UniTeD: Unified Temporal Diffusion for Joint Perception and Planning in Autonomous Driving

Bo Zhao[1⋆], Xinting Zhao[1⋆], Naifan Li[1], Erkang Cheng[1✉], and Haibin Ling[2]

[1] Nullmax, China
[2] Westlake University, China
{zhaobo,zhaoxinting,linaifan,chengerkang}@nullmax.ai,
linghaibin@westlake.edu.cn

Abstract. Diffusion models have shown strong potential for multi-modal planning in end-to-end autonomous driving. However, most existing methods confine diffusion to the planning module, conditioning on fixed outputs from separate discriminative perception networks. This decoupled design propagates perception errors to the planner, increasing optimization difficulty and reducing robustness. To overcome these limitations, we propose *UniTeD*, a *Uni*fied *Te*mporal *D*iffusion framework that jointly models perception and planning through iterative denoising in a shared generative space. By enabling bidirectional information exchange, the framework facilitates mutual refinement between tasks and improves robustness via noise-conditioned multi-task training. We further extend this unified diffusion paradigm to a streaming setting by incorporating temporal context. A Temporal Transition Module (TTM) is introduced to resolve the noise-level mismatch between historical and current frames. In addition, we propose an Anchor Refresh Strategy (ARS) to alleviate the training–inference distribution shift commonly observed in sparse diffusion-based end-to-end driving frameworks. Without bells and whistles, UniTeD achieves state-of-the-art performance across multiple benchmarks, surpassing both recent discriminative end-to-end methods and diffusion-based planning approaches.

Keywords: End-to-End Autonomous Driving · Unified Task Denoising · Temporal Interaction

1 Introduction

Autonomous driving is undergoing a fundamental paradigm shift, moving from traditional modular pipelines, with perception [18, 26, 29, 30, 32, 33, 35], motion prediction [15, 16, 22, 49], and planning [7, 10, 11, 38] as separate components, toward fully *end-to-end* (E2E) systems [9,17,23,31]. E2E approaches reformulate autonomous driving as a fully differentiable process that directly optimizes the final planning objective.

⋆ Equal contribution.
✉ Corresponding author.

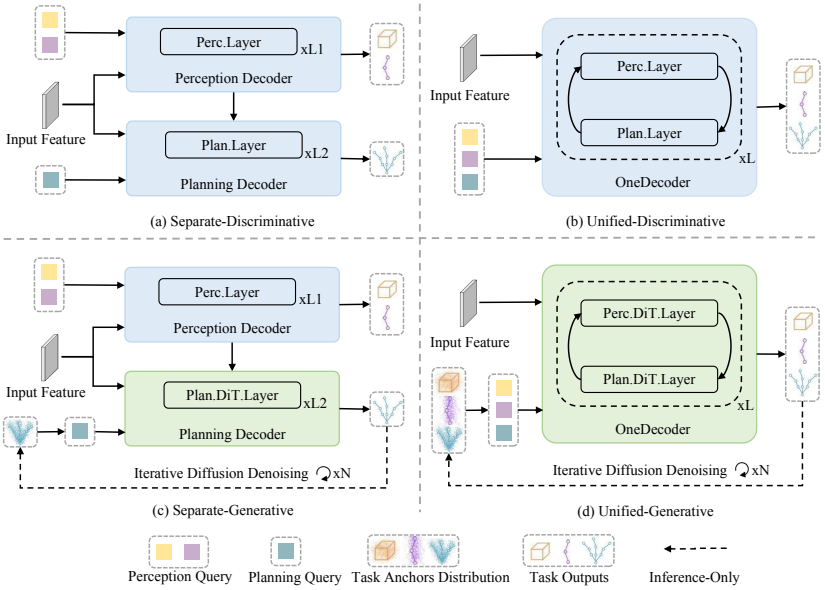

Fig. 1: Comparison of existing paradigms for end-to-end autonomous driving. (a) Separate-Discriminative: separate perception and planning with discriminative modeling for both tasks. (b) Unified-Discriminative: unified perception and planning with discriminative modeling for both tasks. (c) Separate-Generative: separate perception and planning with generative modeling for planning task only. (d) Unified-Generative (ours): unified perception and planning with generative modeling for both tasks.

Most recent E2E approaches adopt a discriminative formulation [9,17,21,23, 42,43,46]. They typically treat perception and planning separately (Fig. 1 (a)), either sequentially [17,23,42] or in parallel [9,46]. For instance, UniAD [17] and VAD [23] employ sequential pipelines, where perception outputs are passed to a downstream planning module. In contrast, TransFuser [9] and PARA-Drive [46] process perception and planning tasks concurrently, but with limited cross-task interaction. To enable deeper integration, DriveTransformer [21] and HiP-AD [43] introduce unified decoders that jointly process all task queries. As shown in Fig. 1 (b), perception (agent and map) and planning queries interact at every layer and directly attend to shared feature representations, substantially improving E2E driving performance. However, these discriminative methods share a core limitation rooted in their optimization objective: minimizing supervised loss encourages predictions toward mean or majority behaviors, limiting their ability to capture the multi-modal nature of driving.

To address this issue, recent research has explored generative modeling, particularly diffusion-based approaches [8,31,48,50] (Fig. 1 (c)). DiffusionDrive [31] proposes a truncated diffusion policy that stabilizes generation while reducing denoising steps for real-time efficiency. ResAD [50] addresses spatiotemporal imbalance by reformulating trajectory prediction as residual learning. DiffRe-

finer [48] further introduces a two-stage framework, where a diffusion-based refiner denoises coarse trajectories generated in an initial stage. Despite advances achieved, existing generative frameworks limit diffusion modeling to the planning task, treating perception outputs as fixed conditions. This strategy may propagate perception errors to the generative process, leading to inaccurate guidance and increased optimization complexity. It thus prevents joint refinement between perception and planning, restricting the full potential of generative modeling.

Furthermore, existing approaches use only single-frame information for diffusion-based planning and ignore temporal dynamics. In addition, sparse query-based diffusion planning methods [31, 48] often suffer from severe training-inference inconsistency: during training, only a small subset of planning queries matched to the ground truth are optimized, while the remaining queries receive little supervision, causing a query distribution shift at inference time. Such behavior fundamentally contradicts the iterative refinement principle of diffusion models, where outputs are expected to progressively improve as denoising proceeds.

