Imagine-2-Drive: Leveraging High-Fidelity World Models via Multi-Modal Diffusion Policies

Anant Garg[1], **K. Madhava Krishna**[1]

{garg.anant205@gmail.com, mkrishna@iiit.ac.in}
[1]Robotics Research Center, IIIT Hyderabad

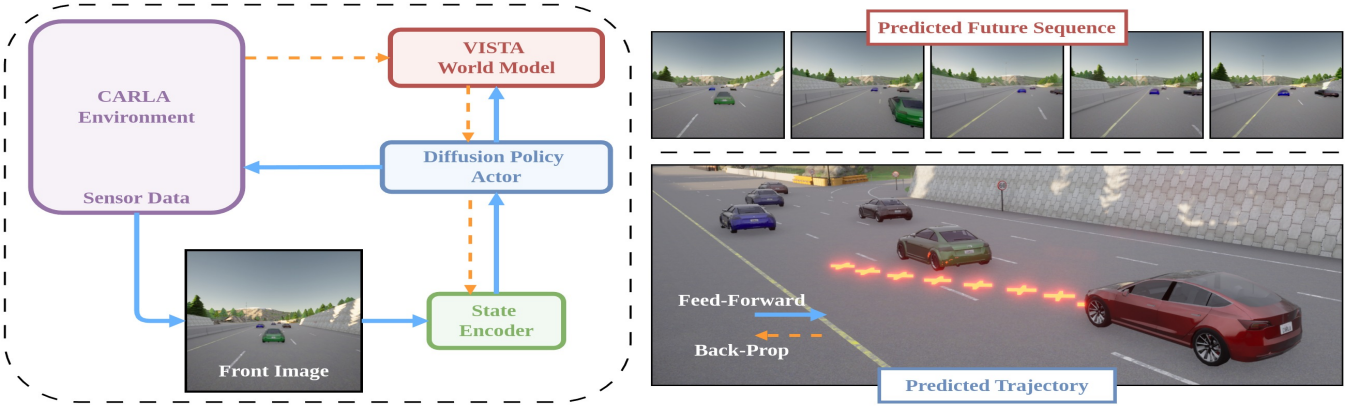

Figure 1: **Overview**: Using front camera RGB image as the sole input modality, *Imagine-2-Drive* provides a framework to combine *VISTAPlan*, VISTA based World Model with *DPA*, a multi-modal diffusion based policy actor. Given a trajectory output by *DPA*, shown in *( )* and the corresponding predicted future observations from *VISTAPlan*, the *DDPO* tries to find an optimal policy by maximizing the cumulative sum of rewards from future states. The proposed architecture is shown along with the gradient flow for joint end-to-end training.

Abstract

In autonomous driving with image based state space, accurate prediction of future events and modeling diverse behavioral modes are essential for safety and effective decision-making. World model-based Reinforcement Learning (WMRL) approaches offers a promising solution by simulating future states from current state and actions. However, utility of world models is often limited by typical RL policies being limited to deterministic or single gaussian distribution. By failing to capture the full spectrum of possible actions, reduces their adaptability in complex, dynamic environments. In this work, we introduce *Imagine-2-Drive*, a framework that consists of two components, VISTAPlan, a high-fidelity world model for accurate future prediction and Diffusion Policy Actor (DPA), a diffusion based policy to model multi-modal behaviors for trajectory prediction. We use VISTAPlan to simulate and evaluate trajectories from DPA and use Denoising Diffusion Policy Optimization (DDPO) to train DPA to maximize the cumulative sum of rewards over the trajectories. We analyze the benefits of each component and the framework as a whole in CARLA with standard driving metrics. As a consequence of our twin novelties- VISTAPlan and DPA, we significantly outperform the state of the art (SOTA) world models on standard driving metrics by 15% and 20% on Route Completion and Success Rate respectively.

Project website: https://anantagrg.github.io/Imagine-2-Drive.github.io/

INTRODUCTION

Autonomous driving systems must operate safely and effectively in complex, dynamic environments, where accurate prediction of future events and diverse behavioral modeling are critical for informed decision-making. The ability to foresee potential obstacles, navigate uncertain traffic conditions, and make proactive adjustments to driving strategies relies on robust future prediction capabilities.

