WPT: World-to-Policy Transfer via Online World Model Distillation

Guangfeng Jiang[1], Yueru Luo[2], Jun Liu[1✉], Yi Huang[2], Yiyao Zhu[3],
Zhan Qu[4], Dave Zhenyu Chen[4], Bingbing Liu[4], Xu Yan[4✉]
[1]University of Science and Technology of China,
[2]CUHK-SZ, [3]HKUST, [4]Huawei Foundation Model Dept

Abstract

*Recent years have witnessed remarkable progress in world models, which primarily aim to capture the spatio-temporal correlations between an agent's actions and the evolving environment. However, existing approaches often suffer from tight runtime coupling or depend on offline reward signals, resulting in substantial inference overhead or hindering end-to-end optimization. To overcome these limitations, we introduce **WPT**, a World-to-Policy Transfer training paradigm that enables online distillation under the guidance of an end-to-end world model. Specifically, we develop a trainable reward model that infuses world knowledge into a teacher policy by aligning candidate trajectories with the future dynamics predicted by the world model. Subsequently, we propose policy distillation and world reward distillation to transfer the teacher's reasoning ability into a lightweight student policy, enhancing planning performance while preserving real-time deployability. Extensive experiments on both open-loop and closed-loop benchmarks show that our WPT achieves state-of-the-art performance with a simple policy architecture: it attains a **0.11 collision rate** (open-loop) and achieves a **79.23 driving score** (closed-loop), surpassing both world-model-based and imitation-learning methods in accuracy and safety. Moreover, the student sustains up to **4.9× faster** inference, while retaining most of the gains.*

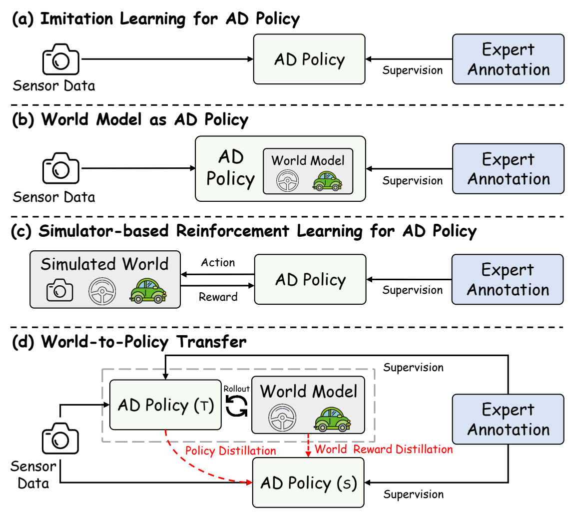

Figure 1. **Different training paradigms of AD policy with world model.** (a) Imitation learning where the policy is trained using expert supervision. (b) World model (WM) directly integrated into the AD policy for enhanced feature evolution and trajectory reasoning. (c) Simulator-based reinforcement learning for AD policy training using a simulated world. (d) Our WPT, where the policy interacts with the WM during training, with both the teacher policy (T) and the student policy (S) leveraging the WM for knowledge transfer. After training, the WM will be discarded.

1. Introduction

Recent advances in autonomous driving have increasingly centered on world models that learn to capture the spatiotemporal dynamics of complex driving environments [30, 37, 45, 54, 59, 65]. Unlike imitation-based frameworks [18, 19, 24] that focus on replicating expert behavior, world-model-based methods explicitly model the causal interactions between agents and their surroundings, enabling anticipatory reasoning and more reliable long-horizon planning.

A world model (WM) generally predicts future scenarios from past observations and was originally introduced for simulated control and robotic applications [6, 13–16, 36, 47]. Based on their role in enhancing autonomous driving (AD) policies (Fig. 1(a)), world models can be broadly categorized into two types: The first category of methods directly integrates the world model into the driving policy (Fig. 1(b)), enabling more powerful feature evolution [54, 59] and trajectory reasoning [30, 37, 45, 65]. Specifically, Drive-OccWorld [54] jointly models occupancy and flow prediction to support safe and interpretable trajectory planning. Alternatively, methods such as WoTE [30] and

✉ Corresponding authors.

DriveDPO [37] integrate an implicit bird-eye-view (BEV) world model with a learned reward evaluator to assess multiple trajectory candidates according to their predicted future states. Despite their performance gains, these methods exhibit a strong dependence on accurate future predictions and autoregressive rollouts, where sequential dependencies between prediction steps substantially hinder real-time efficiency. The second category of methods treats the world model as a simulator to enable closed-loop reinforcement learning for AD policy training [3, 7, 8, 34, 42, 60], as shown in Fig. 1(c). Nevertheless, such approaches are highly dependent on the fidelity of the simulator's generated data and are evaluated primarily in synthetic environments.

A natural question arises: *how can the AD policy leverage the world knowledge while avoiding extra computational overhead?* To this end, we present **WPT** (World-to-Policy Transfer in Fig. 1(d)), a novel training paradigm where the policy interacts with a world model during training to acquire predictive awareness of future dynamics, while maintaining real-time efficiency through a simple policy network during deployment. Concretely, for a given AD policy, we utilize a world model to capture the spatiotemporal evolution of the environment from its learned representations. We then design a trainable, interaction-based reward model that evaluates each candidate trajectory according to its consistency with the predicted future world states, thereby enabling the selection of the optimal trajectory. The proposed interaction mechanism allows the end-to-end model to anticipate future environmental dynamics, enabling the policy to internalize world-model knowledge through predicted evolutions. Furthermore, to satisfy real-time requirements, we introduce policy distillation and world reward distillation to transfer the reasoning capability of the large world model to a lightweight policy network, enabling fast inference while enhancing performance through knowledge transfer.