To address the aforementioned challenges, we present *UniTeD*, a *Uni*fied *Te*mporal *D*iffusion framework that jointly models perception and planning through iterative denoising in a shared generative space. Specifically, UniTeD adopts a unified diffusion decoder that jointly processes perception queries (agents and map elements) and planning queries within a single generative process. By denoising them simultaneously in a shared generative space, our approach fully leverages the strong uncertainty modeling capability of diffusion models and enables comprehensive cross-task information exchange. In this way, UniTeD not only supports multi-modal trajectory generation for the ego vehicle but also strengthens perception tasks, including dynamic agent prediction and static map understanding. Importantly, optimization under a noise-conditioned diffusion paradigm inherently improves multi-task robustness. Each task is trained to ensure accurate predictions even when conditioned on noisy or imperfect intermediate outputs from other tasks, leading to greater stability and improved generalization at inference time.

Additionally, to better capture temporal dynamics that are often overlooked in existing diffusion-based planners, we incorporate a memory bank that stores historical task queries, enabling a streaming unified diffusion framework. We introduce a Temporal Transition Module (TTM) to solve the noise-level mismatch between historical and current frames. Furthermore, we propose an Anchor Refresh strategy (ARS) to mitigate the training–inference distribution shift in sparse diffusion frameworks. During inference, high-confidence queries are retained, while low-confidence task queries are refreshed by resampling from the original noise distribution. This mechanism preserves query diversity while maintaining alignment with the training distribution, ensuring stable and reliable performance even in complex multi-modal settings.

Extensive experiments demonstrate that UniTeD achieves strong performance across major benchmarks, outperforming both discriminative E2E approaches and recent diffusion-based planning ones. In particular, our method attains 87.25 DS on Bench2Drive, as well as 90.24 PDMS and 90.13 EPDMS on NAVSIM.

The main contributions of this work are summarized as follows:

1. We propose a novel unified diffusion framework for E2E autonomous driving. Our solution jointly denoises perception and planning queries, enabling deep bidirectional information exchange within a shared generative space.
2. We extend unified diffusion to a streaming setting via a memory bank and introduce a Temporal Transition Module to resolve noise-level mismatches between historical and current frames.
3. We present an Anchor Refresh Strategy to alleviate training–inference distribution shifts in sparse diffusion-based autonomous driving frameworks.
4. We show superior performance across major benchmarks, surpassing both state-of-the-art discriminative E2E methods and diffusion-based planners.

2 Related Works

Discriminative End-to-End Autonomous Driving. Most recent E2E approaches typically apply discriminative methods for both the planning and perception tasks. UniAD [17] represents a milestone in this category, integrating multiple tasks into a single model, while VAD [23] further improves efficiency by replacing dense rasterized scene representations with vectorized ones. SparseDrive [42] eliminates the dependency on dense BEV features through a fully sparse query-based framework. To mitigate the sequential dependencies, several works explore parallel architectures. TransFuser [9] fuses LiDAR and camera inputs to construct a BEV representation, executing perception and planning tasks concurrently. PARA-Drive [46] extends this by jointly modeling mapping, motion, and occupancy prediction alongside planning. DriveTransformer [21] introduces a unified decoder that jointly processes all task queries. HiP-AD [43] similarly employs a unified decoder for both perception and planning tasks, and further enhances performance by incorporating multi-granularity planning queries.

Generative Methods in Autonomous Driving. In the perception domain of autonomous Driving, diffusion-based frameworks have redefined traditional tasks as a denoising process. DiffusionDet [6] first redefines object detection as a denoising diffusion process from noisy boxes to object boxes. MotionDiffuser [24] utilizes diffusion processes to model the multi-modal distribution of multi-agent motion prediction. DiffusionTrack [36] introduces a noise-to-tracking paradigm, reformulating multi-object tracking as a conditional denoising task to achieve more robust data association. DiffMap [19] utilizes generative models for map segmentation. PolyDiffuse [4] further advances this by generating vectorized map elements through a denoising process. LaneDiffusion [45] strengthens the representation of complex road networks by effectively modeling lane centerlines and topological connections.

In the planning domain, DiffusionDrive [31] proposes a truncated diffusion policy to stabilize generation while enhancing real-time efficiency through reduced denoising steps. ResAD [50] reformulates trajectory prediction as residual

learning to address spatiotemporal imbalances. DiffRefiner [48] employs a two-stage framework, utilizing a diffusion-based refiner to polish coarse trajectories from the initial stage. DiffusionDriveV2 [53] incorporates reinforcement learning to prune low-quality modes and explore superior trajectories. DiffAD [44] reformulates autonomous driving as a conditional image generation task, predicting future scenes in the form of rasterized BEV images.

Being a generative solution, our UniTeD jointly formulates perception and planning within a shared generative space, while previous methods retain diffusion-based models for planning while treating discriminative perception outputs as fixed conditions for the planner. Furthermore, UniTeD extends the generative framework to a streaming scheme.

Temporal Modeling in Autonomous Driving. Existing E2E frameworks typically incorporate temporal dynamics through two main streams: Global Feature Fusion [17, 23] and Query-based Propagation [21, 40, 42, 43]. Early methods like UniAD [17] and VAD [23] utilize stacked historical BEV features to provide a global spatial-temporal context for downstream tasks. SparseDrive [42], DriveTransformer [21], and HiP-AD [43] all utilize a memory queue to cache historical queries, leveraging temporal attention to integrate historical information. To ensure planning stability, MomAD [40] further introduces a momentum mechanism to select the optimal planning query, which is then integrated with historical queries via cross-attention. Despite the success of these discriminative frameworks, incorporating temporal modeling into diffusion-based E2E frameworks remains challenging. Existing diffusion planners, such as DiffusionDrive [31] and ResAD [50], primarily focus on single-frame generation or naive concatenation of historical features. Different from these methods, UniTeD introduces a unified diffusion-based streaming paradigm that propagates joint task queries across frames. A novel module is also proposed to solve the noise-level mismatch between historical and current frames.