World model-based Reinforcement Learning approaches (Pan et al. 2022; Li et al. 2024) have emerged as a promising solution to this challenge by simulating future states based on current observations and actions. These models enable autonomous vehicles (AVs) to internally "imagine" possible future scenarios, facilitating more efficient exploration and reducing the risks and costs associated with real-world interactions. However, the utility of these world models is often limited by the nature of traditional RL policies. Most RL policies are constrained to deterministic outputs or single gaussian distributions, which fail to capture the full range of possible behaviors. This undermines the adaptability of the world models and their ability to handle the complexity

and variability found in driving environments. The significant advantages of world models also highlights the importance of learning an accurate world model.

Current WMRL (Hafner et al. 2020, 2022, 2024; Pan et al. 2022; Deng, Jang, and Ahn 2021) approaches, model the environment dynamics in latent space using a Recurrent Neural Network (RNN) based network. A common limitation of these approaches is their reliance on single-step transition models, where errors accumulate over multi-step planning, causing planned states to drift from the on-policy distribution. Prior works (Wang et al. 2023; Ding et al. 2024; Gao et al. 2024b) leverage video diffusion (Blattmann et al. 2023) based approaches to predict the future states in a single pass, with VISTA (Gao et al. 2024b) being the most versatile and accurate.

To overcome these limitations, we present ***Imagine-2-Drive*** , a novel framework that incorporates two key innovations: VISTAPlan, a high-fidelity world model designed for precise future prediction, and DPA, a diffusion-based policy actor that models diverse behavioral modes for trajectory planning. VISTAPlan extends upon the VISTA's prediction capabilities by incorporating additional modules to predict reward and discount factors. These enhancements enable VISTAPlan to function as a comprehensive world model, facilitating more effective planning and decision-making.

Similar to (Janner et al. 2022; Saha et al. 2023; Ze et al. 2024; Jiang et al. 2023; Yang et al. 2024), DPA utilizes a diffusion based model to predict the trajectory (a sequence of actions) in a single pass. Being a generative model, it allows DPA to model multiple behavioral modes. By incorporating the diverse behavior patterns inherent to diffusion policies, our framework can explore a broader range of behaviors, thereby improving its overall performance and robustness. DPA is trained with DDPO (Black et al. 2024) using VISTAPlan to simulate and evaluate trajectories to maximize the cumulative reward over the trajectories

We validate our framework in the autonomous driving domain in CARLA (Dosovitskiy et al. 2017) simulator, demonstrating its superiority over existing methods. A comprehensive evaluation across diverse driving scenarios highlights the contributions of each component, and how they effectively complement one another to enhance the overall performance.

To summarize, our key contributions are:

1. We propose VISTAPlan, a high fidelity world model based reinforcement learning framework for autonomous driving, with image as the sole input modality. VISTAPlan extends VISTA by incorporating additional reward and discount factor heads enabling it for effective planning and decision-making.
2. DPA, a novel diffusion based policy actor which can model multiple behavioral modes, thereby enabling it to explore a broader range of behaviors. It is trained using DDPO to maximize the cumulative sum of rewards over trajectories.
3. Thorough analysis of our approach and baselines in complex CARLA driving scenarios over standard driving metrics to understand the effectiveness of our framework and validating each component.

RELATED WORK

World Models for Autonomous Driving

For autonomous driving, world models follow two key paradigms: one leverages world models as neural driving simulators, while the other unifies action prediction with future generation. Think-2-Drive (Li et al. 2024) adapts Dreamer-V3 (Hafner et al. 2024) for autonomous driving using BEV state-space representations.

Generative world models have been extensively explored for future driving sequence prediction based on various input modalities. Models like DriveGAN (Kim et al. 2021), DriveDreamer (Wang et al. 2023), and MagicDrive (Gao et al. 2024a) focus on generating future driving scenarios using actions as inputs. GAIA-I (Hu et al. 2023) expands this approach by incorporating text commands alongside actions. VISTA (Gao et al. 2024b), the most versatile, accepts inputs including actions, trajectories, text commands, and goal points, demonstrating high generalizability for sequence prediction.

