

DiffRefiner: Coarse to Fine Trajectory Planning via Diffusion Refinement with Semantic Interaction for End to End Autonomous Driving

Liuhan Yin^{1,2*†}, Runkun Ju^{2*}, Guodong Guo², Er kang Cheng^{2‡}

¹Polytechnic Institute, Zhejiang University, China

²Nullmax, China

liuhanyin@zju.edu.cn, {jurunkun, guogudong, chengerkang}@nullmax.ai

Abstract

Unlike discriminative approaches in autonomous driving that predict a fixed set of candidate trajectories of the ego vehicle, generative methods, such as diffusion models, learn the underlying distribution of future motion, enabling more flexible trajectory prediction. However, since these methods typically rely on denoising human-crafted trajectory anchors or random noise, there remains significant room for improvement. In this paper, we propose DiffRefiner, a novel two-stage trajectory prediction framework. The first stage uses a transformer-based *Proposal Decoder* to generate coarse trajectory predictions by regressing from sensor inputs using predefined trajectory anchors. The second stage applies a *Diffusion Refiner* that iteratively denoises and refines these initial predictions. In this way, we enhance the performance of diffusion-based planning by incorporating a discriminative trajectory proposal module, which provides strong guidance for the generative refinement process. Furthermore, we design a fine-grained denoising decoder to enhance scene compliance, enabling more accurate trajectory prediction through enhanced alignment with the surrounding environment. Experimental results demonstrate that DiffRefiner achieves state-of-the-art performance, attaining 87.4 *EPDMS* on NAVSIM v2, and 87.1 *DS* along with 71.4 *SR* on Bench2Drive, thereby setting new records on both public benchmarks. The effectiveness of each component is validated via ablation studies as well.

Code — <https://github.com/nullmax-vision/DiffRefiner>

Introduction

End-to-end autonomous driving (E2E-AD) has achieved significant progress in recent years, which directly maps raw sensor inputs to driving actions or trajectory planning (Chitta et al. 2022; Sun et al. 2024; Chen et al. 2024a; Weng et al. 2024; Chitta, Prakash, and Geiger 2021; Muhammad et al. 2020; Hawke et al. 2020). Compared to traditional approaches that rely on separated perception, prediction, and planning modules, end-to-end methods optimize the entire

*These authors contributed equally.

†Work done during an internship at Nullmax.

‡Corresponding author.

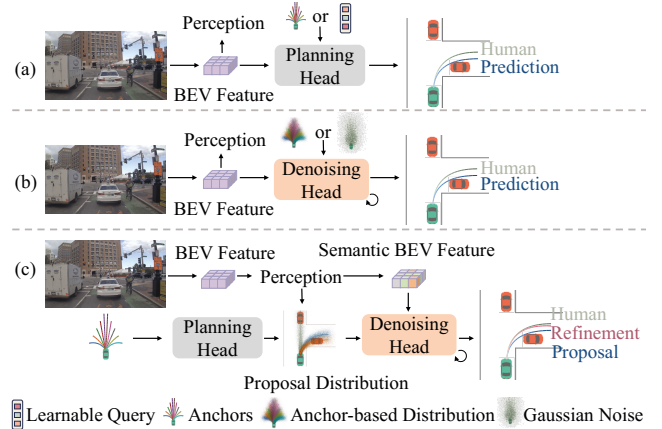


Figure 1: Comparison of different paradigms for end-to-end planning: (a) single-stage discriminative approach, (b) single-stage generative diffusion method, and (c) our proposed coarse-to-fine framework integrates discriminative proposal construction with generative diffusion refinement.

driving process in a unified manner, resulting in improved robustness and simpler deployment pipelines.

Prior approaches for ego-vehicle trajectory prediction typically employ single-pass regression on perception features or camera inputs (Hu et al. 2023; Casas, Sadat, and Urtasun 2021; Jiang et al. 2023a; Shao et al. 2023) (see Figure 1 (a)). While computationally efficient, these regression-based methods are fundamentally limited in their ability to handle the multimodal nature of trajectory prediction tasks. The regression optimization process averages across multimodal behaviors, yielding suboptimal predictions particularly at complex intersections, and consequently exhibits poor generalization in real-world driving scenarios. To address these challenges, recent work has investigated discretized solution spaces. For instance, several scoring-based approaches (Chen et al. 2024b; Li et al. 2024) employ offline-clustered trajectory anchors as discrete motion candidates, framing trajectory prediction as a classification problem. The distribution of future motions is then learned by evaluating each anchor either through its similarity to human demonstrations (Chen et al. 2024b) or via scores assigned

from offline simulation-based assessments (Yao et al. 2025; Li et al. 2025a). However, the computational complexity increases significantly with the size of the anchor set, limiting the feasibility of such methods for deployment in latency-sensitive autonomous driving systems (Liao et al. 2025).

A growing research direction leverages diffusion models to address the multimodal challenges of driving behavior. Building on their remarkable success in image and video generation (Yang, Srivastava, and Mandt 2023; Peebles and Xie 2023; Kim et al. 2023; Yang et al. 2023), these methods show strong potential for trajectory prediction (Jiang et al. 2023b; Wu et al. 2023). As shown in Figure 1 (b), these models provide a continuous, generative framework for trajectory prediction. Through iterative denoising of Gaussian-distributed samples (Chi et al. 2023), they can generate diverse, physically plausible trajectories while naturally capturing the multimodality inherent in driving decisions. DiffusionDrive (Liao et al. 2025) achieves this by generating diverse real-time trajectories through denoising of samples drawn from an anchor-parameterized Gaussian mixture model. However, current diffusion-based approaches suffer from critical limitations in their initialization process. These methods rely on unstructured Gaussian noise (Zheng et al. 2025) or fixed trajectory-derived anchors (Bae, Park, and Jeon 2024), both of which lack scene adaptability. When the initial samples deviate from feasible motion distributions, this necessitates excessive denoising iterations, which in turn leads to increased computational latency.