3 Method

3.1 Overview

The overall architecture of UniTeD is illustrated in Fig. 2, which operates as a unified diffusion process. First, the image backbone extracts multi-scale features $\mathbf{F} = \{\mathbf{F}_v\}_{v=1}^{V}$ from V multi-view images at k-th frame. Second, we define $\mathbf{A} = \{\mathbf{A}^a \in \mathbb{R}^{N_a \times D_a}, \mathbf{A}^m \in \mathbb{R}^{N_m \times D_m}, \mathbf{A}^p \in \mathbb{R}^{N_p \times D_p}\}$ as a joint anchor set for agent, map, planning tasks, respectively. N_a, N_m, and N_p denote the numbers of anchors for the different tasks, and D_a, D_m, and D_p denote their corresponding anchor dimensions. Following the truncated diffusion [31], the noisy anchors $\mathbf{A}_{t_k}$ are sampled from a Gaussian distribution centered around the prior $\mathbf{A}$, where t_k denotes the noise level associated with the k-th frame. Then $\mathbf{A}_{t_k}$ are projected into a shared latent query space by an Anchor Embedding module, yielding $\mathbf{Q}_k = \{\mathbf{Q}^a, \mathbf{Q}^m, \mathbf{Q}^p\} \in \mathbb{R}^{(N_a+N_m+N_p)\times C}$. These queries are then fed into the Unified Diffusion Decoder (Sec. 3.3). To incorporate temporal context, the Temporal Transition Module (TTM) (Sec. 3.4) modulates a temporal queue of

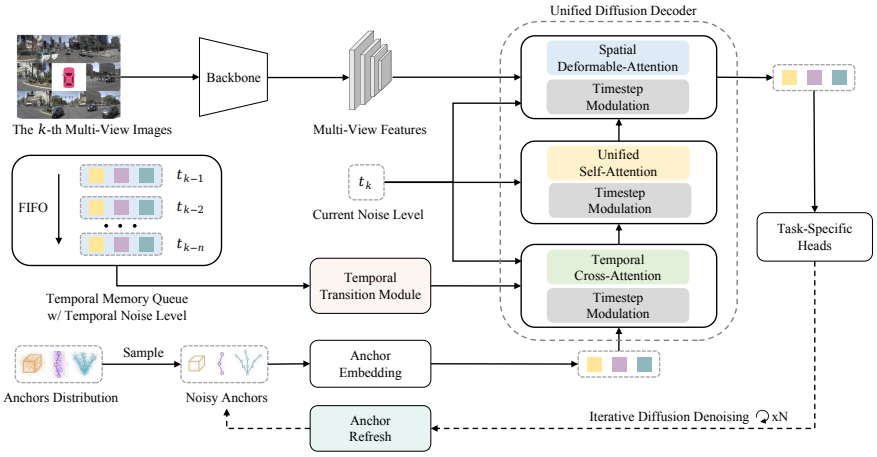

Fig. 2: Overview of UniTeD. UniTeD contains three core components: (a) Unified Diffusion Decoder that models perception and planning queries through iterative denoising in a shared generative space; (b) Temporal Transition Module (TTM) that resolves the noise-level mismatch between historical and current frames; and (c) Anchor Refresh Strategy that alleviates the training–inference distribution shift.

historical features $\{\mathbf{Q}_{k-i}\}_{i=1}^{n}$ to generate aligned historical features $\{\tilde{\mathbf{Q}}_{k-i}\}_{i=1}^{n}$ consistent with the current noise level t_k. Together with the multi-view features $\mathbf{F}$, the modulated temporal features serve as conditional inputs to the decoder. Finally, after denoising, task-specific heads transform the refined queries into final predictions for the agent, map, and planning tasks.

During inference, an Anchor Refresh Strategy (Sec. 3.5) updates the predicted outputs to initialize the subsequent denoising step, enabling iterative refinement until the final timestep is reached.

3.2 Unified Modeling via Diffusion

We follow the truncated diffusion policy in DiffusionDrive [31], but extend it to a joint multi-task modeling form. The Unified Diffusion Decoder simultaneously models the distribution of all task instances by denoising the noisy anchors $\mathbf{A}_{t_k}$ into their final clean outputs. The forward and reverse processes are as follows:

Unified Forward Diffusion Process. During the forward process, we apply the same level of noise to all tasks simultaneously. For each anchor $a \in \mathbf{A}$, the unified forward process at noise timestep t_k is defined as:

$$q(a_{t_k}|a) = \mathcal{N}(a_{t_k}; \sqrt{\bar{\alpha}_{t_k}}a, (1-\bar{\alpha}_{t_k})\mathbf{I}), \quad t_k \in [1, T_{\text{trunc}}] \tag{1}$$

where $T_{\text{trunc}} \ll T_{\text{max}}$ denotes the maximum number of truncated diffusion steps. The resulting noisy anchors serve as the input to the decoder. This synchronized noise injection ensures that all task categories are embedded with a shared uncertainty scale, promoting consistent representation learning across tasks.

Unified Reverse Denoising Process. In the reverse process, the decoder recovers the clean unified all-task instances from the noisy anchor inputs. We follow the DDIM sampling [39] to iteratively refine the task states. During each denoising step, for k-th frame, the decoder takes the current step noisy anchors $\mathbf{A}_{t_k}$, multi-view visual features $\mathbf{F}$, and modulated historical context queue $\{\tilde{\mathbf{Q}}_{k-i}\}_{i=1}^{n}$ as inputs to estimate the clean state of each instance.

3.3 Unified Diffusion Decoder

The Unified Diffusion Decoder serves as the generative core of our framework (Fig. 2). To progressively recover clean task states, the decoder refines the joint query set $\mathbf{Q}_k$ through N denoising iterations across L stacked blocks, each consisting of the specialized interaction layers details below:

Conditional Modulation. To condition the denoising process on the generative stage, every interaction layer within the decoder block is governed by a Conditional Modulation module. Following the Adaptive Layer Normalization (AdaLN) design in DiT [37], the condition $\mathbf{C}_k$, by encoding current timestep t_k, fed into an MLP to produce modulation parameters $\{\alpha, \beta, \gamma\}$, which dynamically projects task features onto the corresponding distribution level via AdaLN.

Temporal Cross-Attention Layer. After the initial modulation, a Temporal Cross-Attention layer integrates the modulated historical context queue $\{\tilde{\mathbf{Q}}_{k-i}\}_{i=1}^{n}$, which is aligned by the TTM to ensure noise-level consistency. The complete temporal interaction pipeline are elaborated in Sec. 3.4.

Unified Self-Attention Layer. To enable cross-task reasoning, the joint queries at k-th frame $\mathbf{Q}_k = \{\mathbf{Q}^a, \mathbf{Q}^m, \mathbf{Q}^p\}$ are fed into a Unified Self-Attention Layer, allowing each token to attend to all others regardless of task origin (agent, map, or planning). This all-to-all interaction jointly updates agent motion, map geometry and ego-planning states, enabling the model to exploit cross-task geometric priors during generation. For example, ego planning $\mathbf{Q}^p$ and agent intentions $\mathbf{Q}^a$ impose spatial constraints that assist map reconstruction $\mathbf{Q}^m$ in occluded regions, while map topology regularizes multi-modal trajectory generation. A single joint attention pass thus promotes globally plausible, physically and logically consistent scene synthesis.