Diffusion Policy for Planning

Following the success of Diffuser (Janner et al. 2022) on D4RL benchmark (Fu et al. 2021), diffusion-based policies (Chi et al. 2024; Reuss et al. 2023; Ankile et al. 2024; Wang et al. 2024; Ze et al. 2024; Sridhar et al. 2023; Pearce et al. 2023) have been successfully applied in robotics for motion planning. Most typically, these policies are trained from human demonstrations through a supervised objective, and enjoy both high training stability and strong performance in modeling complex and multi-modal trajectory distributions. In the autonomous driving domain, successful works like (Yang et al. 2024; Jiang et al. 2023) sample trajectories using diffusion models while (Liu et al. 2024) formulate Constrained MDP (CMDP) to incorporate constraints.

Training Diffusion Models with Reinforcement Learning

Training diffusion models using reinforcement learning (RL) techniques has gained traction in recent research, particularly for applications such as text-to-image generation (Fan et al. 2023; Black et al. 2024; Clark et al. 2024; Wallace et al. 2023). DDPO (Black et al. 2024) formulate the denoising process as an MDP and apply PPO update to this formalism. We build upon these earlier works and embed the denoising MDP within the environmental MDP of the dynamics of autonomous driving.

Our approach leverages VISTA's high-fidelity predictive capabilities to build a world model (VISTAPlan) combined with a diffusion based policy (DPA) trained using DDPO method which effectively captures multiple behavioral modes.

METHODOLOGY

Imagine-2-Drive is a World Model based RL framework designed for long-horizon trajectory planning using only front-

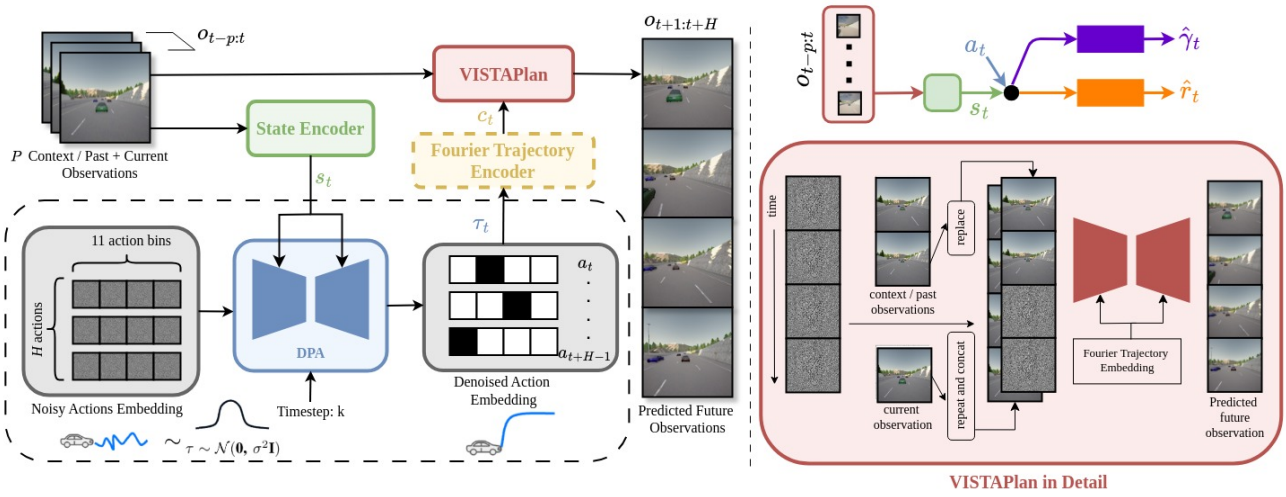

Figure 2: **Architecture**: *Imagine-2-Drive* comprises of a *Diffusion Policy Actor (DPA)* for trajectory prediction τ and *VISTAPlan* as a World Model for future state and reward prediction. Based on the encoded state using current and P past observations, DPA denoises a set of one-hot embeddings for H future discrete actions to get a trajectory τ. The context and current observations, along with the trajectory is passed through *VISTAPlan* to predict future H observations along with the reward for each state and action pair. *DPA* is trained using *DDPO* to maximize the cumulative sum of the rewards over the trajectories.

facing camera image as input. The framework aims to generate a collision-free trajectory over a prediction horizon H, based on a sequence of past and current observations.

