93.1	**100**	80.4	86.6
ARTEMIS (Feng et al., 2026)	ResNet-34	98.3	95.1	94.3	**100**	81.4	87.0
FUMP (Liu et al., 2025)	ResNet-34	98.1	96.2	94.2	**100**	82.0	87.8
DiffusionDrive (Liao et al., 2025)	ResNet-34	98.2	96.2	94.7	**100**	82.2	88.1
WoTE (Li et al., 2025b)	ResNet-34	98.5	96.8	94.9	99.9	81.9	88.3
DIVER (Song et al., 2025)	ResNet-34	98.5	96.5	94.9	**100**	82.6	88.3
ResAD (Zheng et al., 2026)	ResNet-34	98.0	97.5	94.1	**100**	83.3	88.8
DriveSuprim (Yao et al., 2026)	ResNet-34	97.8	97.3	93.6	**100**	86.7	89.9
DiffusionDriveV2 (Zou et al., 2025)	ResNet-34	98.3	97.9	94.8	99.9	**87.5**	**91.2**
Hydra-MDP (Li et al., 2024)	V2-99	98.4	97.8	93.9	**100**	86.5	90.3
ResAD (Zheng et al., 2026)	V2-99	**98.9**	97.8	94.9	**100**	87.0	90.6
CausalDrive (Ours)	ResNet-34	98.4	**98.1**	**95.9**	**100**	86.0	**91.2**

Table 2: Comparison on NAVSIM v2 *navtest* with extended closed-loop metrics. EPDMS* denotes results computed using the original NAVSIM v2 evaluation code, in which human-behavior filtering was not applied when calculating EPDMS. This version is reported for fair comparison with prior works. EPDMS denotes results computed using the updated official implementation after the issue was resolved.

Method	Img. Backbone	NC ↑	DAC ↑	DDC ↑	TL ↑	EP ↑	TTC ↑	LK ↑	HC ↑	EC ↑	EPDMS* ↑	EPDMS ↑
Ego Status MLP	ResNet-34	93.1	77.9	92.7	99.6	86.0	91.5	89.4	**98.3**	85.4	64.0	–
Transfuser (Chitta et al., 2023)	ResNet-34	96.9	89.9	97.8	99.7	87.1	95.4	92.7	**98.3**	87.2	76.7	–
Hydra-MDP++ (Li et al., 2025a)	ResNet-34	97.2	97.5	99.4	99.6	83.1	96.5	94.4	98.2	70.9	81.4	–
DriveSuprim (Yao et al., 2026)	ResNet-34	97.5	96.5	99.4	99.6	88.4	96.6	95.5	**98.3**	77.0	83.1	–
ARTEMIS (Feng et al., 2026)	ResNet-34	98.3	95.1	98.6	99.8	81.5	97.4	96.5	**98.3**	**98.3**	83.1	–
DiffusionDriveV2 (Zou et al., 2025)	ResNet-34	97.7	96.6	99.2	99.8	88.9	97.2	96.0	97.8	91.0	85.5	87.5
Hydra-MDP++ (Li et al., 2025a)	V2-99	98.4	**98.0**	99.4	99.8	87.5	97.7	95.3	**98.3**	77.4	85.1	–
DriveSuprim (Yao et al., 2026)	V2-99	97.8	97.9	**99.5**	**99.9**	**90.6**	97.1	**96.6**	**98.3**	77.9	**86.0**	–
CausalDrive (Ours)	ResNet-34	**98.6**	97.9	**99.5**	**99.9**	87.2	**98.1**	96.2	98.2	83.3	**86.0**	**89.5**

Performance in Challenging Interactive Scenarios. To assess robustness on a harder distribution, we evaluate on *navhard*, a challenging NAVSIM v2 split with unprotected turns, dense traffic, and related cases. Stage 2 is camera-only; for a fair comparison with camera–LiDAR methods, we report Stage-1 results only. As shown in Table 3, CausalDrive attains the best Stage-1 EPDMS^* (71.5) and EPDMS (78.2). Relative to DiffusionDriveV2, the EPDMS margin grows from $+2.0$ on *navtest* to $+6.9$ on navhard, indicating that the decomposition helps mainly when nominal intention and driving style are jointly difficult, rather than from trajectory fitting on easy scenes. We remain strongest on DAC (92.0), DDC (99.1), LK (96.2), and TTC (96.2).

Quality. We further report Top-K PDMS to measure the quality of the K sampled candidates, not only the executed plan. As shown in Table 4, Top-K PDMS remains high as K increases, indicating that high-quality behavior is shared across sampled modes rather than concentrated in the executed trajectory. This is consistent with the decomposition: every candidate shares the same nominal intention, and the sampled modes differ only in driving style.