To address these challenges, we propose DiffRefiner, a novel two-stage trajectory prediction framework that adopts a coarse-to-fine architecture. A transformer-based Proposal Decoder first generates coarse trajectory predictions by regressing from a bank of predefined anchors, producing structured priors that serve as high-level guidance for subsequent refinement. The second stage employs a conditional Diffusion Refiner that iteratively refines these initial predictions via a generative diffusion process, thereby capturing trajectory details. This hybrid approach significantly improves diffusion-based planning performance by incorporating discriminative trajectory proposals that provide strong initialization for the subsequent generative refinement process. To enhance scene compliance, we propose a fine-grained denoising decoder that employs constrained diffusion to achieve precise alignment between predicted trajectories and the surrounding environment. Specifically, we propose a Fine-Grained Semantic Interaction Module that systematically integrates environmental constraints into the trajectory refinement process. The module operates in three stages: first, cross-attention layers establish dense correspondences between trajectory features and BEV semantic regions (e.g., drivable areas and obstacles) to encode holistic global context and scene-level dependencies; second, deformable attention selectively aligns trajectory endpoints with critical region semantics to extract fine-grained local structures and interaction cues; and third, an adaptive gating network dynamically fuses the global scene representation with localized semantic information, enabling the model to balance coarse contextual understanding and precise spatial alignment. This hierarchical design facilitates accurate, context-

aware trajectory optimization within the diffusion-based refinement framework.

In the experiments, we evaluate DiffRefiner on the open-loop real-world dataset NAVSIM and the closed-loop simulation benchmark Bench2Drive (Jia et al. 2024). Experimental results demonstrate that DiffRefiner achieves state-of-the-art performance, attaining 87.4 *EPDMS* on NAVSIM v2, and 87.1 *DS* along with 71.4 *SR* on Bench2Drive, thereby setting new records on both public benchmarks.

The main contributions of the paper can be summarized as follows:

- We propose a coarse-to-fine planning framework that first generates efficient anchor-based trajectory proposals as strong priors, then optimizes them through diffusion-based refinement.
- We present a fine-grained denoising decoder with a Scene-Aware Semantic Interaction Module that achieves precise trajectory refinement through optimized environment alignment during denoising.
- Our DiffRefiner achieves state-of-the-art (SOTA) performance on the open-loop real-world benchmark NAVSIM v2 and the closed-loop simulation benchmark Bench2Drive.

Related Work

End-to-End Autonomous Driving

Early end-to-end autonomous driving approaches (Hawke et al. 2020; Jiang et al. 2023a; Hu et al. 2023) employ single-modal prediction, directly mapping sensor inputs to deterministic trajectories. Subsequent methods (Chen et al. 2024b; Li et al. 2024, 2025a; Yao et al. 2025; Sun et al. 2024) address the multimodal nature of human driving through discriminative frameworks, formulating trajectory prediction either as anchor-based classification (using predefined or clustered anchors) or as offset regression from these anchors. Although these methods improve diversity over single-modal approaches, their performance remained limited by anchor coverage and the inability to model complex motion distributions.

Recently, diffusion-based generative methods have emerged as a promising alternative for trajectory prediction. Originally developed for image generation (Rombach et al. 2022; Peebles and Xie 2023), these approaches demonstrate strong capacity for modeling complex multi-modal motion distributions (Kondo et al. 2024). By learning to reverse a Gaussian corruption process applied to ground truth trajectories (Janner et al. 2022), they generate diverse, physically plausible motions (Zhu et al. 2023). For improved efficiency, DiffusionDrive (Liao et al. 2025) employs a truncated diffusion process initialized from clustered trajectory anchors, while Diffusion Planner (Zheng et al. 2025) combines transformer architectures with classifier guidance. Compared to deterministic regression, these diffusion-based approaches demonstrate superior ability to capture multimodality and maintain robustness in diverse driving scenarios.