As depicted in Fig. 2, the framework consists of three key components:

1. **State Encoder:** Encodes the sequence of P past front-view RGB camera observations, together with the current observation, to provide contextual information for current state.
2. **VISTAPlan:** A high-fidelity world model that facilitates efficient planning by accurately predicting future observations.
3. **DPA:** A diffusion based policy actor to produce a trajectory of actions which maximizes the cumulative sum of rewards over the prediction horizon H.

Notation: This paper differentiates between two categories of timesteps: diffusion timesteps, indicated by superscripts $k \in \{1, \ldots, T\}$, and environment timesteps, denoted by subscripts $t \in \{1, \ldots, N\}$.

Approximate POMDP with MDP

Using a single frame as the state in autonomous driving lacks the necessary temporal context to capture object motion and changes over time. This limitation makes it challenging to formulate the problem as a Markov Decision Process (MDP), which assumes that the current state captures all relevant information for decision-making. To address this, we approximate an MDP by incorporating a sequence of previous observations, thereby adding temporal information to the state representation. This approach effectively transforms the problem from a POMDP to an MDP by including a fixed-length history, which provides a richer context for understanding the dynamics of the driving environment.

Following ViNT (Shah et al. 2023), we use a *State Encoder* which tokenize each observation image $\{o_i\}_{t-P}^{t}$ into an embedding of size $d_{model} = 512$ with an EfficientNet-B0 (Tan and Le 2020) model which outputs a feature vector $\zeta(o_i)$. The individual tokens are then combined with a positional encoding and fed into a transformer backbone $\mathcal{F}_{sa}$. We use a decoder-only transformer with $n_L = 4$ multi-headed attention blocks, each with $n_H = 4$ heads and $d_{FF} = 2048$ hidden units. The output tokens are concatenated and flattened, then passed through MLP layers to give a final state embedding $s_t \in R^{32}$

$$s_t = MLP\left(\mathcal{F}_{sa}\left(\zeta(o)_{t-P:t}\right)\right) \tag{1a}$$

VISTAPlan

We leverage VISTA's high-fidelity prediction capabilities as the foundation for building our world model. Given the P past and current observations $\{o_i\}_{t-P}^{t}$, we predict H future observations $\{o_i\}_{t+1}^{t+H}$ conditioned on a action trajectory $\tau \in R^{H\times 2}$.

Fourier Trajectory Encoding: The trajectory is encoded using the Fast Fourier Transform (FFT) (Mildenhall et al. 2020) which offers a compact frequency-domain representation that captures the trajectory's underlying patterns.

$$\tau_t = (a_t, a_{t+1}, \ldots, a_{t+H}) \tag{2a}$$

$$c_t = FFT\left(\tau_t\right) \tag{2b}$$

Future State Prediction: To predict the future observations $\{o_i\}_{t+1}^{t+H}$ corresponding to τ_t, we use the Stable Video Diffusion model (SVD) (Blattmann et al. 2023) used in

VISTA.

$$o_{t+1:t+H} = SVD(o_{t-P:t}, c_t) \tag{3a}$$

Using the state encoder 1a and $\{o_i\}_{t+1}^{t+H}$, we can get future states $\{s_i\}_t^{t+H}$. This capability allows the model to anticipate the future state of the environment dependent on the trajectory, which is crucial for safe and effective decision making in complex driving scenarios.

Additional Required MDP Components: To use VISTA as a world model, we add additional sub-modules to predict future rewards $\hat{r}_t$ and $\hat{\gamma}_t$ using s_t and a_t

$$\hat{r}_t \sim p_\phi\left(\hat{r}_t \mid s_t, a_t\right), \quad \hat{\gamma}_t \sim p_\phi\left(\hat{\gamma}_t \mid s_t, a_t\right) \tag{4}$$

This provides us with all the components required for MDP and to use VISTAPlan as a world model for planning.