Overall, our **WPT** framework offers three key advantages: **(1) Interpretability.** It achieves end-to-end optimization through a reward-guided mechanism that bridges prediction and planning, resulting in world-consistent and explainable driving behaviors. **(2) Efficiency.** It preserves real-time performance through distillation, transferring the reasoning ability of the large world model to a lightweight planner for fast inference (up to **4.9**$\times$ speedup). **(3) Effectiveness.** Experiments across both open-loop and closed-loop benchmarks show that WPT delivers state-of-the-art results (open-loop: **0.11%** collision, closed-loop: **79.23** driving score) on different lightweight policies.

The main contributions are summarized as follows:

- We propose **WPT**, an online distillation paradigm that integrates world modeling with a trainable reward model for interpretable and world-consistent planning.
- We design policy distillation and world reward distillation, transferring reasoning capabilities from a large model to a lightweight one.
- Extensive experiments on diverse benchmarks and baselines demonstrate that WPT achieves state-of-the-art performance, with 0.61m L2 / 0.11% collision in open-loop and 79.23 driving score in closed-loop, outperforming existing world-model-based and imitation-learning-based methods. Moreover, these gains transfer to our student model with up to 4.9$\times$ faster inference.

2. Related Work

2.1. End-to-End Autonomous Driving

End-to-end autonomous driving directly maps raw multi-sensor inputs to future trajectories or low-level control commands [1]. From the perspective of output modalities, recent research can be broadly categorized into single-modal and multi-modal trajectory planning approaches.

Single-modal planning predicts one "best" trajectory, as seen in approaches like UniAD [19], which integrates multi-task modules in a unified framework optimized for planning. Further, VAD [24] replaces dense scene representations with vectorized ones. Subsequent works refine this line along several axes: *e.g.*, sparse representation to improve efficiency [11, 23, 40, 58], temporal modeling to stabilize trajectory outputs [38, 57], and handling of online map uncertainty to increase planning reliability [10, 53].

To handle multi-future ambiguities, recent works explore multi-modal planning. VADv2 [4] builds an anchor vocabulary for action spaces together with probabilistic planning to capture diverse future options. Hydra-MDP [27, 31] advances this via multi-expert distillation into a multi-head student, yielding trajectories aligned with distinct criteria. More recently, diffusion-based planners [37, 49, 64] have emerged as a strong paradigm for modeling multi-modal trajectory distributions, providing diversity naturally. Alongside generative modeling, the community has explored preference- or reward-driven selection to better align outputs with different objectives: safety-targeted selection [37], human-style alignment [26], *etc*.

2.2. World Models for Driving

The world models aim to learn a compact representation of the environment and predict future states based on an agent's actions and past observations [9, 33, 44, 45, 52, 61]. **Video World Model.** Video world models generate future visual frames or videos conditioned on candidate actions or trajectories, enabling planners to "see" what would happen under each choice and score options with perceptual (image-level) rewards. Particularly, Drive-WM [45] controllably generates multi-view videos under different maneuvers, and then selects trajectories via image-level rewards, showing the feasibility of WM-guided planning.

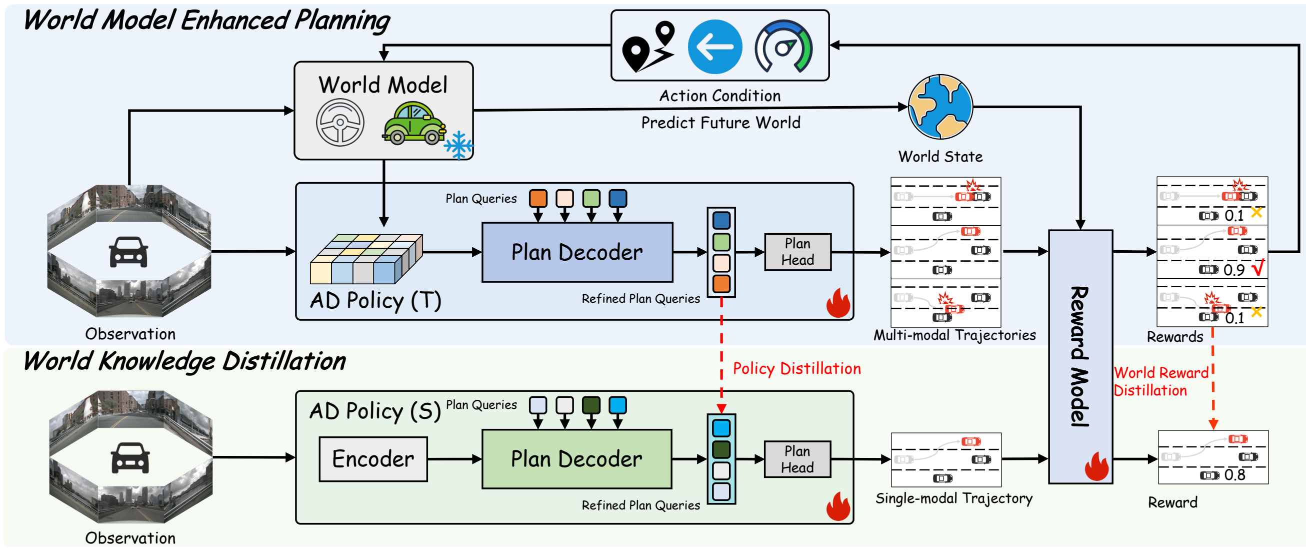

Figure 2. **Overview of WPT framework.** During training (top), the pretrained world model predicts future world under given action conditions, and the teacher AD policy (T) generates multi-modal trajectories. The *reward model* evaluates these trajectories to produce world reward. During distillation (bottom), the student AD policy (S) learns from the teacher through two mechanisms: (1) *policy distillation*, which aligns the planning representations between teacher and student; and (2) *world reward distillation*, which encourages the student to match the teacher's optimal reward trajectory in the predicted future world.

Most recently, Epona [59] adopts an autoregressive diffusion WM with spatial–temporal factorization for long-horizon rollouts, and integrates trajectory prediction with video generation for end-to-end planning.