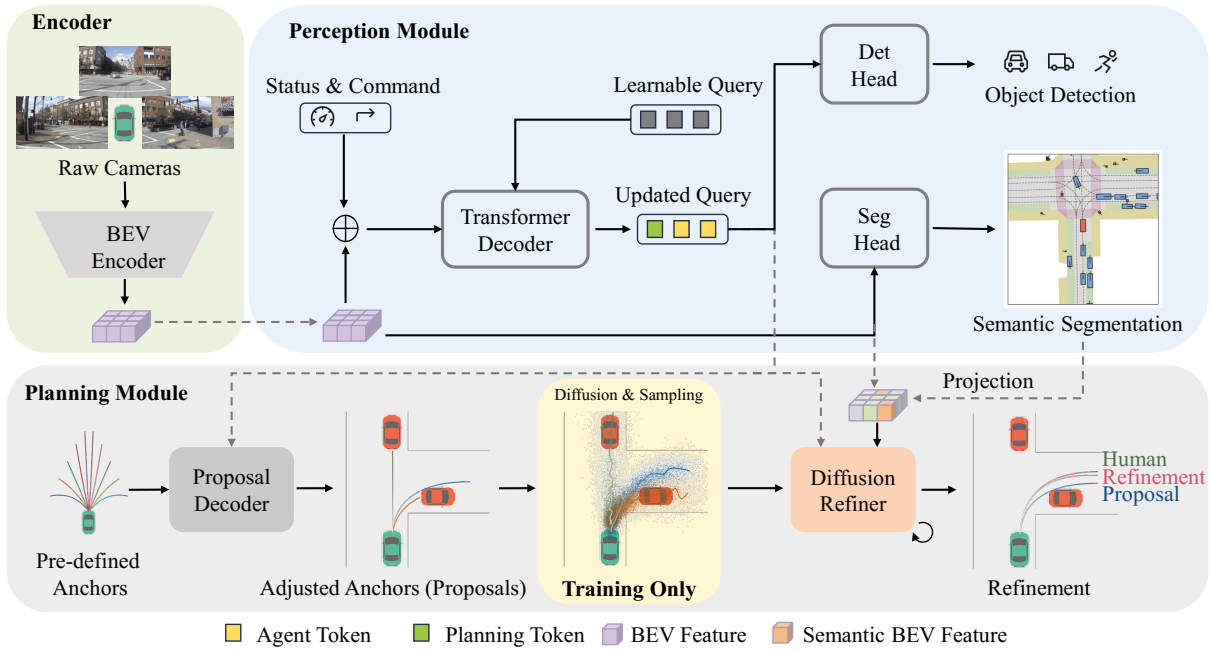


Figure 2: Overview of the proposed DiffRefiner. The DiffRefiner architecture comprises three primary components: a BEV encoder, an perception module, and a planning module, which sequentially perform scene representation learning, perception, and motion planning. The planning module is further decomposed into two submodules: (a) a proposal decoder, which employs a discriminative approach to produce coarse proposals that capture the overall motion trend; and (b) a diffusion refiner, which refines the proposals by leveraging a fine-grained denoising decoder conditioned on explicitly modeled scene semantics, thereby generating a final trajectory that better complies with environmental constraints.

Coarse-To-Fine Trajectory Planning

The coarse-to-fine framework has become a predominant architecture for trajectory planning, effectively reducing search space while enabling progressive refinement (Xing et al. 2025; Jia et al. 2025; Guo et al. 2025). These approaches first generate coarse trajectory candidates and subsequently refine them using specialized modules. For example, some methods employ diverse planning strategies, including approaches that first detect key agents and construct scene representations before generating final outputs (Su, Wu, and Yan 2024). Alternative frameworks utilize dual-decoder architectures to select coarse trajectory candidates for subsequent attention-based refinement (Yao et al. 2025). Additional methods enhance prediction accuracy by generating initial rough paths and refining them through iterative optimization (Wang et al. 2024; Xing et al. 2025).

Perception aware Trajectory Planning

Early approaches to end-to-end autonomous driving (Bojarski et al. 2016) directly map sensor inputs to planning outputs without incorporating explicit perception modules. Subsequent methods (Jiang et al. 2023a; Hu et al. 2023; Sun et al. 2024; Weng et al. 2024) substantially enhance planning performance through multi-task learning, where auxiliary perception tasks provide rich supervisory signals for the planning module. These methods typically process multi-modal sensor data through shared representations and leverage implicit feature interactions between perception and

planning. Another line of work (Li et al. 2025b,c) investigates self-supervised perception through temporal modeling of sensory inputs. While these approaches achieve greater computational efficiency, they lack mechanisms for explicit semantic understanding of the driving environment. However, a common limitation of these approaches is their inability to enable fine-grained interaction between perception and planning, which can result in unsafe behaviors such as collisions or violations of traffic constraints. In contrast, our framework establishes explicit semantic grounding for trajectory generation, where structured scene understanding provides fine-grained planning guidance to effectively mitigate these safety-critical failures.

Method

End-to-end autonomous driving takes raw sensor inputs and directly predicts the future trajectory of the ego vehicle. The predicted trajectory is denoted as $Y = \{Y_i\}_{i=1}^{T_f}$, where T_f indicates the prediction horizon, and Y_i represents the state of the ego vehicle at time step i , including its position and heading.

Overall Framework

The proposed DiffRefiner framework integrates three key components: a perception module, proposal decoder, and diffusion-based refiner, in a unified coarse-to-fine trajectory planning architecture. As illustrated in Figure 2, the

framework consists of three major components: (1) a BEV-centric perception module that processes sensor inputs and is trained with auxiliary tasks to enhance scene understanding; (2) a coarse trajectory proposal decoder that employs a lightweight Transformer (Vaswani et al. 2017) to adjust anchors and generate initial path predictions; and (3) a diffusion-based trajectory refiner that iteratively denoises and refines the proposals to produce optimized trajectories that better capture real-world driving complexity.

Perception Module

The perception module utilizes a BEV encoder (Jaeger, Chitta, and Geiger 2023) to generate bird’s-eye-view (BEV) features F_{bev} from raw sensor inputs. The module processes these features through two complementary heads: a sparse agent head for detecting individual objects and a dense segmentation head for comprehensive scene understanding. This dual-head architecture enables simultaneous object-level agent understanding and pixel-wise environment segmentation.

The segmentation head transforms the BEV features through a semantic segmentation network:

$$\hat{S} = \mathcal{F}_{\text{seg}}(F_{\text{bev}}), \quad (1)$$

where \hat{S} denotes the predicted semantic maps containing road elements, dynamic agents, and static obstacles.

For the sparse agent computation, to incorporate the ego vehicle’s state information, we encode its dynamic status (including velocity and acceleration) and navigation commands into a compact latent representation. This representation is combined with the scene context and processed by a transformer-based decoder that operates on a set of learnable queries Q . The decoder produces updated queries that are divided into two distinct types: a *Planning Token* T_p for trajectory generation and an *Agent Token* T_a for sparse detection tasks. The detection head processes the agent token to predict surrounding objects:

$$\hat{D} = \mathcal{F}_{\text{det}}(T_a), \quad (2)$$

where \hat{D} represents the detected agents’ positions and categories in the environment.

Proposal Decoder

In the first stage, we employ a lightweight Transformer-based approach that predicts offsets to adjust the predefined anchors (Li et al. 2025c), yielding the adjusted anchors as trajectory proposals. The method takes a set of pre-defined trajectory anchors, typically obtained through offline clustering as discrete motion candidates, and predicts the trajectory output via a Transformer decoder.