Spatial Deformable Attention layer. To anchor latent features in 3D space, each query $q \in \mathbf{Q}_k$ is linked to a 3D anchor $a \in \mathbf{A}$, which is projected onto the v-th camera image plane via the view transformation $\mathcal{P}_v$. We then apply the Spatial Deformable Attention, implemented via Multi-Scale Deformable Attention [51], to efficiently sample and aggregate multi-view, multi-scale image features $\mathbf{F}$.

3.4 Temporal Interaction via Temporal Transition Module

The Temporal Interaction via Temporal Transition Module (TTM) (Fig. 3) is the key component enabling our streaming unified diffusion framework. While many diffusion-based models neglect temporal dynamics, we use a memory bank for long-range reasoning. A naive fusion of historical features, however, suffers from noise-level mismatch between historical queries $\{\mathbf{Q}_{k-i}\}_{i=1}^{n}$ and current queries

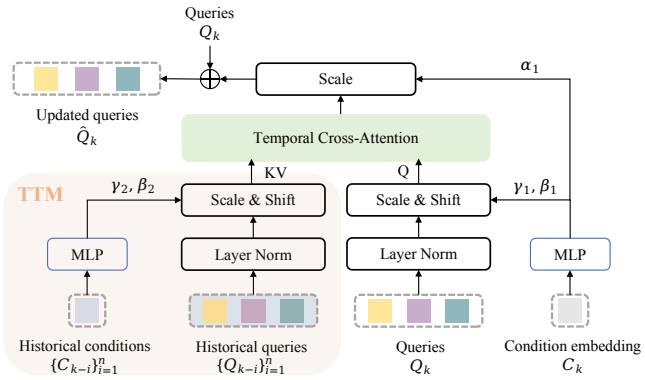

Fig. 3: Temporal Interaction. The core TTM enables a streaming unified diffusion framework by aligning historical queries with the current noising level.

$\mathbf{Q}_k$, as their denoising time steps are sampled stochastically and independently. Addressing this issue, we implement a structured interaction pipeline:
Input. The four inputs at the k-th frame with noise level t_k are:

- Queries $\mathbf{Q}_k$: joint noisy queries targeted for denoising at noise level t_k.
- Condition embedding $\mathbf{C}_k$: $\mathbf{C}_k = \texttt{MLP}(\texttt{TE}(t_k))$, where $\texttt{TE}(\cdot)$ is timestep embedding following DiT [37] and ADM [13].
- Historical queries $\{\mathbf{Q}_{k-i}\}_{i=1}^n$: queued past queries from the memory bank.
- Historical conditions $\{\mathbf{C}_{k-i}\}_{i=1}^n$: each condition is embedded as:

$$\mathbf{C}_{k-i} = \texttt{MLP}\big([\texttt{TE}(t_{k-i}), \texttt{TE}(\Delta t_i), \texttt{TE}(\Delta k_i)]\big) \tag{2}$$

  where t_{k-i} is the historical noise level, $\Delta t_i = t_k - t_{k-i}$ is the relative noise scale shift, and Δk_i is the frame interval.

Temporal Transition Module (TTM). TTM aligns historical queries with the current denoising stage. Through an MLP, each historical condition $\mathbf{C}_{k-i}$ generates scale and shift parameters, which then transform the historical queries:

$$\begin{aligned} \gamma_{k-i}, \beta_{k-i} &= \texttt{MLP}(\mathbf{C}_{k-i}), \\ \tilde{\mathbf{Q}}_{k-i} &= (1+\gamma_{k-i}) \odot \text{LayerNorm}(\mathbf{Q}_{k-i}) + \beta_{k-i}. \end{aligned} \tag{3}$$

This way, the generated temporal queries $\{\tilde{\mathbf{Q}}_{k-i}\}_{i=1}^n$ effectively re-project historical features into the current noise manifold for optimized cross-frame reasoning.
Temporal Fusion. With the inputs aligned, the interaction is executed through the Temporal Cross-Attention layer introduced in Sec. 3.3. This layer takes the modulated current queries $\tilde{\mathbf{Q}}_k$ and the TTM-aggregated features $\{\tilde{\mathbf{Q}}_{k-i}\}_{i=1}^n$ for interaction. This allows each task instance to inherit motion and structural priors from the past n frames without being disrupted by the noise-level mismatch across different physical timestamps.

Algorithm 1 Anchor Refresh Strategy at the j-th Iteration

Input: Noisy anchors a_t at diffusion step t; anchor priors $\mathbf{A} = \{a^{(i)}\}_{i=1}^{N_{\text{all}}}$, where $N_{\text{all}} = N_a + N_m + N_p$; step size m; confidence threshold τ.
Output: Refined anchors $\{a_{t-m}^{(i)}\}_{i=1}^{N_{\text{all}}}$ for the subsequent diffusion step.

// *Prediction and confidence estimation*
$\{\hat{y}_t^{(i)}, s_t^{(i)}\} = \Phi(\{a_t^{(i)}\})$;
for each anchor index $i \in \{1, \ldots, N_{\text{all}}\}$ **do**
  if $s_t^{(i)} > \tau$ **then**
    // *Selective denoising for high-confidence predictions*
    $a_{t-m}^{(i)} \leftarrow \text{DDIM}(a_t^{(i)}, \hat{y}_t^{(i)}, t, t-m)$;
  else
    // *Anchor refresh for low-confidence predictions*
    $a_{t-m}^{(i)} \sim \mathcal{N}\Big(\sqrt{\bar{\alpha}_{t-m}}\, a^{(i)}, (1-\bar{\alpha}_{t-m})\, \mathbf{I}\Big)$;
  end if
end for
return $\{a_{t-m}^{(i)}\}_{i=1}^{N_{\text{all}}}$;

3.5 Anchor Refresh Strategy

Training Inference Inconsistencies. Sparse query-based diffusion planning approaches [31, 48] often suffer from severe training–inference inconsistencies. During training, only a small subset of planning queries matched to the ground truth is supervised, while the remaining queries receive little or no guidance. At inference, however, all queries are updated at each denoising step, including previously unexplored queries that are fed back into subsequent iterations. This causes a progressive drift in the query distribution, gradually deviating from the anchor distribution observed during training. As identified in DiffusionDet [6], directly propagating these unconstrained, stochastic outputs through iterative steps causes a significant training-inference misalignment. Therefore, without explicit mechanisms to correct this misalignment, the discrepancy accumulates over denoising steps, leading to increasingly degraded predictions.