Occupancy World Model. Occupancy world models evolve the 3D scene volumetric states over time, offering planner-friendly rollouts in physical space. OccWorld [61] tokenizes 3D occupancy and autoregressively predicts future occupancy and ego motion, enabling fine-grained scene evolution without dense labels. Drive-OccWorld [54] forecasts 4D occupancy and evaluates trajectories with occupancy costs. RenderWorld [51] encodes 3D occupancy into tokens, which a world model uses to forecast future 4D occupancy and ego motion.

Latent World Model. Latent world models roll out future latent features instead of pixels or voxels. LAW [29] self-supervises an action-conditioned latent WM by predicting visual latent features from current features and trajectories. WoTE [30] performs online trajectory evaluation using a BEV WM that forecasts future BEV states for multiple trajectory candidates. Their BEV WM predictions further are supervised via a traffic simulator, which also enables evaluating these trajectories.

Unlike these approaches, which rely on rollouts or online evaluation during deployment, WPT is a purely training-time paradigm. It provides interaction-based rewards and facilitates world knowledge transfer, enabling the deployed policy to plan in real time without additional overhead.

3. Method

This section outlines the four core components of our proposed WPT framework: (1) the Autonomous Driving policy (Sec. 3.1), (2) the World Model (Sec. 3.2), (3) the Reward Model (Sec. 3.3), and (4) the World Knowledge Distillation (Sec. 3.4). As shown in Fig. 2, WPT leverages the world model during training to guide the teacher AD policy in generating future-aware trajectories. The student policy then distills knowledge from both the teacher's planning representations and the world model's reward supervision.

3.1. Autonomous Driving Policy

This paper focuses on improving the planning performance of AD policies, employing a standard end-to-end architecture as depicted in Fig. 2. The model takes as input a temporal sequence of multi-view camera images and encodes them into a unified world representation F^{w} (*e.g.*, BEV features) that captures the spatial context of the driving environment.

Subsequently, a planning decoder $\mathcal{P}_D$ receives both the world representation F^{w} and a set of learnable planning queries Q, and refines the queries through cross-attention interaction:

$$\tilde{Q} = \mathcal{P}_D\left(Q, F^{\mathrm{w}}\right), \tag{1}$$

where $\mathcal{P}_D$ denotes a *CrossAttention*-based module that facilitates interaction between planning queries and spatial scene representations.

Finally, a lightweight MLP-based plan head $\mathcal{P}_h$ decodes the refined queries into predicted future trajectories $\hat{\mathcal{T}}$:

$$\hat{\mathcal{T}} = \mathcal{P}_h\left(\tilde{Q}\right). \tag{2}$$

To support both performance and real-time deployment, we design two types of policies as shown in Fig. 2.

Multi-modal Policy as a Teacher. The multi-modal policy first generates a set of candidate trajectories $\mathcal{T} = \{\tau_1, \ldots, \tau_N\}$ through the planning decoder, where N is the number of candidate trajectories [19]. These trajectories are evaluated through interaction with the world model to select the optimal trajectory under predicted future states.

Single-modal Policy as a Student. To support real-time inference, we introduce a lightweight policy that predicts the entire future trajectory $\hat{\mathcal{T}}$ in a single forward process. For each frame, the plan queries Q^S are initialized. The trajectory is decoded via:

$$\hat{\mathcal{T}} = \mathcal{P}_h\left(\mathcal{P}_D\left(Q^S, F^{\mathrm{w}}\right)\right). \tag{3}$$

Unlike the teacher policy, the student policy does not rely on multi-modal trajectory generation or world model interaction during inference. Instead, it directly extracts planning-relevant cues from world representation, enabling low-latency trajectory prediction.

3.2. World Model

Preliminaries. Autonomous driving world models $\mathcal{W}$ are generative models that predict future driving environment states s conditioned on observational data o and action condition c (*e.g.*, historical trajectories, navigation commands, and ego-vehicle states):

$$\mathcal{W}\left(\{o_{t-h}, ..., o_t\}, \{c_{t-h}, ..., c_t\}\right) = s_{t+1}, \tag{4}$$

where h denotes the number of historical observations, t is the current time. Subsequently, s can be decoded into the corresponding world representation through the decoder of the corresponding mode, such as images, occupancy, *etc*.

Next, we describe a general world model structure, which consists of three main components: observation encoder, feature aggregation, and world decoder.

(1) Observation Encoder $\mathcal{W}_E$. This component processes historical observations $\{o_{t-h}, ..., o_t\}$ to extract features, which are then transformed into a world embedding F_t^{w}:

$$F_t^{\mathrm{w}} = \mathcal{W}_E\left(\{o_{t-h}, ..., o_t\}\right). \tag{5}$$

(2) Feature Aggregation $\mathcal{A}_M$. The feature aggregation module aggregates historical world embeddings $F_{t-h:t}^{\mathrm{w}}$ to capture temporal context and ensure consistency across the historical sequence. Then, these features yield enhanced world representations $\tilde{F}_t^{\mathrm{w}}$ for future prediction:

$$\tilde{F}_t^{\mathrm{w}} = \mathcal{A}_M\left(F_{t-h:t}^{\mathrm{w}}\right). \tag{6}$$

(3) World Decoder $\mathcal{W}_D$. The world decoder $\mathcal{W}_D$ is an autoregressive method that predicts the future world embedding F_{t+1}^{w} based on action conditions c (e.g., driving commands and historical trajectories) and the enhanced world features $\tilde{F}_t^{\mathrm{w}}$ from $\mathcal{A}_M$. This process is formulated as:

$$F_{t+1}^{\mathrm{w}} = \mathcal{W}_D\left(\tilde{F}_t^{\mathrm{w}}, c_t\right). \tag{7}$$