We define the offline-clustered discrete trajectory vocabulary as $\mathcal{V}c$. Each anchor is position-encoded and projected by an MLP to form initial proposal queries, which are subsequently contextualized via cross-attention with the Planning Token T_p :

$$Q_{\text{proposal}} = \text{CrossAttn}(Q = \text{MLP}(\text{pos}(\mathcal{V}c)), K = V = T_p) \quad (3)$$

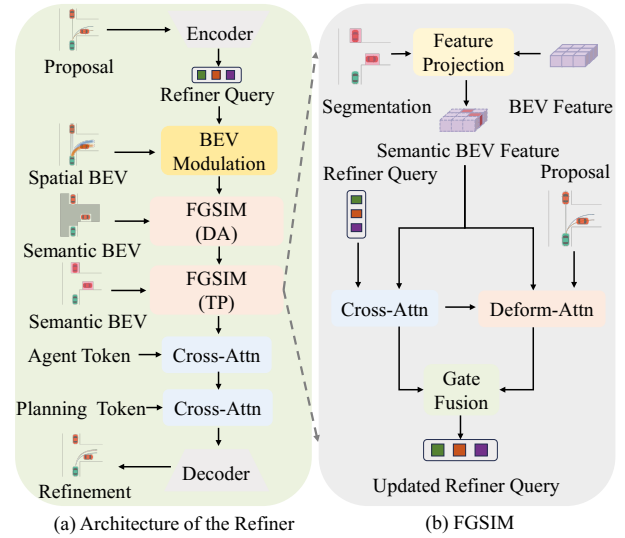


Figure 3: Illustration of the detailed architecture of the refiner and the Fine-Grained Semantic Interaction Module (FGSIM).

where $\text{pos}(\cdot)$ denotes sinusoidal positional encoding, $\text{MLP}(\cdot)$ projects each encoded anchor to a latent query space, T_p provides the planning-aware context, and Q_{proposal} represents the resulting context-enhanced trajectory queries.

Diffusion Refiner

In the second stage, our diffusion-based refiner optimizes all trajectory proposals through conditional denoising, generating more realistic and context-aware predictions. Specifically, as shown in Figure 3, we develop a fine-grained denoising decoder that explicitly enforces trajectory-environment alignment constraints during the iterative refinement process, ensuring enhanced compliance with scene semantics and dynamics.

Training Phase. During training, we simulate the forward diffusion process (Ho, Jain, and Abbeel 2020) by progressively adding Gaussian noise to Y_{proposal} over T steps. At a randomly sampled step t , the noisy trajectory \tilde{Y} is computed as:

$$\tilde{Y} = \sqrt{\bar{\alpha}_t} Y_{\text{proposal}} + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (4)$$

where $\bar{\alpha}_t$ is the cumulative product of the noise schedule (Song, Meng, and Ermon 2021). The noisy sample \tilde{Y} is encoded into refinement queries through positional encoding and MLPs as a multi-modal ego query:

$$Q_{\text{refiner}} = \text{Enc}(\tilde{Y}) = \text{MLP}(\text{pos}(\tilde{Y})) \quad (5)$$

Conditional Denoising with Scene-Aware Interaction. The scene-aware conditional denoising module enhances trajectory prediction through a hierarchical refinement process that integrates multi-level environmental context, as illustrated in Figure 3.

The architecture begins with a spatial BEV modulation module (Liao et al. 2025) that extracts trajectory-conditioned spatial features from the BEV representation.

These features initialize our Fine-Grained Semantic Interaction Module (FGSIM), which operates through two sequential refinement stages.

The first stage performs road-aware refinement by combining refiner queries with BEV features and drivable area segmentation, thereby constraining predictions to physically navigable road regions. Building upon this, the second stage conducts interaction-aware refinement by incorporating dynamic agent features, enabling explicit modeling of traffic participant interactions and proactive collision avoidance.

Subsequent cross-attention layers further refine predictions by capturing inter-agent relationships and ego-vehicle motion constraints. Finally, parallel MLP heads predict the refined trajectories and their confidence score, ensuring physical feasibility and contextual consistency.

Fine-Grained Semantic Interaction Module As map-based and interaction-based reasoning represent two fundamental aspects of autonomous planning, we introduce a semantic-aware interaction module that explicitly accounts for both. The module sequentially enhances the planner by aligning it with map semantics and dynamic agent interactions, while hierarchically integrating global scene context and local structural cues into trajectory decoding to improve scene understanding. To enable such targeted interaction, the module first extracts semantically critical regions from the perception output, providing region-level guidance for subsequent map- and agent-based reasoning.

To identify critical regions that are highly relevant to downstream planning, we leverage the semantic segmentation output $\hat{\mathcal{S}}$ to extract semantically meaningful areas of interest:

$$\mathcal{R} = \left\{ \left\{ r_{ij} \right\}_{i=1}^I \right\}_{j=1}^J, \quad (6)$$

where r_{ij} denotes the i -th region of the j -th semantic category (e.g., lane boundaries, crosswalks). Category-specific semantic masks are applied to filter the segmentation map, and the resulting regions are projected into the BEV feature space to enable region-aware feature extraction:

$$F_{R_i} = \left\{ \text{proj}(F_{\text{bev}}, \mathcal{R}_j) \right\}_{j=1}^J. \quad (7)$$

The module then processes each semantic feature F_{R_i} through complementary attention mechanisms designed to jointly capture global scene context and local geometric details. Here, Q_{refiner} denotes the updated refiner queries passed from the previous interaction stage. This is achieved via a two-step attention process: first, a global cross-attention operation integrates scene-wide contextual information, followed by a local deformable attention mechanism that adaptively focuses on regions spatially relevant to the trajectory:

$$Q_r^{(c)} = \text{CrossAttn}(Q = Q_{\text{refiner}}, K = V = F_{R_i}), \quad (8)$$

$$Q_r^{(d)} = \text{DeformAttn}(Q = Q_r^{(c)}, K = V = F_{R_i}; \tilde{Y}_T), \quad (9)$$

where \tilde{Y}_T provides trajectory-adaptive spatial reference.