Qualitative Comparison. As shown in Figure 4, we compare DiffusionDrive, DiffusionDriveV2, and CausalDrive in a sharp turn and in dense traffic. Both baselines directly imitate the expert trajectory, confusing where the vehicle should go and how to execute the intention. They still produce straight-going or off-road modes in the turn and miss follow or yield responses in dense traffic. In the sharp turn, CausalDrive instead keeps a single style-free nominal trajectory that follows the re-

Table 3: Stage-1 evaluation on NAVSIM v2 *navhard*. We omit the camera-only Stage 2 to compare fairly with camera–LiDAR methods.

Method	NC↑	DAC↑	DDC↑	TL↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS* ↑	EPDMS↑
DiffusionDrive (Liao et al., 2025)	**96.1**	86.0	98.6	**99.3**	83.9	94.8	95.7	**97.5**	**78.6**	66.4	73.4
DiffusionDriveV2 (Zou et al., 2025)	95.8	90.8	96.7	**99.3**	**86.0**	95.1	93.3	93.7	29.3	64.8	71.3
CausalDrive (Ours)	96.0	**92.0**	**99.1**	**99.3**	83.5	**96.2**	**96.2**	97.3	70.7	**71.5**	**78.2**

Table 4: Comparison of Top-K PDMS of sampled candidates on *navtest*. PDMS@K denotes the PDMS score evaluated on the Top-K ranked trajectories.

Method	PDMS@1	PDMS@5	PDMS@10	PDMS@15	PDMS@20
DiffusionDrive (Liao et al., 2025)	93.5	84.3	75.3	73.3	60.4
DiffusionDriveV2 (Zou et al., 2025)	94.9	91.1	84.4	76.6	63.9
CausalDrive (Ours)	**95.5**	**95.0**	**94.6**	**94.4**	**94.1**

Table 5: Ablation of each component on NAVSIM v1.

NIP	SEP	Ranker	NC↑	DAC↑	TTC↑	Comf. ↑	EP↑	PDMS↑
			98.2	96.2	94.7	100	82.2	88.1
✓			94.6	97.3	70.7	100	82.5	78.7
✓	✓		98.3	97.5	95.7	100	84.4	89.9
✓	✓	✓	98.4	98.1	95.9	100	86.0	91.2

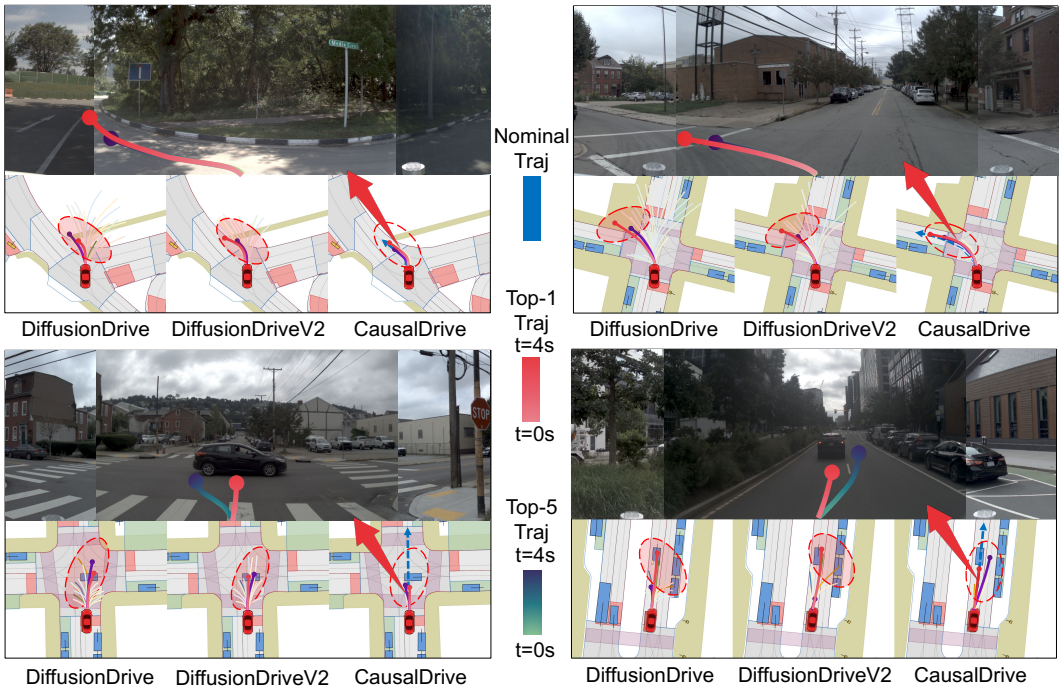


Figure 4: Qualitative comparison on a sharp turn and in dense traffic. DiffusionDrive and DiffusionDriveV2 still produce straight-going or off-road modes in the turn and fail to adapt to surrounding agents in dense traffic; CausalDrive stays on-route through the turn and follows, yields, or overtakes in dense traffic.

quired curvature and stays on-route and inside the drivable area. In dense traffic, conditioned on this nominal trajectory and, additionally, on surrounding agents, the SEP models following, yielding, and overtaking as multimodal driving styles. CausalDrive therefore produces a plan that follows the nominal intention while executing it in the driving style required by traffic.