World Model Enhanced Planning. To exploit the model's ability to anticipate future scene evolution, the multi-modal AD policy sets the world embedding F^{w} of its planning decoder $\mathcal{P}_D$ as the predicted world state F_{t+1}^{w} from the world decoder $\mathcal{W}_D$. This autoregressive approach allows the planner to generate future trajectories that are aligned with the predicted world state, as depicted in Fig. 2. The planning formulation in Eqs. (1) and (2) is then updated as:

$$\mathcal{T}_{t+1} = \mathcal{P}_h\left(\mathcal{P}_D\left(Q^T, F_{t+1}^{\mathrm{w}}\right)\right). \tag{8}$$

This approach leverages the world model's predictive capability to refine the AD policy's planning, providing more accurate trajectory predictions while accounting for future environmental dynamics. However, while this method enhances the AD policy's capabilities, the predicted trajectories still primarily mimic expert behavior without fully accounting for the changing dynamics of the future world. To address this limitation, we develop a trainable reward model that transfers world knowledge into the teacher policy by aligning candidate trajectories with the future dynamics predicted by the world model. This allows for a more refined evaluation of the trajectories, enabling the policy to better adapt to the evolving environment, which we will discuss in the next section.

3.3. Reward Model

In this section, we introduce a world model reward-based distillation mechanism that transfers the predictive capability of the world model into the AD policy. We design two complementary forms of reward supervision. The first is an imitation reward, where the reward model evaluates which trajectory aligns best with human driving preferences under future world evolution. The second is a simulation reward, where the reward model assigns scores based on predicted world states and driving quality metrics, such as PDM scores in NAVISIM [5].

The reward model serves as the key mechanism for transferring the future predictive knowledge of the world model into the AD policy. As shown in Fig. 3, we evaluate each candidate trajectory τ_i by combining it with the predicted future world state F_{t+1}^{w}:

$$F_{w,i} = \text{RewardModel}\left(F_{t+1}^{\mathrm{w}}, \tau_i\right), \tag{9}$$

where $F_{w,i}$ denotes the joint trajectory–world interaction representation. Two types of reward heads are then applied

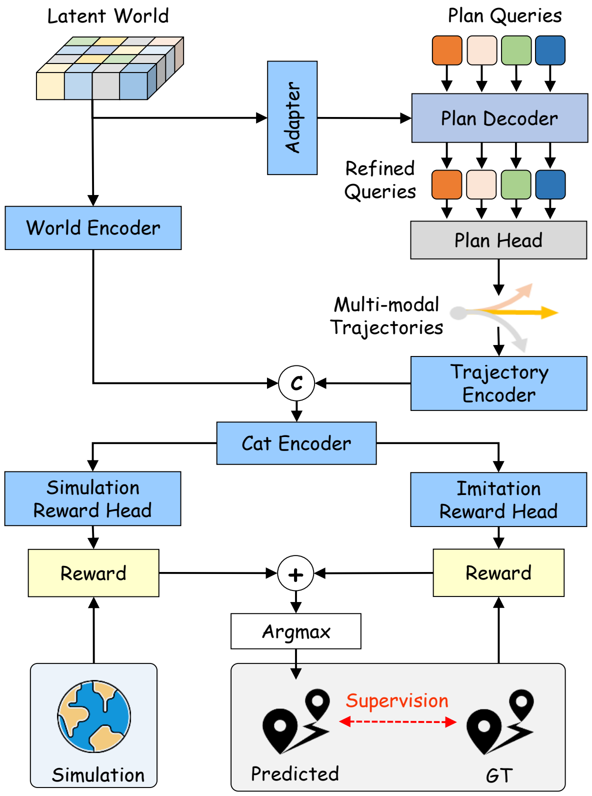

Figure 3. **Overview of reward model.** The reward model consists of multiple components: the world encoder processes the latent world representation, while the plan queries are refined through the plan decoder and plan head to generate multi-modal candidate trajectories. These trajectories are then passed to the trajectory encoder, which encodes them for evaluation by two distinct reward heads: the simulation reward head and the imitation reward head. The final reward is computed by combining these reward values, with the best trajectory selected via the argmax operation. The supervisory signals of the reward model come from simulation and imitation. For the detailed process, please refer to Sec. 3.3.

to obtain trajectory rewards, and the final best trajectory is:

$$\tau^* = \arg\max_i \left(w_1 r_{\text{im},i} + w_2 r_{\text{sim},i} \right), \tag{10}$$

where r_{im} and r_{sim} represent the imitation and simulation reward respectively, w_1 and w_2 are balancing coefficients.

Imitation Reward. The imitation reward aims to assess how well each candidate trajectory aligns with expert driving behavior. For each candidate trajectory τ_i, we first compute its L2 distance d_i from the corresponding expert trajectory, and then normalize it via a softmax function to obtain the target imitation score:

$$r^*_{\text{im},i} = \text{softmax}\left(\frac{-d_i}{\sum_{j=1}^{N} -d_j} \right). \tag{11}$$

The predicted imitation reward r_{im} from the reward model is supervised by minimizing the cross-entropy loss with respect to the softmax-normalized target:

$$\mathcal{L}_{\text{im}} = \text{CrossEntropy}\left(r_{\text{im}}, r^*_{\text{im}} \right). \tag{12}$$

This formulation encourages the reward model to assign higher scores to trajectories that closely match human driving patterns, thereby transferring human-like preferences to the AD policy.

Simulation Reward. Unlike the imitation reward, which captures human driving preferences, the simulation reward evaluates candidate trajectories from an environment-centered perspective, emphasizing safety, comfort, and driving efficiency. Inspired by the NAVISIM [5] score, we construct five metrics in the predicted future world: no collisions (NC), drivable area compliance (DAC), time-to-collision (TTC), ego progress (EP), and comfort (Comf). For implementation details, refer to appendix. The final simulation reward is expressed as:

$$r^*_{\text{sim}} = \{ r_{\text{NC}}, r_{\text{DAC}}, r_{\text{TTC}}, r_{\text{EP}}, r_{\text{Comf}} \}.$$

The predicted simulation reward r_{sim} is supervised using binary cross entropy:

$$\mathcal{L}_{\text{sim}} = \text{BCE}\left(r_{\text{sim}}, r^*_{\text{sim}} \right). \tag{13}$$

This supervision encourages the model to align its trajectory evaluation with safety and physical constraints derived from the predicted world.