A gated fusion mechanism dynamically balances these representations:

$$\text{Gate} = \sigma(W_{\text{gate}}(Q_r^{(c)}, Q_r^{(d)})), \quad (10)$$

$$Q_r = Q_r^{(c)} \cdot \text{Gate} + Q_r^{(d)} \cdot (1 - \text{Gate}), \quad (11)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function that maps the input to the range $(0, 1)$, W_{gate} is a learnable linear projection used to compute the gating coefficient, and Q_r denotes the updated refiner query after fusing global and local interactions.

Training Loss

Following prior end-to-end approaches (Hu et al. 2023; Jiang et al. 2023a), we employ a two-stage training scheme to enhance optimization stability.

In the first stage, the perception network is optimized using a Transfuser-style (Chitta et al. 2022) perception loss, denoted as $\mathcal{L}_{\text{perception}}$.

In the second stage, perception and planning are jointly optimized end-to-end. A winner-takes-all strategy selects the trajectory closest to the ground truth, and ego prediction loss is computed as:

$$\mathcal{L}_{\text{planning}} = w_{\text{reg}}\mathcal{L}_{\text{reg}} + w_{\text{cls}}\mathcal{L}_{\text{cls}}, \quad (12)$$

where \mathcal{L}_{reg} is the L1 regression loss and \mathcal{L}_{cls} is the binary cross-entropy classification loss. The final objective combines all components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{proposal}} + \mathcal{L}_{\text{refinement}} + \mathcal{L}_{\text{perception}}, \quad (13)$$

where $\mathcal{L}_{\text{proposal}}$ and $\mathcal{L}_{\text{refinement}}$ are the planning losses from the proposal and refinement modules.

Experiments

Experimental Setup

We use two widely recognized benchmarks: NAVSIM v2 (Cao et al. 2025) for open-loop evaluation and Bench2Drive (Jia et al. 2024) for closed-loop testing.

NAVSIM. NAVSIM (Dauner et al. 2024), based on the Openscene (Peng et al. 2023) dataset, is a real-world, planning-focused benchmark for evaluating autonomous driving models in open-loop settings. We evaluate on the *Navtest* split, comprising 12,146 frames covering diverse scenarios such as intersections, dynamic agents, and varied traffic conditions. To measure planning performance, we use the Extended Predictive Driver Model Score (EPDMS) introduced in NAVSIM v2 (Cao et al. 2025).

Bench2Drive. Bench2Drive (Jia et al. 2024) is a closed-loop evaluation benchmark based on CARLA (Dosovitskiy et al. 2017), designed to assess end-to-end autonomous driving systems in interactive urban scenarios. We evaluate our model on 220 routes spanning 44 diverse, interactive scenarios. Official metrics include Driving Score (DS), Success Rate (SR), and Multi-Ability Score, which collectively measure navigation performance, safety, and rule adherence.

Implementation Details. For the NAVSIM benchmark, we use the standard *navtrain* split for training. Consistent with the NAVSIM v2 Challenge specifications, our model processes synchronized multi-view inputs from the front, left-front, and right-front cameras. We train with a batch size of 384 and a learning rate of $4e-4$ for 100 epochs, applying identical training schedules for both the perception pretraining and end-to-end fine-tuning phases. For Bench2Drive

Method	Backbone	Modality	EPDMS \uparrow	NC \uparrow	DAC \uparrow	DDC \uparrow	TL \uparrow	EP \uparrow	TTC \uparrow	LK \uparrow	HC \uparrow	EC \uparrow
Human Agent	-	-	90.3	100	100	99.8	100	87.4	100	100	98.1	90.1
Transfuser (Chitta et al. 2022)	ResNet34	C+L	76.7	96.9	89.9	97.8	99.7	87.1	95.4	92.7	98.3	87.2
DiffusionDrive* (Liao et al. 2025)	ResNet34	C+L	84.0	98.2	96.2	98.6	-	87.6	97.3	97.4	98.4	-
GaussianFusion* (Liu et al. 2025)	ResNet34	C+L	85.0	98.3	97.3	99.0	-	87.5	97.4	97.4	98.3	-
HydraMDP++ (Li et al. 2025a)	ResNet34	C	81.4	97.2	97.5	99.4	99.6	83.1	96.5	94.4	98.2	70.9
DriveSuprim (Yao et al. 2025)	ResNet34	C	83.1	97.5	96.5	99.4	99.6	88.4	96.6	95.5	98.3	77.0
DiffRefiner (Ours)	ResNet34	C	86.2	98.5	97.4	99.6	99.8	87.6	97.7	97.7	98.3	86.2
HydraMDP++ (Li et al. 2025a)	V2-99	C	85.1	98.4	98.0	99.4	99.8	87.5	97.7	95.3	98.3	77.4
DriveSuprim (Yao et al. 2025)	V2-99	C	86.0	97.8	97.9	99.5	99.9	90.6	97.1	96.6	98.3	77.9
DiffRefiner (Ours)	V2-99	C	87.4	98.6	98.5	99.7	99.9	87.6	98.1	98.1	98.3	87.9

Table 1: Evaluation results on the NAVSIM v2 benchmark. We report the overall score (EPDMS) and a set of detailed driving metrics. Results are grouped by backbone types (ResNet34 and V2-99). Our proposed DiffRefiner consistently outperforms prior methods across most metrics and achieves the best overall performance. * indicates results reported from GaussianFusion (Liu et al. 2025), while results of other baselines are from DriveSuprim (Yao et al. 2025).