Inference via Anchor Refresh Strategy. Inspired by DiffusionDet [6], we propose an Anchor Refresh strategy (ARS) to address the training–inference inconsistencies. During inference, high-confidence queries are retained, while low-confidence queries are refreshed by resampling from the original noise distribution (Algorithm 1).

- **Selective Denoising:** At each denoising step, we identify high-confidence predictions using a task-specific threshold τ. For queries with confidence $s_t^{(i)} > \tau$, we apply the standard DDIM reverse operator to transition from $a_t^{(i)}$ to $a_{t-m}^{(i)}$ using on the predicted clean state $\hat{y}_t^{(i)}$. This ensures that only task-relevant information is propagated.
- **Anchor-based Distribution Refresh:** For low-confidence queries $s_t^{(i)} \leq \tau$, we discard the unconstrained outputs and replace them with their initial anchors $a^{(i)}$, re-projected to the prescribed noise level at step $t-m$.

With the above operations, ARS preserves query diversity while maintaining alignment with the training distribution, enabling stable and robust inference in complex multi-modal scenarios.

3.6 Loss Function

The entire UniTeD framework is trained end-to-end in a fully differentiable way. To ensure the overall performance of perception and planning in driving, we define a multi-task optimization objective that covers four primary tasks: detection, motion prediction, mapping, and planning. Each task can be optimized using both classification and regression losses with corresponding weight. The total loss $\mathcal{L}_{\text{total}}$ is defined as a weighted combination of task-specific components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{mot}}\mathcal{L}_{\text{mot}} + \lambda_{\text{map}}\mathcal{L}_{\text{map}} + \lambda_{\text{plan}}\mathcal{L}_{\text{plan}}. \tag{4}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate UniTeD on three benchmarks to validate mid-term simulation and interactive closed-loop performance.

NAVSIM v1 & v2. Built on the nuPlan [2] dataset, NAVSIM [12] provides a non-reactive mid-term simulation with a 4-second horizon, bridging the gap between open-loop and closed-loop evaluation. Performance is measured via the Predictive Driver Model Score (PDMS) [12] and Extended PDMS (EPDMS) [3], which incorporate safety penalties alongside efficiency and comfort rewards.

Bench2Drive. Bench2Drive [20] is a large-scale, CARLA-based [14] closed-loop benchmark featuring 44 scenarios across 220 intersections. It tests interactive driving and long-tail robustness. We report the Driving Score (DS) and Success Rate (SR) as primary indicators.

Implementation Details. We employ ResNet-34 as the backbone for feature extraction. Input images are resized to a resolution of 640×352. For the NAVSIM benchmark [3, 12], the model processes inputs from eight multi-view cameras, while for Bench2Drive [20], six multi-view cameras are utilized. Our decoder consists of six cascaded diffusion decoder layers. Following the design of DiffusionDrive [31], we adopt a truncated diffusion policy operating on 1,024 object anchors, categorized into 900 agent anchors, 100 map anchors, and 24 planning anchors. During training, the diffusion schedule is truncated at 50 out of 1,000 steps to diffuse the anchors. For inference, we use 2 denoising steps with step size m of 10 and select the top-1 scoring trajectory as the final prediction. The confidence threshold τ for each task is detailed in the Appendix. The model is trained on 8 NVIDIA L40 GPUs with a total batch size of 64. We use the AdamW optimizer with an initial learning rate of 4×10^{-4} and train for 100 epochs to ensure convergence across benchmarks.

Table 1: Performance comparison with other state-of-the-art methods on the NAVSIMv1 benchmark. All compared methods utilize ResNet-34 as the CNN backbone. The best results are highlighted in **bold**. The second best is underline, and the third best is in *italics*. Dis and Gen denote discriminative and generative formulations, respectively. Sep and Uni indicate whether perception and planning are formulated separately or in a unified manner. C and L denote Camera and LiDAR modalities.

Method	Paradigm	Modality	Planning Metrics					Overall
			NC↑	DAC↑	EP↑	TTC↑	COMF↑	PDMS↑
VADv2 [5]	Dis-Sep	C	97.2	89.1	76.0	91.6	100.0	80.9
UniAD [17]	Dis-Sep	C	97.8	91.9	78.8	92.9	100.0	83.4
TransFuser [9]	Dis-Sep	C+L	97.7	92.8	79.2	92.8	100.0	84.0
Hydra-MDP [28]	Dis-Sep	C+L	98.3	96.0	78.7	94.6	100.0	86.5
Hydra-MDP++ [25]	Dis-Sep	C	97.6	96.0	80.4	93.1	100.0	86.6
WoTE [27]	Dis-Sep	C	98.5	*96.8*	81.9	94.9	99.9	88.3
DiffusionDrive [31]	Gen-Sep	C+L	98.2	96.2	82.2	94.7	100.0	88.1
GaussianFusion [34]	Gen-Sep	C+L	98.3	97.2	*83.0*	94.6	100.0	*88.8*
DiffRefiner [48]	Gen-Sep	C	*98.4*	**97.4**	83.4	*95.3*	100.0	89.4
PARA-Drive [46]	Dis-Uni	C	97.9	92.4	79.3	93.0	99.8	84.0
HiP-AD [21]	Dis-Uni	C	**98.9**	96.7	81.2	96.3	99.9	88.6
UniTeD	Gen-Uni	C	**98.9**	97.2	**84.1**	**96.6**	**100.0**	**90.2**

Table 2: Performance comparison with other state-of-the-art methods on the NAVSIMv2 benchmark. The best results are highlighted in **bold**. The second best is underline, and the third best is in *italics*.

Method	Basic Planning ↑				Safety ↑			Comfort ↑		Overall
	NC	DAC	DDC	TL	EP	TTC	LK	HC	EC	EPDMS↑
TransFuser [9]	96.9	89.9	97.8	*99.7*	87.1	95.4	92.7	*98.3*	87.2	76.7
Hydra-MDP++ [25]	97.2	**97.5**	99.4	99.6	83.1	96.5	94.4	98.2	70.9	81.4
DriveSuprim [47]	97.5	96.5	99.4	99.6	**88.4**	96.6	95.5	*98.3*	77.0	83.1
DiffusionDrive [31]	98.2	96.2	98.6	-	87.6	97.3	*97.4*	98.4	-	84.0
GaussianFusion [34]	*98.3*	*97.3*	*99.0*	-	*87.5*	*97.4*	*97.4*	*98.3*	-	*85.0*
DiffRefiner [48]	98.5	97.4	**99.6**	99.8	87.6	97.7	97.7	*98.3*	*86.2*	86.2
UniTeD	**99.1**	97.2	**99.6**	**99.9**	86.9	**98.5**	**98.4**	**99.9**	**87.3**	**90.1**