Final Reward. To balance human-like driving preference and environment-aware evaluation, the final reward integrates both imitation and simulation rewards as:

$$\begin{aligned} r_{\text{final}} &= \alpha_1 \log r_{\text{im}} + \alpha_2 \log r_{\text{NC}} + \alpha_3 \log r_{\text{DAC}} \\ &\quad + \alpha_4 \log\left(5\, r_{\text{TTC}} + 5\, r_{\text{EP}} + 2\, r_{\text{Comf}} \right), \end{aligned} \tag{14}$$

where $\alpha_1, \ldots, \alpha_4$ are balancing coefficients.

This fusion strategy jointly accounts for human preference, environmental safety, and motion smoothness, enabling the policy to generate trajectories that are both human-like and physically feasible.

3.4. World Knowledge Distillation

To enable real-time inference without relying on the world model, we transfer the teacher policy (T) reasoning ability into a lightweight student policy (S) through two distillation strategies as shown in Fig. 2: Policy Distillation and World Reward Distillation. Both distillation strategies are applied only during training. After training, the teacher policy and the world model are removed, leaving a compact, real-time AD policy.

Policy Distillation. The teacher policy produces a set of planning queries Q^T that encode future-aware reasoning guided by the world model. The student policy generates its own set of queries Q^S. To align the student's planning

Table 1. **End-to-end planning performance on nuScenes validation set.**

Method	Input	Auxiliary Supervision	L2 (m) ↓				Collision (%) ↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
IL [35]	LiDAR	None	0.44	1.15	2.47	1.35	0.08	0.27	1.95	0.77
NMP [55]	LiDAR	Box & Motion	0.53	1.25	2.67	1.48	0.04	0.12	0.87	0.34
FF [17]	LiDAR	Freespace	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO [25]	LiDAR	Freespace	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
BevFormer [32]+OccWorld [61]	Camera	3D-Occ	0.43	0.87	1.31	0.87	-	-	-	-
BevFormer [32]+Occ-LLM [50]	Camera	3D-occ	0.26	0.67	0.98	0.64	-	-	-	-
ST-P3 [18]	Camera	Map & Box & Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [19]	Camera	Map & Box & Motion & Tracklets & Occ	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD-Base [24]	Camera	Map & Box & Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
UAD [12]	Camera	Box	0.39	0.81	1.50	0.90	0.01	0.12	0.43	0.19
PARA-Drive [46]	Camera	Map & Box & Motion & Tracklets & Occ	0.40	0.77	1.31	0.83	0.07	0.25	0.60	0.30
OccNet [41]	Camera	3D-Occ & Map & Box	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
GenAD [62]	Camera	Map & Box & Motion	0.36	0.83	1.55	0.91	0.06	0.23	1.00	0.43
Epona [28]	Camera	None	0.61	1.17	1.98	1.25	0.01	0.22	0.85	0.36
SSR [28]	Camera	None	**0.24**	0.65	1.36	0.75	**0.00**	0.10	0.36	0.15
OccWorld [61]	Camera	None	0.43	1.08	1.99	1.17	0.07	0.38	1.35	0.60
RenderWorld [51]	Camera	None	0.48	1.30	2.67	1.48	0.14	0.55	2.23	0.97
GaussianAD [63]	Camera	3D-Occ & Map & Box	0.40	0.66	0.92	0.66	0.49	0.38	0.61	0.49
GaussianAD [63]	Camera	3D-Occ & Map & Box & Motion	0.40	0.64	**0.88**	0.64	0.09	0.38	0.81	0.42
Drive-OccWorld [54]	Camera	4D-Occ	0.32	0.75	1.49	0.85	0.05	0.17	0.64	0.29
Baseline	Camera	3D-Occ	0.29	0.79	1.56	0.88	0.70	0.76	1.71	1.06
WPT-Student (Ours)	Camera	4D-Occ	**0.24**	0.58	1.17	0.66	0.14	0.16	0.42	0.24
WPT-Teacher (Ours)	Camera	4D-Occ	0.25	**0.58**	1.01	**0.61**	0.16	**0.08**	**0.10**	**0.11**

intent with the teacher's world-conditioned reasoning, we minimize the L2 distance between the two query sets:

$$\mathcal{L}_{\text{policy}} = \|Q^S - Q^T\|_2. \tag{15}$$

This alignment transfers the structured planning intent from the teacher into the student, allowing the student to inherit future-aware reasoning without autoregressive rollout.

World Reward Distillation. The teacher generates multi-modal candidate trajectories $\mathcal{T}^T = \{\tau_1^T, \ldots, \tau_N^T\}$, which are evaluated by the reward model to obtain the optimal trajectory τ_T^* based on the final reward score (see Sec. 3.3). Meanwhile, the student generates a single trajectory τ_S, which is also evaluated by the reward model. To transfer the world-model-based evaluation knowledge, we minimize the difference between the reward scores of the student's trajectory and the teacher's best trajectory:

$$\mathcal{L}_{\text{reward}} = \|r_{\text{final}}(\tau_S) - r_{\text{final}}(\tau_T^*)\|_2. \tag{16}$$

This formulation allows the student to implicitly learn the teacher's world-informed decision criteria, effectively bridging the gap between explicit world-model reasoning and lightweight policy inference.

4. Experiments

4.1. Datasets and Metrics

NuScenes (open-loop). We evaluate open-loop planning on nuScenes [2], which contains 1,000 driving scenes, each 20s long, collected with a full sensor suite providing 360° coverage. The dataset includes ∼1.4M images and 3D bounding boxes for 23 classes annotated at 2Hz keyframes; semantic maps are available. We follow the standard split of 700/150/150 scenes for train/val/test. Following existing practice on nuScenes [19, 24], we report L2 displacement error and collision rate for planning quality.