Method	Paradigm	Overall \uparrow		Multi-Ability \uparrow					
		DS	SR(%)	Merge	Overtake	EBrake	GiveWay	Tsign	Mean
PDM-Lite (Sima et al. 2024)	Rule based	97.0	92.3	88.8	93.3	98.3	90.0	93.7	92.8
UniAD (Hu et al. 2023)	Discriminative	45.8	16.4	14.1	17.8	21.7	10.0	14.2	15.6
ThinkTwice (Jia et al. 2023)	Discriminative	62.4	33.2	27.4	18.4	35.8	50.0	54.4	37.2
DriveTransformer (Jia et al. 2025)	Discriminative	63.5	35.0	17.6	35.0	48.4	40.0	52.1	38.6
TF++ (Zimmerlin et al. 2024)	Discriminative	84.2	67.3	58.8	57.8	83.3	40.0	82.1	64.4
HiPAD (Tang et al. 2025)	Discriminative	86.8	69.1	50.0	84.4	83.3	40.0	72.1	66.0
Orion (Fu et al. 2025)	Generative	77.7	54.6	25.0	71.1	78.3	30.0	69.2	54.7
GaussianFusion (Liu et al. 2025)	Generative	79.4	59.5	40.0	66.7	66.7	50.0	63.7	57.4
DiffRefiner (Ours)	Hybrid	87.1	71.4	63.8	60.0	85.0	50.0	86.3	69.0

Table 2: Performance comparison on the Bench2Drive benchmark. Our proposed DiffRefiner achieves the best overall performance among existing approaches, demonstrating significant improvements in DS, SR, and most multi-ability tasks.

evaluation, we follow the dataset configuration and pre-processing pipeline established in TF++ (Zimmerlin et al. 2024). Both benchmarks employ 20 clustered trajectory anchors for proposal initialization follow DiffusionDrive (Liao et al. 2025), and all proposals are then passed to the refinement module. All experiments are conducted on a cluster of 8 NVIDIA RTX 4090 GPUs.

Comparison with State-of-the-Art Methods

Results on NAVSIM. We perform comprehensive open-loop evaluations on the NAVSIM v2 benchmark. As demonstrated in Table 1, our framework establishes new state-of-the-art performance, surpassing the previous best method (DriveSuprim (Yao et al. 2025)) by significant margins of 3.7% (ResNet34 backbone) and 1.6% (V2-99 backbone). The results show particular improvements in safety critical metrics, validating the efficacy of our architectural design.

Results on Bench2Drive. We further evaluate our approach in a closed-loop setting on the Bench2Drive benchmark. As shown in Table 2, our method outperforms all existing learning-based baselines, achieving state-of-the-art re-

sults. Without model ensembling, it improves DS by 0.3 and SR by 2.3 over the previous best, HiPAD (Tang et al. 2025). Unlike prior discriminative or fully generative methods, our framework employs a hybrid generative paradigm that integrates coarse anchor-based proposals with diffusion-based refinement. This design yields consistent gains across most multi-ability metrics, demonstrating robustness and effectiveness in diverse interactive driving scenarios.

Ablation Study

Ablation on Planning Framework. The impact of each stage in our planning framework is evaluated in Table 3. A comparison between rows 1 and 6 indicates that incorporating the refiner improves EPDMS by 1.2, confirming that it substantially enhances the quality of trajectory proposals. Rows 2 and 6 further show that higher-quality proposals raise the refinement upper bound, where row 2 applies two denoising iterations while row 6 uses a single iteration. Meanwhile, rows 1 and 5 indicate that proper refiner supervision benefits proposal learning. Finally, the comparison of rows 4 and 6 highlights the advantage of our gen-

ID	Ref	Pro	RT	Src	Param	Latency (ms)	EPDMS \uparrow
1		✓	-	Pro	57.2M	12	85.0
2	✓		Gen	Ref	73.7M	40	86.0
3	✓	✓	Dis	Pro	74.8M	12	85.5
4	✓	✓	Dis	Ref	74.8M	27	78.3
5	✓	✓	Gen	Pro	74.8M	12	85.8
6	✓	✓	Gen	Ref	74.8M	27	86.2

Table 3: Ablation study of the proposed planning framework. Ref: Refinement; Pro: Proposal; RT: Refiner type; Src: Input source of downstream control; Gen: Generative; Dis: Discriminative; Param: Total number of model parameters; Latency: End-to-end planning latency.

ID	P	A	M	DA	TP	EPDMS \uparrow	NC \uparrow	DAC \uparrow
1	✓					82.4	97.9	93.8
2	✓	✓				82.9	98.0	94.5
3	✓		✓			83.3	98.0	94.9
4	✓	✓	✓			83.5	98.1	95.1
5	✓	✓	✓	✓		84.3	98.0	95.7
6	✓	✓	✓	✓	✓	85.0	98.4	96.3

Table 4: Ablation study of refiner components. P: Planning token; A: Agent token; M: BEV modulation; DA: Drivable area in FGSIM; TP: Traffic participant in FGSIM.

erative refiner over discriminative alternatives in performing fine-grained trajectory adjustments.

Ablation on Refiner Components. As shown in Table 4, we perform a systematic ablation to assess the contribution of each module in the refiner decoder. Results show consistent gains from all components, validating the overall design. Comparing rows 4–6, the semantic interaction mechanism progressively improves scene understanding and mitigates collision-related errors by exploiting fine-grained semantic cues.