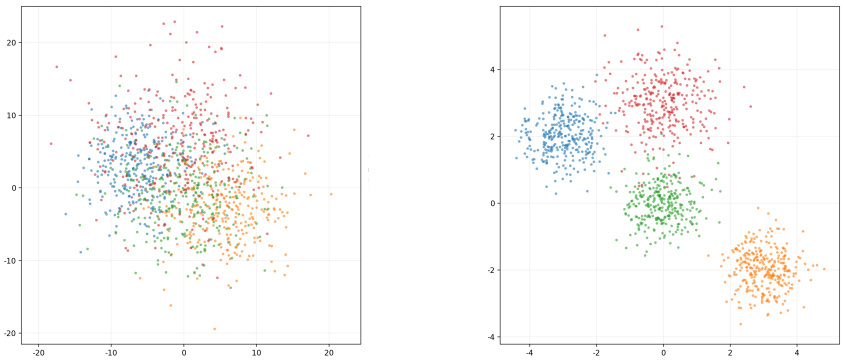


Figure 5: t-SNE of final trajectories from DiffusionDrive (left) and interaction residuals from CausalDrive (right). Different colors denote cluster result of CausalDrive. CausalDrive yields compact and separable clusters with high intra-category consistency, whereas DiffusionDrive trajectories form a diffuse distribution.

4.4 Ablation Studies and Analysis

Component Analysis. To validate the efficacy of each proposed component in CausalDrive, we conduct a comprehensive ablation study, with the results detailed in Table 5. The baseline is DiffusionDrive, which takes the expert trajectory as its only supervision and therefore entangles the nominal intention with the driving style, reaching 88.1 PDMS. Attaching the NIP to the DiffusionDrive backbone, supervised by a driving-style-free counterfactual and without surrounding agents, favors a trajectory inside the drivable area, raising DAC from 96.2 to 97.3. However, those agents are unobserved, so the plan cannot avoid them: TTC falls from 94.7 to 70.7, NC falls from 98.2 to 94.6, and PDMS falls to 78.7. Conditioned on this fixed nominal trajectory and on surrounding agents, the SEP models following, yielding, and overtaking as driving styles for interacting with surrounding traffic. TTC then recovers from 70.7 to 95.7 and NC from 94.6 to 98.3, with DAC at 97.5 and ego progress at 84.4. NIP and SEP are complementary: the NIP specifies the nominal intention, while the SEP supplies the driving styles required to execute it in traffic, raising PDMS to 89.9, above DiffusionDrive. The ranker then selects among these plausible multimodal driving styles, further raising PDMS to 91.2, DAC to 98.1, TTC to 95.9, and ego progress to 86.0.

Representation Analysis of Modeling Paradigms. To compare the representations induced by the two trajectory modeling paradigms, we apply t-SNE to the complete trajectories generated by DiffusionDrive and the driving-style outputs of CausalDrive across 500 selected complex interaction scenarios. As shown in Figure 5, the DiffusionDrive trajectories form a diffuse distribution. This entanglement arises because directly generating complete trajectories requires encoding the nominal intention and driving styles in a entangled distribution. In contrast, the CausalDrive style outputs form compact and separable structures with high intra-category consistency. This provides qualitative evidence that the SEP learns distinct driving styles conditioned on a fixed nominal intention, rather than an entangled distribution over complete trajectories.

5 Conclusion

In this work, we show that expert trajectories entangle a largely deterministic nominal intention with multimodal driving styles, leaving expert trajectory supervision methods confounded in causal learning. We propose CausalDrive to separate them: the NIP is supervised by counterfactual trajectories constructed from the driving command, map topology, and ego state, and predicts a style-independent nominal trajectory; conditioned on that trajectory and surrounding agents, the SEP models a multimodal residual on that intention, capturing diverse styles of executing it. Comprehensive experiments and qualitative comparisons show that CausalDrive achieves competitive performance, providing a new perspective for end-to-end autonomous driving system.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. The study involves neither human-subject research nor animal experimentation. All experiments are conducted on the publicly released NAVSIM benchmark and follow its license and usage terms. We collect no additional data and access no personally identifiable information, and the experiments raise no further privacy or security concerns. We have also taken care to avoid bias and discriminatory outcomes in the research process. The authors accept responsibility for the transparency and integrity of this work.