Bench2Drive (closed-loop). We use Bench2Drive [22], a CARLA-based [7] benchmark designed for multi-ability closed-loop E2E AD assessment, covering diverse driving scenarios. Following the official protocol, we train on 1000 clips (950 for training and 50 for open-loop validation) and compare closed-loop results on the predefined 220 routes.

4.2. Implementation Details

Training on NuScenes Setting. We adopt the pre-trained Drive-OccWorld [54] model as our world model and freeze its weights during training. WPT is trained for 12 epochs with a batch size of 8, using AdamW and a cosine annealing learning rate schedule.

Training on Bench2Drive Setting. We adopt an instance-based world model that forecasts future agent states and lane topology instead of occupancy (detailed in our appendix). WPT is trained for 6 epochs with a batch size of 16, also using AdamW and a cosine annealing learning rate schedule.

Table 2. **Open-loop and closed-loop planning performance on Bench2Drive.** Avg. L2 is averaged over the predictions in 2 seconds under 2Hz. * denotes expert feature distillation.

Method	Avg. L2 (m) ↓	Driving Score ↑	Success Rate (%) ↑	Efficiency ↑	Comfortness ↑
AD-MLP [56]	3.64	18.05	0.00	48.45	22.63
UniAD-Base [19]	0.73	45.81	16.36	129.21	43.58
UniAD-Tiny [19]	0.80	40.73	13.18	123.92	47.04
VAD-Base [24]	0.91	42.35	15.00	157.94	46.01
VAD-Tiny [24]	1.15	34.28	10.45	70.04	**66.86**
SparseDrive [40]	0.87	44.54	16.71	170.21	48.63
GenAD [62]	-	44.81	15.90	-	-
DiFSD [39]	0.70	52.02	21.00	178.30	-
DriveTransformer [23]	**0.62**	63.46	35.01	100.64	20.78
DiffAD [43]	-	67.92	38.64	-	-
WoTE [30]	-	61.71	31.36	-	-
DriveDPO [37]	-	62.02	30.62	166.80	26.79
BridgeAD [57]	0.71	50.06	22.73	-	-
Baseline	0.79	65.23	34.10	184.86	23.44
WPT-Student (Ours)	0.75	72.61	45.45	188.52	17.80
WPT-Teacher (Ours)	0.76	**79.23**	**54.54**	**188.63**	16.39
TCP-traj* [48]	1.70	59.90	30.00	76.54	18.08
ThinkTwice* [21]	0.95	62.44	31.23	69.33	16.22
DriveAdapter* [20]	1.01	64.22	33.08	70.22	16.01

4.3. Main Results

Open-Loop Results on NuScenes. We compare WPT with recent end-to-end planners on nuScenes, shown in Tab. 1. As a reference, our Baseline uses the same student structure without the proposed reward or distillation. Against this baseline, WPT-Teacher improves Avg. L2 to 0.61m and collision to 0.11%, while WPT-Student retains most of the gains without test-time world-model overhead.Compared with strong world-model methods, WPT-Teacher also achieves the best Avg. L2 and the lowest collision (*e.g.*, vs. Drive-OccWorld [54]: 0.85m, 0.29%). By horizon, WPT-Teacher leads at 2s (0.58m) and markedly reduces 3s collisions to 0.10%, indicating stronger long-horizon foresight, benefiting from our world-aware training. Importantly, our lightweight WPT-Student preserves these performance improvements while achieving world-model-free inference. These results validate that training-time interaction and world-aware distillation transfer predictive awareness into a compact policy that effectively improves the student performance.

Closed-Loop Results on Bench2Drive. We evaluate WPT on Bench2Drive [22] under both open-loop (Avg. L2) and closed-loop metrics, as summarized in Tab. 2. Additional multi-modality analyses across diverse scenarios are provided in our appendix. WPT-Teacher achieves the best Driving Score (79.23), Success Rate (54.54%), and Efficiency (188.63), surpassing recent E2E planners such as DriveTransformer [23] (63.46 DS / 35.01% SR / 100.64 Eff.) and DriveDPO [37] (62.02 DS / 30.62% SR / 166.80 Eff.). Relative to the Baseline (same student architecture, no rewards/distillation), WPT-Teacher boosts DS by +14.00 and SR by +20.44 points and the distilled WPT-Student retains most of the gains. Despite the lower open-loop Avg. L2 (0.62m) of DriveTransformer, both WPT variants yield stronger closed-loop results, consistent with Bench2Drive's emphasis on closed-loop ability. Compared with methods that rely on expert feature distillation, our models deliver higher DS and SR without requiring expert knowledge. We note a known efficiency–comfort trade-off; improving smoothness without compromising efficiency and success is left for future reward-shaping refinements. Overall, the gains indicate that our interaction-based rewards and world-aware distillation yield a policy that is effective in closed-loop rollouts.

4.4. Ablation Study

To validate each design in our method, we conduct comprehensive studies on nuScenes dataset using average L2 error and collision rate as planning metrics. Baseline-T denotes the teacher planner WPT trained without our reward model.

Effect of Reward Model. We evaluate when interaction-based rewards are applied, during training and at inference, as shown in Tab. 3. Starting from Baseline-T (0.72m / 0.71%), adding the imitation reward in training reduces collisions to 0.23% (0.70m L2). Applying the same reward at inference leads to better results (0.69m / 0.22%). Adding simulation rewards gives the largest gains: training-only reaches 0.62m / 0.14%, and enabling inference-time scoring attains the best 0.61m / 0.11%. The results show the effectiveness of our reward model. Meanwhile, we find that most benefits come from training-time supervision; optional inference-time scoring adds a small margin at the cost of runtime coupling (see our Appx.).

Table 3. **Ablation study of reward model.** We compare different reward equipment at different usage stages (training stage or also at inference). "Im. Rwd." is an imitation reward, while "Sim.Rwd." means simulation reward.