Analysis of FGSIM Components. As illustrated in Table 5, both global context and local target cues independently lead to performance improvements, demonstrating their complementary roles. However, a naive additive fusion of the two results in performance degradation due to conflicting information. In contrast, our gating mechanism adaptively balances their contributions, achieving the best performance and confirming the benefit of adaptive feature integration.

ID	Global	Local	Fusion	EPDMS \uparrow	NC \uparrow	DAC \uparrow
1	✓		-	85.9	98.4	97.2
2		✓	-	85.9	98.4	97.2
3	✓	✓	Addition	85.9	98.4	97.2
4	✓	✓	Gating	86.2	98.5	97.4

Table 5: Ablation study of FGSIM components. “Global” and “Local” refer to global cross-attention and local deformable attention, respectively. “Fusion” compares additive and gating-based fusion strategies.

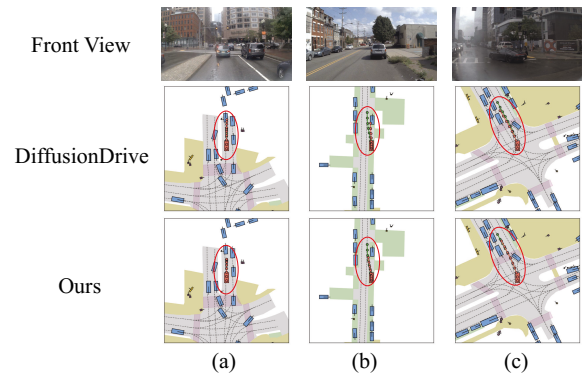


Figure 4: Visualization of representative examples of DiffusionDrive (Liao et al. 2025) and our method. Green points denote the ground truth, and red points denote the prediction. (a) and (b) illustrate cases in which our method achieves better collision avoidance compared with DiffusionDrive, whereas (c) demonstrates case where our method exhibits improved compliance with lane constraints.

Steps	EPDMS \uparrow	NC \uparrow	DAC \uparrow	DDC \uparrow	EP \uparrow	LK \uparrow
1	86.20	98.47	97.36	99.64	87.59	97.79
2	86.22	98.47	97.37	99.63	87.58	97.74
5	86.17	98.48	97.34	99.63	87.58	97.78

Table 6: Ablation study on the number of denoising steps.

Ablation on Denoising Steps. The results in Table 6 demonstrate that near-optimal performance can be achieved with just a single denoising step. This finding underscores the effectiveness of high-quality proposals as strong priors, enabling efficient diffusion-based refinement and highlighting the suitability of our framework for real-time end-to-end autonomous driving.

Qualitative Analysis

As illustrated in Figure 4, our method surpasses DiffusionDrive (Liao et al. 2025) in complex interactive scenarios by better attending to fine-grained scene details. It reduces collisions with surrounding agents and adheres more strictly to map constraints, resulting in higher-quality trajectories.

Conclusions

In this work, we present DiffRefiner, a novel two-stage trajectory prediction framework for end-to-end autonomous driving planning. DiffRefiner incorporates a transformer-based proposal decoder to generate coarse trajectories, which provide strong guidance for subsequent generative refinement. The diffusion refiner further improves these proposals via iterative denoising with a fine-grained decoder, thereby enhancing scene compliance and producing more accurate and realistic trajectories. Extensive experiments on NAVSIM v2 and Bench2Drive demonstrate that DiffRefiner achieves state-of-the-art performance.

References

- Bae, I.; Park, Y.-J.; and Jeon, H.-G. 2024. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17890–17901.
- Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.
- Cao, W.; Hallgarten, M.; Li, T.; Dauner, D.; Gu, X.; Wang, C.; Miron, Y.; Aiello, M.; Li, H.; Gilitschenski, I.; et al. 2025. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218*.
- Casas, S.; Sadat, A.; and Urtasun, R. 2021. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14403–14412.
- Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2024a. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Chen, S.; Jiang, B.; Gao, H.; Liao, B.; Xu, Q.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024b. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*.
- Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.
- Chitta, K.; Prakash, A.; and Geiger, A. 2021. Neat: Neural attention fields for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15793–15803.
- Chitta, K.; Prakash, A.; Jaeger, B.; Yu, Z.; Renz, K.; and Geiger, A. 2022. Transfuser: Fusion with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11): 12878–12895.
- Dauner, D.; Hallgarten, M.; Li, T.; Weng, X.; Huang, Z.; Yang, Z.; Li, H.; Gilitschenski, I.; Ivanovic, B.; Pavone, M.; et al. 2024. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37: 28706–28719.
- Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.
- Fu, H.; Zhang, D.; Zhao, Z.; Cui, J.; Liang, D.; Zhang, C.; Zhang, D.; Xie, H.; Wang, B.; and Bai, X. 2025. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*.
- Guo, K.; Liu, H.; Wu, X.; Pan, J.; and Lv, C. 2025. iPad: Iterative Proposal-centric End-to-End Autonomous Driving. *arXiv preprint arXiv:2505.15111*.
- Hawke, J.; Shen, R.; Gurau, C.; Sharma, S.; Reda, D.; Nikolov, N.; Mazur, P.; Micklethwaite, S.; Griffiths, N.; Shah, A.; et al. 2020. Urban driving with conditional imitation learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 251–257. IEEE.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17853–17862.
- Jaeger, B.; Chitta, K.; and Geiger, A. 2023. Hidden biases of end-to-end driving models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8240–8249.
- Janner, M.; Du, Y.; Tenenbaum, J. B.; and Levine, S. 2022. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*.
- Jia, X.; Wu, P.; Chen, L.; Xie, J.; He, C.; Yan, J.; and Li, H. 2023. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21983–21994.
- Jia, X.; Yang, Z.; Li, Q.; Zhang, Z.; and Yan, J. 2024. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37: 819–844.
- Jia, X.; You, J.; Zhang, Z.; and Yan, J. 2025. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. In *The Thirteenth International Conference on Learning Representations*.
- Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023a. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8350.
- Jiang, C. M.; Cornman, A.; Park, C.; Sapp, B.; Zhou, Y.; and Anguelov, D. 2023b. MotionDiffuser: Controllable Multi-Agent Motion Prediction using Diffusion. *arXiv:2306.03083*.
- Kim, S. W.; Brown, B.; Yin, K.; Kreis, K.; Schwarz, K.; Li, D.; Rombach, R.; Torralba, A.; and Fidler, S. 2023. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8496–8506.
- Kondo, K.; Tagliabue, A.; Cai, X.; Tewari, C.; Garcia, O.; Espitia-Alvarez, M.; and How, J. P. 2024. Cgd: Constraint-guided diffusion policies for uav trajectory planning. *arXiv preprint arXiv:2405.01758*.
- Li, K.; Li, Z.; Lan, S.; Xie, Y.; Zhang, Z.; Liu, J.; Wu, Z.; Yu, Z.; and Alvarez, J. M. 2025a. Hydra-mdp++: Advancing end-to-end driving via expert-guided hydra-distillation. *arXiv preprint arXiv:2503.12820*.
- Li, Y.; Fan, L.; He, J.; Wang, Y.; Chen, Y.; Zhang, Z.; and Tan, T. 2025b. Enhancing End-to-End Autonomous Driving with Latent World Model. In Yue, Y.; Garg, A.; Peng,