REPRODUCIBILITY STATEMENT

We have taken the following steps to make the reported results reproducible. The training procedure, model configurations, and hardware setup are specified, and the method section describes the proposed approach, including counterfactual nominal-trajectory construction, the Nominal Intention Planner, and the diffusion-based Style Execution Policy. All results are evaluated on the publicly available NAVSIM benchmark, covering both NAVSIM v1 and NAVSIM v2, under their standard protocols. These descriptions should allow other researchers to replicate our experiments and build on this work.

REFERENCES

Mayank Bansal, Alex Krizhevsky, and Abhijit S. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *ArXiv*, abs/1812.03079, 2018.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric M. Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *ArXiv*, abs/2106.11810, 2021.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Conference on Robot Learning (CoRL)*, 2025.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning (CoRL)*, 2019.

Shaoyu Chen, Bo Jiang, Hao Gao, Ben-Cheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *ArXiv*, abs/2402.13243, 2024.

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.

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation*, pp. 4693–4700, 2018.

OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Vancouver, Canada*, pp. 18–22, 2023.

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 *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Renju Feng, Ning Xi, Duanfeng Chu, Rukang Wang, Zejian Deng, Anzheng Wang, Liping Lu, Jinxiang Wang, and Yanjun Huang. Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving. *IEEE Robotics and Automation Letters*, 11(1):226–233, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

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*, 2022.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Kailin Li, Zhenxin Li, Shiyi Lan, Yuan Xie, Zhizhong Zhang, Jiayi Liu, Zuxuan Wu, Zhiding Yu, and Jose M Alvarez. Hydra-mdp++: Advancing end-to-end driving via expert-guided hydra-distillation. *arXiv preprint arXiv:2503.12820*, 2025a.

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

Zhen-Xin Li, Kailin Li, Shihao Wang, Shi-Yi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, Yu-Gang Jiang, and José M. Álvarez. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *ArXiv*, abs/2406.06978, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Lin Liu, Caiyan Jia, Ziying Song, Hongyu Pan, Bencheng Liao, Wenchao Sun, Yongchang Zhang, Lei Yang, and Yandan Luo. Fully unified motion planning for end-to-end autonomous driving. *arXiv preprint arXiv:2504.12667*, 2025.

Seokha Moon, Minseung Lee, Joon Seo, Jinkyu Kim, and Jungbeom Lee. Causality-aware end-to-end autonomous driving via ego-centric joint scene modeling. *arXiv preprint arXiv:2605.13646*, 2026.

Brian A. Paden, Michal Cáp, Sze Zheng Yong, Dmitry S. Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1):33–55, 2016.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7073–7083, 2021.

Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:187–210, 2018.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Ziying Song, Lin Liu, Hongyu Pan, Bencheng Liao, Mingzhe Guo, Lei Yang, Yongchang Zhang, Shaoqing Xu, Caiyan Jia, and Yadan Luo. Breaking imitation bottlenecks: Reinforced diffusion powers diverse trajectory generation. *arXiv e-prints*, pp. arXiv–2507, 2025.

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

Jiacheng Tang, Zhiyuan Zhou, Zhuolin He, Jia Zhang, Kai Zhang, and Jian Pu. Causalvad: Deconfounding end-to-end autonomous driving via causal intervention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32124–32133, 2026.

Xinshuo Weng, B. Ivanovic, Yan Wang, Yue Wang, Marco Pavone, and Nvidia Research. Paradrive: Parallelized architecture for real-time autonomous driving. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15449–15458, 2024.

Penghao Wu, Li Chen, Hongyang Li, Xiaosong Jia, Junchi Yan, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6119–6132, 2022.

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1602–1611, 2025.

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

Chengran Yuan, Zhanqi Zhang, Jiawei Sun, Shuo Sun, Zefan Huang, Christina Dao Wen Lee, Dongen Li, Yuhang Han, Anthony Wong, Keng Peng Tee, et al. Drama: An efficient end-to-end motion planner for autonomous driving with mamba. *arXiv preprint arXiv:2408.03601*, 2024.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 87–104, 2024.

Zhiyu Zheng, Shaoyu Chen, Haoran Yin, Xinbang Zhang, Jialv Zou, Xinggang Wang, Qian Zhang, and Lefei Zhang. Resad: Normalized residual trajectory modeling for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3729–3739, 2026.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Diffusiondrivev2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.

A AI USE STATEMENT

During manuscript preparation, the authors used Large Language Models (LLMs) solely as a writing aid. Their role was limited to language editing: smoothing wording, correcting grammar, and improving the readability and coherence of the text. The models did not contribute to the research idea, the problem formulation, the methodology, the experimental design, the implementation, or the analysis and interpretation of results. All scientific claims, technical decisions, and empirical findings originate from the authors. The authors have reviewed every model-assisted passage and accept full responsibility for the final manuscript. We have ensured that any text produced by the LLMs conforms to ethical guidelines and does not constitute plagiarism or scientific misconduct.