Method	Stage		Planning	
	Train	Infer.	Avg. L2(m)↓	Avg. Col.(%)↓
Baseline-T	-	-	0.72	0.71
+ Im. Rwd.	✓	-	0.78	0.23
	✓	✓	0.69	0.22
+ Im.&Sim. Rwd.	✓	-	0.62	0.14
	✓	✓	**0.61**	**0.11**

Table 4. **Ablation study of different rewards.** Simulation reward consists of five signals: NC (No Collision), EP (Ego Progress), DAC (Drivable Area Compliance), TTC (Time-to-Collision), and Comf. (Comfort).

Im. Rwd.	Sim. Rwd.					Planning	
	NC	DAC	EP	TTC	Comf.	L2 (m)↓	Col. (%)↓
-	-	-	-	-	-	0.72	0.71
✓	-	-	-	-	-	0.76	0.23
✓	✓	✓	✓	✓	✓	**0.61**	**0.11**
✓	-	✓	✓	✓	✓	0.62	0.22
✓	✓	-	✓	✓	✓	0.62	0.23
✓	✓	✓	-	✓	✓	0.62	0.23
✓	✓	✓	✓	-	✓	0.69	0.25
✓	✓	✓	✓	✓	-	0.63	0.23

Effect of Different Rewards. We analyze the composition of our reward signals, as shown in Tab. 4. With imitation reward only, performance is 0.76m / 0.23%. Further aggregating all simulation rewards (NC, DAC, EP, TTC, Comf.) achieves the best 0.61m / 0.11%, indicating complementary effects. Ablating each simulation reward degrades performance; removing the TTC reward signal causes the largest collision increase (to 0.25%), highlighting its importance for safety, while other simulation reward signals contribute smaller but steady gains.

Effect of Interaction Sources. We compare using ground-truth occupancy (GT-Occ) versus world-model rollouts (WM-Occ) to drive reward signals, shown in Tab. 5. With imitation reward only, GT-Occ offers stronger signals (0.64m / 0.16% vs. 0.69m / 0.22%). With equipping simulation rewards, WM-Occ achieves the best overall (0.61m / 0.11%) compared to GT-Occ (0.65m / 0.11%), suggesting that training on world-model rollouts with WM-Occ aligns the policy with the predictive structure compared to the deterministic GT-Occ.

Effect of Distillation Strategies. We study the lightweight student with different distillation strategies, shown in Tab. 6. Plain training yields 0.88m / 1.06%. Query-level distillation alone provides a sizable step (0.69m / 0.86%). Adding imitation-reward distillation sharply reduces collisions to 0.25% (0.68m). Combining imitation and simulation reward distillation attains the best student performance, 0.66m / 0.24%, demonstrating that world-aware signals effectively transfer predictive awareness without test-time world-model calls.

Table 5. **Interaction occupancy source ablation.** GT-Occ denotes using ground truth occupancy for interaction, while WM-Occ denotes using the occupancy generated by WM.

Method	Interaction Source		Planning	
	GT-Occ	WM-Occ	L2 (m)↓	Col. (%)↓
Baseline-T	-	✓	0.72	0.71
+ Im. Rwd.	✓	-	0.64	0.16
	-	✓	0.69	0.22
+ Im.&Sim. Rwd.	✓	-	0.65	0.11
	-	✓	**0.61**	**0.11**

Table 6. **Ablation study of distillation.** "Query" denotes the distillation of the plan query.

Distillation Type			Planning	
Query	Im. Rwd.	Sim. Rwd.	Avg. L2 (m) ↓	Avg. Col. (%) ↓
-	-	-	0.88	1.06
✓	-	-	0.69	0.86
✓	✓	-	0.68	0.25
✓	✓	✓	**0.66**	**0.24**

Computation and Latency. We report full training compute (GPU hours) and planning inference latency in the appendix. Briefly, the distilled WPT-Student matches Baseline latency (64 ms) and is $4.9\times$ faster than WPT-Teacher (312 ms), while improving planning from 0.88m / 1.06% to 0.66m / 0.24%.

Overall, these ablations show that: 1) interaction-based rewards are the primary driver of safety gains; 2) simulator reward signals provide complementary improvements; 3) supervision from world-model rollouts is preferable to deterministic GT; 4) world-aware distillation consolidates these effects into a compact, deployable planner.

5. Conclusion

In this paper, we present **WPT**, a World-to-Policy Transfer training paradigm for end-to-end AD. Through a trainable reward model and dual distillation schemes, WPT distills world-model knowledge into a lightweight policy during training. This eliminates runtime world-model dependencies, ensuring real-time inference. Extensive experiments on both open-loop and closed-loop benchmarks demonstrate that WPT achieves state-of-the-art performance in planning accuracy, driving safety, and efficiency.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Contract 62471450.

References

[1] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 2

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 6

[3] Dian Chen, Vladlen Koltun, and Philipp Krähenbühl. Learning to drive from a world on rails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15590–15599, 2021. 2

[4] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024. 2

[5] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024. 4, 5

[6] Anna Dawid and Yann LeCun. Introduction to latent variable energy-based models: a path toward autonomous machine intelligence. *Journal of Statistical Mechanics: Theory and Experiment*, 2024, 2023. 1

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2, 6

[8] Hao Gao, Shaoyu Chen, Bo Jiang, Bencheng Liao, Yiang Shi, Xiaoyang Guo, Yuechuan Pu, Haoran Yin, Xiangyu Li, Xinbang Zhang, et al. Rad: Training an end-to-end driving policy via large-scale 3dgs-based reinforcement learning. *arXiv preprint arXiv:2502.13144*, 2025. 2

[9] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024. 2

[10] Xunjiang Gu, Guanyu Song, Igor Gilitschenski, Marco Pavone, and Boris Ivanovic. Producing and leveraging online map uncertainty in trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14521–14530, 2024. 2