4.2 Main Results

Results on NAVSIM v1. To ensure a fair comparison, all baseline models are evaluated using a ResNet-34 backbone. Methods based on reinforcement learning or Vision-Language Models (VLMs) are excluded from the main analysis, with detailed comparisons provided in the Appendix. As summarized in Tab. 1, our proposed UniTeD achieves state-of-the-art performance on the NAVSIM v1 benchmark, yielding a PDMS of 90.2. UniTeD consistently outperforms existing approaches across different methodological paradigms. Specifically, UniTeD surpasses the best discriminative-separate models, including WoTE [27] (88.3) and Hydra-MDP++ [25] (86.6), by margins of +1.9 and +3.6 PDMS, respectively. These improvements highlight the advantage of generative modeling over deterministic regression in capturing complex and diverse driving behaviors. Compared with the diffusion-based DiffRefiner [48] (89.4), which adopts a generative-separate pipeline, UniTeD still achieves a further gain of +0.8 PDMS. In compar-

Table 3: Comparison on the Bench2Drive online leaderboard.

Method	Training Data	DS ↑	SR(%) ↑
VAD [23]	B2D (200K)	42.2	15.0
UniAD [17]	B2D (200K)	45.8	16.4
SparseDrive [42]	B2D (200K)	44.5	16.7
DriveTransformer [21]	B2D (200K)	63.5	35.0
HiP-AD [43]	B2D (200K)	86.8	69.1
TF++ [52]	TF++ (500K)	84.2	67.3
DiffusionDrive [31]	TF++ (500K)	77.7	52.7
DiffRefiner [48]	TF++ (500K)	87.1	**71.4**
UniTeD	B2D (200K)	**87.3**	70.0

Table 4: Perception and Prediction performance on the nuScenes validation set.

Method	Detection		Map	Track	Motion
	mAP↑	NDS↑	mAP↑	AMOTA↑	minADE↓
UniAD [17]	0.380	0.359	-	-	-
VAD [23]	0.276	0.397	0.476	-	-
SparseDrive-S [42]	0.418	0.525	0.551	0.386	0.62
DiFSD [41]	0.410	0.528	0.560	-	-
HiP-AD [43]	**0.424**	0.535	0.571	0.406	0.61
UniTeD	**0.424**	**0.537**	**0.596**	**0.419**	**0.58**

ison to the unified discriminative counterpart HiP-AD [43] (88.6), our method delivers a +1.6 improvement, demonstrating the effectiveness of formulating both perception and planning queries within a unified diffusion framework. Notably, relying solely on camera (C) inputs, UniTeD outperforms all competitive LiDAR-based (C+L) approaches, demonstrating its efficacy and robustness.

Results on NAVSIM v2. As shown in Tab. 2, UniTeD maintains its superior performance under the more challenging NAVSIM v2 benchmark, achieving a state-of-the-art EPDMS of 90.1. In particular, our method outperforms the strongest generative baseline, DiffRefiner [48] (86.2), by a substantial margin of +3.9 points. Such consistent and significant gains over all major competitors further validate the enhanced planning capability and robustness of our framework in complex, real-world driving scenarios.

Results on Bench2Drive. To evaluate the closed-loop planning capability of our model in reactive environments, we conduct experiments on the Bench2Drive leaderboard. As reported in Tab. 3, UniTeD achieves a state-of-the-art Driving Score (DS) of 87.3, outperforming all existing approaches. When trained on the standard Bench2Drive [20] 200K dataset, UniTeD consistently surpasses contemporary unified methods, including HiP-AD [43] (86.8 DS) and DriveTransformer [21] (63.5 DS), achieving improvements of +0.5 and +23.8 DS, respectively. Notably, even compared with models trained on the substantially larger and higher-quality TF++ [52] (500K) dataset, UniTeD maintains its advantage. It outperforms the recent generative model DiffRefiner [48] by +0.2 DS and DiffusionDrive [31] by +9.6 DS. These results highlight the strong robustness and generalization capability of UniTeD.

Results on nuScenes. We further evaluate the perception capability of UniTeD on nuScenes [1]. As shown in Tab. 4, the results demonstrate that UniTeD is not only a strong planner but also a versatile framework capable of maintaining high-fidelity environmental understanding. Specifically, UniTeD achieves 0.596 mAP on map segmentation, outperforming HiP-AD [43] by +2.5% in map mAP. For 3D object detection, it attains a leading 0.537 NDS. In motion prediction, UniTeD achieves the lowest minADE of 0.58. Compared with SparseDrive-S [42], it reduces motion prediction error by 6.4% (from 0.62 to 0.58). Moreover, UniTeD reaches 0.419 AMOTA, surpassing HiP-AD by +1.3% and SparseDrive-S by

Table 5: Ablation study of different paradigms and task components. Perc and Plan denote perception and planning modules. Sep and Uni indicate whether perception and planning are formulated separately or in a unified manner. Regr and Diff denote regression-based and diffusion-based approaches, respectively.

ID	Paradigm		Perc	Plan	NAVSIMv1 Metrics					
	Sep	Uni			NC	DAC	EP	TTC	C	PDMS
0	✓		Regr	Regr	97.8	91.9	79.2	92.8	100.0	84.1
1	✓		Regr	Diff	97.6	96.0	80.4	93.1	100.0	86.7
2		✓	Regr	Regr	98.8	96.7	81.1	96.3	99.9	88.5
3		✓	Regr	Diff	98.8	97.0	82.7	**96.7**	99.9	89.5
4		✓	Diff	Diff	**99.0**	**97.2**	**84.1**	96.6	**100.0**	**90.2**

Table 6: Ablation study of TTM in unified paradigm. Regr and Diff are defined the same as in Tab. 5. Memory denotes the Memory Queue.

ID	Unified Paradigm		Memory	TTM	NAVSIMv1 Metrics					
	Regr	Diff			NC	DAC	EP	TTC	C	PDMS
0	✓		✗	-	98.9	96.0	81.0	96.4	99.9	88.2
1	✓		✓	-	98.8	96.7	81.1	96.3	99.9	88.5
2		✓	✗	✗	98.8	96.8	82.2	96.4	99.9	89.0
3		✓	✓	✗	99.0	97.1	82.5	96.5	99.9	89.4
4		✓	✓	✓	**99.0**	**97.2**	**84.1**	**96.6**	**100.0**	**90.2**

Table 7: Ablation study on the Anchor Refresh Strategy.