- N.; Sha, F.; and Yu, R., eds., *International Conference on Representation Learning*, volume 2025, 42942–42959.
- Li, Y.; Wang, Y.; Liu, Y.; He, J.; Fan, L.; and Zhang, Z. 2025c. End-to-end driving with online trajectory evaluation via bev world model. *arXiv preprint arXiv:2504.01941*.
- Li, Z.; Li, K.; Wang, S.; Lan, S.; Yu, Z.; Ji, Y.; Li, Z.; Zhu, Z.; Kautz, J.; Wu, Z.; et al. 2024. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*.
- Liao, B.; Chen, S.; Yin, H.; Jiang, B.; Wang, C.; Yan, S.; Zhang, X.; Li, X.; Zhang, Y.; Zhang, Q.; et al. 2025. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12037–12047.
- Liu, S.; Liang, Q.; Li, Z.; Li, B.; and Huang, K. 2025. GaussianFusion: Gaussian-Based Multi-Sensor Fusion for End-to-End Autonomous Driving. *arXiv preprint arXiv:2506.00034*.
- Muhammad, K.; Ullah, A.; Lloret, J.; Del Ser, J.; and De Albuquerque, V. H. C. 2020. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 22(7): 4316–4336.
- Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.
- Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 815–824.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Shao, H.; Wang, L.; Chen, R.; Li, H.; and Liu, Y. 2023. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, 726–737. PMLR.
- Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, 256–274. Springer.
- Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
- Su, H.; Wu, W.; and Yan, J. 2024. Difs-d: Ego-centric fully sparse paradigm with uncertainty denoising and iterative refinement for efficient end-to-end autonomous driving. *arXiv e-prints*, arXiv–2409.
- Sun, W.; Lin, X.; Shi, Y.; Zhang, C.; Wu, H.; and Zheng, S. 2024. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*.
- Tang, Y.; Xu, Z.; Meng, Z.; and Cheng, E. 2025. Hip-ad: Hierarchical and multi-granularity planning with deformable attention for autonomous driving in a single decoder. *arXiv preprint arXiv:2503.08612*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, K.-L.; Tsao, L.-W.; Wu, J.-C.; Shuai, H.-H.; and Cheng, W.-H. 2024. TrajFine: Predicted trajectory refinement for pedestrian trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4483–4492.
- Weng, X.; Ivanovic, B.; Wang, Y.; Wang, Y.; and Pavone, M. 2024. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15449–15458.
- Wu, K.; Zhou, Y.; Shi, H.; Li, X.; and Ran, B. 2023. Graph-based interaction-aware multimodal 2D vehicle trajectory prediction using diffusion graph convolutional networks. *IEEE Transactions on Intelligent Vehicles*, 9(2): 3630–3643.
- Xing, Z.; Zhang, X.; Hu, Y.; Jiang, B.; He, T.; Zhang, Q.; Long, X.; and Yin, W. 2025. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1602–1611.
- Yang, B.; Luo, Y.; Chen, Z.; Wang, G.; Liang, X.; and Lin, L. 2023. Law-diffusion: Complex scene generation by diffusion with layouts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22669–22679.
- Yang, R.; Srivastava, P.; and Mandt, S. 2023. Diffusion probabilistic modeling for video generation. *Entropy*, 25(10): 1469.
- Yao, W.; Li, Z.; Lan, S.; Wang, Z.; Sun, X.; Alvarez, J. M.; and Wu, Z. 2025. DriveSuprim: Towards Precise Trajectory Selection for End-to-End Planning. *arXiv preprint arXiv:2506.06659*.
- Zheng, Y.; Liang, R.; Zheng, K.; Zheng, J.; Mao, L.; Li, J.; Gu, W.; Ai, R.; Li, S. E.; Zhan, X.; et al. 2025. Diffusion-based planning for autonomous driving with flexible guidance. *arXiv preprint arXiv:2501.15564*.
- Zhu, Z.; Zhao, H.; He, H.; Zhong, Y.; Zhang, S.; Guo, H.; Chen, T.; and Zhang, W. 2023. Diffusion models for reinforcement learning: A survey. *arXiv preprint arXiv:2311.01223*.
- Zimmerlin, J.; Beißwenger, J.; Jaeger, B.; Geiger, A.; and Chitta, K. 2024. Hidden biases of end-to-end driving datasets. *arXiv preprint arXiv:2412.09602*.