[11] Ke Guo, Haochen Liu, Xiaojun Wu, Jia Pan, and Chen Lv. ipad: Iterative proposal-centric end-to-end autonomous driving. *arXiv preprint arXiv:2505.15111*, 2025. 2

[12] Mingzhe Guo, Zhipeng Zhang, Yuan He, Ke Wang, and Liping Jing. End-to-end autonomous driving without costly modularization and 3d manual annotation. *arXiv preprint arXiv:2406.17680*, 2024. 6

[13] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018. 1

[14] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[15] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[16] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 1

[17] Peiyun Hu, Aaron Huang, John Dolan, David Held, and Deva Ramanan. Safe local motion planning with self-supervised freespace forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12732–12741, 2021. 6

[18] Shengchao Hu, Li Chen, Peng Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 2022. 1, 6

[19] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023. 1, 2, 4, 6, 7

[20] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7953–7963, 2023. 7

[21] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. 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*, pages 21983–21994, 2023. 7

[22] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024. 6, 7

[23] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. *arXiv preprint arXiv:2503.07656*, 2025. 2, 7

[24] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 8340–8350, 2023. 1, 2, 6, 7

[25] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *European Conference on Computer Vision*, pages 353–369. Springer, 2022. 6

[26] Derun Li, Changye Li, Yue Wang, Jianwei Ren, Xin Wen, Pengxiang Li, Leimeng Xu, Kun Zhan, Peng Jia, Xianpeng Lang, et al. Learning personalized driving styles via reinforcement learning from human feedback. *arXiv preprint arXiv:2503.10434*, 2025. 2

[27] Kailin Li, Zhenxin Li, Shiyi Lan, Yuan Xie, Zhizhong Zhang, Jiayi Liu, Zuxuan Wu, Zhiding Yu, and Jose M Alvarez. Hydra-mdp++: Advancing end-to-end driving via expert-guided hydra-distillation. *arXiv preprint arXiv:2503.12820*, 2025. 2

[28] Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. *arXiv preprint arXiv:2409.18341*, 2024. 6

[29] Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint arXiv:2406.08481*, 2024. 3

[30] Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27137–27146, 2025. 1, 3, 7

[31] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024. 2

[32] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024. 6

[33] Hao Lu, Zhuang Ma, Guangfeng Jiang, Wenhang Ge, Bohan Li, Yuzhan Cai, Wenzhao Zheng, Yunpeng Zhang, and Yingcong Chen. 4d driving scene generation with stereo forcing. *arXiv preprint arXiv:2509.20251*, 2025. 2

[34] Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin Sapp, Brandyn White, Aleksandra Faust, Shimon Whiteson, et al. Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7553–7560. IEEE, 2023. 2

[35] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736, 2006. 6

[36] Jürgen Schmidhuber. On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv preprint arXiv:1511.09249*, 2015. 1

[37] Shuyao Shang, Yuntao Chen, Yuqi Wang, Yingyan Li, and Zhaoxiang Zhang. Drivedpo: Policy learning via safety dpo for end-to-end autonomous driving. *arXiv preprint arXiv:2509.17940*, 2025. 1, 2, 7

[38] Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei Yang, and Yadan Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22432–22441, 2025. 2

[39] Haisheng Su, Wei Wu, and Junchi Yan. 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. 7

[40] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8795–8801. IEEE, 2025. 2, 7

[41] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8406–8415, 2023. 6

[42] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7153–7162, 2020. 2

[43] Tao Wang, Cong Zhang, Xingguang Qu, Kun Li, Weiwei Liu, and Chang Huang. Diffad: A unified diffusion modeling approach for autonomous driving. *arXiv preprint arXiv:2503.12170*, 2025. 7

[44] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024. 2

[45] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024. 1, 2

[46] Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15449–15458, 2024. 6

[47] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and P. Abbeel. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning*, 2022. 1

[48] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for

end-to-end autonomous driving: A simple yet strong baseline. *Advances in Neural Information Processing Systems*, 35:6119–6132, 2022. 7

[49] Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1602–1611, 2025. 2

[50] Tianshuo Xu, Hao Lu, Xu Yan, Yingjie Cai, Bingbing Liu, and Yingcong Chen. Occ-llm: Enhancing autonomous driving with occupancy-based large language models. *arXiv preprint arXiv:2502.06419*, 2025. 6

[51] Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Haiyang Liu, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al. Renderworld: World model with self-supervised 3d label. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6063–6070. IEEE, 2025. 3, 6

[52] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14662–14672, 2024. 2

[53] Pengxuan Yang, Yupeng Zheng, Qichao Zhang, Kefei Zhu, Zebin Xing, Qiao Lin, Yun-Fu Liu, Zhiguo Su, and Dongbin Zhao. Uncad: Towards safe end-to-end autonomous driving via online map uncertainty. *arXiv preprint arXiv:2504.12826*, 2025. 2

[54] Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9327–9335, 2025. 1, 3, 6, 7

[55] Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8660–8669, 2019. 6

[56] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. 7

[57] Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6854–6863, 2025. 2, 7

[58] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024. 2

[59] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27220–27230, 2025. 1, 3

[60] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15222–15232, 2021. 2

[61] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023. 2, 3, 6

[62] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pages 87–104. Springer, 2024. 6, 7

[63] Wenzhao Zheng, Junjie Wu, Yao Zheng, Sicheng Zuo, Zixun Xie, Longchao Yang, Yong Pan, Zhihui Hao, Peng Jia, Xianpeng Lang, et al. Gaussianad: Gaussian-centric end-to-end autonomous driving. *arXiv preprint arXiv:2412.10371*, 2024. 6

[64] Yinan Zheng, Ruiming Liang, Kexin Zheng, Jinliang Zheng, Liyuan Mao, Jianxiong Li, Weihao Gu, Rui Ai, Shengbo Eben Li, Xianyuan Zhan, et al. Diffusion-based planning for autonomous driving with flexible guidance. *arXiv preprint arXiv:2501.15564*, 2025. 2

[65] Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28632–28642, 2025. 1