ID	ARS	NAVSIMv1 Metrics					
		NC	DAC	EP	TTC	C	PDMS
0	✗	98.4	96.8	81.7	95.2	99.9	88.2
1	✓	**99.0**	**97.2**	**84.1**	**96.6**	**100.0**	**90.2**

+3.3%. These results collectively confirm that UniTeD serves as a comprehensive and unified framework for perception, prediction, and planning.

4.3 Ablation Study

Effect of Unified Diffusion. We conduct a comprehensive ablation study to validate the effectiveness of each design component, as reported in Tab. 5. By progressively incorporating the unified architecture and diffusion-based modeling, performance steadily improves, ultimately reaching a peak PDMS of 90.2. Specifically, using the diffusion-based planning module, comparing the separate perception–planning setup to the unified perception–planning framework (ID 1 vs. ID 3) shows that transitioning to a unified paradigm yields a substantial gain of +2.8 PDMS. This result confirms that joint learning across perception and planning alleviates the information bottleneck and mitigates error accumulation commonly observed in sequential pipelines. Under the separate paradigm (ID 0 vs. ID 1), introducing diffusion-based planning improves PDMS from 84.1

to 86.7 (+2.6). Similarly, within the unified paradigm (ID 2 vs. ID 3), diffusion modeling provides an additional +1.0 gain. These findings highlight the advantage of diffusion processes over deterministic regression in modeling the inherently multi-modal nature of human driving behavior. Finally, extending diffusion modeling to both perception and planning (ID 4) achieves the best overall performance, surpassing the variant with diffusion-based planning only (ID 3) by +0.7 PDMS. This demonstrates that diffusion-based perception yields more robust and high-fidelity feature representations, which in turn establish a stronger foundation for downstream planning. Overall, the synergy introduced by the Unified Diffusion framework enables the model to achieve both precise environmental understanding and safe, human-like trajectory generation, outperforming conventional regression-based unified approaches.

Effect of TTM. We systematically analyze the components of TTM in Tab. 6. The results confirm that both temporal context modeling and timestep alignment are critical for robust planning. Under the regression paradigm (ID 0 vs. ID 1), incorporating Memory Queue yields a +0.3 PDMS improvement. A consistent gain is also observed in the diffusion paradigm (ID 2 vs. ID 3), where PDMS increases from 89.0 to 89.4 (+0.4). The results show that historical information provides essential temporal continuity for both perception and planning, regardless of the underlying framework. Notably, even without temporal memory, the diffusion-based model (ID 2, 89.0) already outperforms the memory-enhanced regression counterpart (ID 1, 88.5) by +0.5 PDMS, further highlighting the superior capacity of generative modeling in capturing complex driving distributions. The most significant improvement arises from the full TTM design. Comparing ID 3 and ID 4, introducing TTM increases PDMS from 89.4 to 90.2. While the Memory Queue supplies raw historical context, TTM addresses feature–denoising misalignment by synchronizing latent representations with the current diffusion timestep t, thereby enabling more coherent temporal reasoning.

Effect of Anchor Refresh Strategy. We examine the impact of ARS in Tab. 7. The results show that ARS is crucial for maintaining consistency within the sparse diffusion architecture. Without this mechanism (ID 0), the model relies solely on the initial trunk anchors, leading to a noticeable training–inference discrepancy. Introducing ARS improves PDMS from 88.2 to 90.2 (ID 0 vs. ID 1). These findings indicate that ARS is a fundamental component for ensuring stable and consistent input distribution.

5 Conclusion

In this paper, we introduced UniTeD, a unified diffusion framework that seamlessly integrates perception and planning within a shared generative denoising process. By enabling bidirectional interaction and mutual refinement, UniTeD alleviates error propagation and strengthens multi-modal reasoning. We further extend the framework to a streaming setting through a Temporal Transition Module that resolves inter-frame noise mismatch, together with an Anchor Refresh Strategy to mitigate training–inference distribution shift. Extensive exper-

iments on challenging benchmarks demonstrate that UniTeD achieves state-of-the-art performance, surpassing both discriminative end-to-end and planning-only diffusion approaches.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
2. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810 (2021)
3. Cao, W., Hallgarten, M., Li, T., Dauner, D., Gu, X., Wang, C., Miron, Y., Aiello, M., Li, H., Gilitschenski, I., et al.: Pseudo-Simulation for Autonomous Driving. arXiv preprint arXiv:2506.04218 (2025)
4. Chen, J., Deng, R., Furukawa, Y.: Polydiffuse: Polygonal Shape Reconstruction via Guided Set Diffusion Models. Advances in Neural Information Processing Systems **36**, 1863–1888 (2023)
5. Chen, S., Jiang, B., Gao, H., Liao, B., Xu, Q., Zhang, Q., Huang, C., Liu, W., Wang, X.: Vadv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning. arXiv preprint arXiv:2402.13243 (2024)
6. Chen, S., Sun, P., Song, Y., Luo, P.: DiffusionDet: Diffusion Model for Object Detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19830–19843 (2023)
7. Cheng, J., Chen, Y., Chen, Q.: PLUTO: Pushing the Limit of Imitation Learning-based Planning for Autonomous Driving. arXiv preprint arXiv:2404.14327 (2024)
8. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025)
9. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A.: TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 12878–12895 (2022)
10. Cui, A., Casas, S., Sadat, A., Liao, R., Urtasun, R.: Lookout: Diverse Multi-Future Prediction and Planning for Self-Driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16107–16116 (2021)
11. Dauner, D., Hallgarten, M., Geiger, A., Chitta, K.: Parting with Misconceptions about Learning-based Vehicle Motion Planning. In: Conference on Robot Learning. pp. 1268–1281. PMLR (2023)
12. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., et al.: NAVSIM: Data-Driven Non-Reactive Autonomous Vehicle Simulation and Benchmarking. Advances in Neural Information Processing Systems **37**, 28706–28719 (2024)
13. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
14. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An Open Urban Driving Simulator. In: Conference on Robot Learning. pp. 1–16. PMLR (2017)

15. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11525–11533 (2020)
16. Gu, J., Hu, C., Zhang, T., Chen, X., Wang, Y., Wang, Y., Zhao, H.: ViP3D: End-to-end Visual Trajectory Prediction via 3D Agent Queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5496–5506 (2023)
17. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17853–17862 (2023)
18. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View. arXiv preprint arXiv:2112.11790 (2021)
19. Jia, P., Wen, T., Luo, Z., Yang, M., Jiang, K., Liu, Z., Tang, X., Lei, Z., Cui, L., Zhang, B., et al.: DiffMap: Enhancing Map Segmentation with Map Prior Using Diffusion Model. IEEE Robotics and Automation Letters **9**(11), 9836–9843 (2024)
20. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-To-End Autonomous Driving. Advances in Neural Information Processing Systems **37**, 819–844 (2024)
21. Jia, X., You, J., Zhang, Z., Yan, J.: DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. arXiv preprint arXiv:2503.07656 (2025)
22. Jiang, B., Chen, S., Wang, X., Liao, B., Cheng, T., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C.: Perceive, Interact, Predict: Learning Dynamic and Static Clues for End-to-End Motion Prediction. arXiv preprint arXiv:2212.02181 (2022)
23. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
24. Jiang, C., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D., et al.: MotionDiffuser: Controllable Multi-Agent Motion Prediction using Diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9644–9653 (2023)
25. Li, K., Li, Z., Lan, S., Xie, Y., Zhang, Z., Liu, J., Wu, Z., Yu, Z., Alvarez, J.M.: Hydra-MDP++: Advancing End-to-End Driving via Expert-Guided Hydra-Distillation. arXiv preprint arXiv:2503.12820 (2025)
26. Li, Q., Wang, Y., Wang, Y., Zhao, H.: HDMapNet: An Online HD Map Construction and Evaluation Framework. In: International Conference on Robotics and Automation. pp. 4628–4634. IEEE (2022)
27. Li, Y., Wang, Y., Liu, Y., He, J., Fan, L., Zhang, Z.: End-to-End Driving with Online Trajectory Evaluation via BEV World Model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27137–27146 (2025)
28. Li, Z., Li, K., Wang, S., Lan, S., Yu, Z., Ji, Y., Li, Z., Zhu, Z., Kautz, J., Wu, Z., et al.: Hydra-MDP: End-to-end Multimodal Planning with Multi-target Hydra-Distillation. arXiv preprint arXiv:2406.06978 (2024)
29. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 2020–2036 (2024)

30. Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C.: MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction. arXiv preprint arXiv:2208.14437 (2022)
31. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., et al.: DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12037–12047 (2025)
32. Liao, B., Chen, S., Zhang, Y., Jiang, B., Zhang, Q., Liu, W., Huang, C., Wang, X.: MapTRv2: An End-to-End Framework for Online Vectorized HD Map Construction. International Journal of Computer Vision **133**(3), 1352–1374 (2025)
33. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4D: Multi-view 3D Object Detection with Sparse Spatial-Temporal Fusion. arXiv preprint arXiv:2211.10581 (2022)
34. Liu, S., Liang, Q., Li, Z., Li, B., Huang, K.: GaussianFusion: Gaussian-Based Multi-Sensor Fusion for End-to-End Autonomous Driving. arXiv preprint arXiv:2506.00034 (2025)
35. Liu, Y., Wang, T., Zhang, X., Sun, J.: PETR: Position Embedding Transformation for Multi-View 3D Object Detection. In: European Conference on Computer Vision. pp. 531–548. Springer (2022)
36. Luo, R., Song, Z., Ma, L., Wei, J., Yang, W., Yang, M.: DiffusionTrack: Diffusion Model For Multi-Object Tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3991–3999 (2024)
37. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
38. Sadat, A., Casas, S., Ren, M., Wu, X., Dhawan, P., Urtasun, R.: Perceive, Predict, and Plan: Safe Motion Planning Through Interpretable Semantic Representations. In: European Conference on Computer Vision. pp. 414–430. Springer (2020)
39. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. arXiv preprint arXiv:2010.02502 (2020)
40. Song, Z., Jia, C., Liu, L., Pan, H., Zhang, Y., Wang, J., Zhang, X., Xu, S., Yang, L., Luo, Y.: Don't Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22432–22441 (2025)
41. Su, H., Wu, W., Yan, J.: DiFSD: Ego-Centric Fully Sparse Paradigm with Uncertainty Denoising and Iterative Refinement for Efficient End-to-End Self-Driving. arXiv preprint arXiv:2409.09777 (2024)
42. Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S.: SparseDrive: End-to-End Autonomous Driving via Sparse Scene Representation. In: 2025 IEEE International Conference on Robotics and Automation. pp. 8795–8801. IEEE (2025)
43. Tang, Y., Xu, Z., Meng, Z., Cheng, E.: HiP-AD: Hierarchical and Multi-Granularity Planning with Deformable Attention for Autonomous Driving in a Single Decoder. arXiv preprint arXiv:2503.08612 (2025)
44. Wang, T., Zhang, C., Qu, X., Li, K., Liu, W., Huang, C.: DiffAD: A Unified Diffusion Modeling Approach for Autonomous Driving. arXiv preprint arXiv:2503.12170 (2025)
45. Wang, Z., Zhang, W., Zhang, W., Tan, X., Liu, H., Wang, Y., Li, G.: LaneDiffusion: Improving Centerline Graph Learning via Prior Injected BEV Feature Generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27052–27062 (2025)

46. Weng, X., Ivanovic, B., Wang, Y., Wang, Y., Pavone, M.: PARA-Drive: Parallelized Architecture for Real-time Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15449–15458 (2024)
47. Yao, W., Li, Z., Lan, S., Wang, Z., Sun, X., Alvarez, J.M., Wu, Z.: DriveSuprim: Towards Precise Trajectory Selection for End-to-End Planning. arXiv preprint arXiv:2506.06659 (2025)
48. Yin, L., Ju, R., Guo, G., Cheng, E.: DiffRefiner: Coarse to Fine Trajectory Planning via Diffusion Refinement with Semantic Interaction for End to End Autonomous Driving. arXiv preprint arXiv:2511.17150 (2025)
49. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: MOTR: End-to-End Multiple-Object Tracking with Transformer. In: European Conference on Computer Vision. pp. 659–675. Springer (2022)
50. Zheng, Z., Chen, S., Yin, H., Zhang, X., Zou, J., Wang, X., Zhang, Q., Zhang, L.: ResAD: Normalized Residual Trajectory Modeling for End-to-End Autonomous Driving. arXiv preprint arXiv:2510.08562 (2025)
51. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv preprint arXiv:2010.04159 (2020)
52. Zimmerlin, J., Beißwenger, J., Jaeger, B., Geiger, A., Chitta, K.: Hidden Biases of End-to-End Driving Datasets. arXiv preprint arXiv:2412.09602 (2024)
53. Zou, J., Chen, S., Liao, B., Zheng, Z., Song, Y., Zhang, L., Zhang, Q., Liu, W., Wang, X.: DiffusionDriveV2: Reinforcement Learning-Constrained Truncated Diffusion Modeling in End-to-End Autonomous Driving. arXiv preprint arXiv:2512.07745 (2025)