Loss Function: In addition to VISTA's loss function $\mathcal{L}_{VISTA}$ defined in equation 6 of (Gao et al. 2024b), we add the additional loss functions for learning reward to get a final loss function $\mathcal{L}_{WM}$:

$$\mathcal{L}_{WM} = \mathcal{L}_{Vista} - E_{\tau\sim p(\tau|\pi)}\left[\sum_{t=1}^{H} \ln p_\phi\left(\hat{r}_t \mid s_t, a_t\right) + \ln p_\phi\left(\hat{\gamma}_t \mid s_t, a_t\right)\right] \tag{5}$$

Diffusion Policy Actor (DPA)

Given the current state s_t, we use a diffusion based policy network (π_θ) to predict the trajectory τ_t 2a. The diffusion model captures complex, multi-modal action distributions, enabling effective long-horizon planning essential for autonomous driving.

However, unlike imitation learning, where diffusion policies are trained with fixed datasets of expert actions, our RL setting lacks predefined targets. Instead, the policy learns to maximize cumulative rewards through exploration and interaction with the environment.

The reinforcement learning (RL) objective is for the agent to maximize $\mathcal{J}_{RL}(\pi)$, the expected cumulative reward over trajectories sampled from its policy.

Without losing generalizibility, we assume the current environment timestep (t) to be 0.

$$\mathcal{J}_{RL}(\pi) = E_{\tau\sim p_\theta(\tau|\pi)}\left[\sum_{t=1}^{H} \hat{\gamma}_t \hat{r}_\phi(s_t, a_t)\right] \tag{6}$$

Following DDPO (Black et al. 2024), we formulate the policy denoising process as an MDP itself and use policy gradients method to train the DPA.

We begin with a randomly initialized diffusion policy network. Sampling from the policy network begins with a noisy trajectory sample drawn from a isotropic gaussian distribution, $\tau^T \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The reverse process is defined by a learned distribution $p_\theta(\tau^{k-1}|\tau^k, s_0)$, which progressively "denoises" the action sequence to produce a sequence of trajectories $\{\tau^T, \tau^{T-1}, \ldots, \tau^0\}$ ending with the sample τ^0.

After getting the final denoised trajectory τ^0, given current state s_t, we query our world model to get future states $\{s_i\}_{t=1}^{H}$ using 3a and 1a.

Denoising as a multi-step MDP: In our formulation, the denoising process itself is viewed as an MDP where:

$$s^k_{DDPO} = (\tau^k, s_0) \tag{7a}$$

$$a^k_{DDPO} = \tau^{k-1} \tag{7b}$$

$$r^k_{DDPO} = \begin{cases} \sum_{t=1}^{H} \hat{\gamma}_t \hat{r}_\phi(s_t, a_t) & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases} \tag{7c}$$

$$\pi_{DDPO}(a^k_{DDPO} \mid s^k_{DDPO}) = p_\theta(\tau^{k-1}|\tau^k, s_0) \tag{7d}$$

Denoising Diffusion RL Objective: The objective for this MDP $\mathcal{J}_{DDRL}(\theta)$ is to maximize the reward signal r_{DDPO}.

$$\mathcal{J}_{DDRL}(\theta) = E_{s_0\sim p(s_0), \tau^0\sim p_\theta(\tau^0|s_0)}\left[p_\theta(r^0_{DDPO} \mid \tau^0)\right] \tag{8}$$

This formulation further enables the estimation of policy gradients. With access to both likelihood and gradients of likelihood, we follow the formulation in (Black et al. 2024; Fan and Lee 2023) to make direct Monte Carlo estimates of $\nabla_\theta \mathcal{J}_{DDRL}$, by sampling, and then performing parameter update. We adopt the importance sampling estimator (Kakade and Langford 2002) and substitute the per-step return with the final reward r^0_{DDPO}.

$$\nabla_\theta \mathcal{J}_{DDRL} = E\left[\sum_{k=0}^{T} \frac{p_\theta(\tau^{k-1}|\tau^k, s_0)}{p_{\theta_{old}}(\tau^{k-1}|\tau^k, s_0)} \times \nabla_\theta \log p_\theta(\tau^{k-1}|\tau^k, s_0)\, r^0_{DDPO}\right] \tag{9}$$

in which the expectation is taken over the trajectories sampled with $p_{\theta_{old}}$, i.e., the previous sampler. The estimator becomes less accurate if $p_{\theta_{old}}$ deviates too much from p_θ. We adopt the trust region (Schulman et al. 2017a) for regularizing the change of θ w.r.t θ_{old}, which in practice we adopt the clipping proposed in proximal policy optimization (Schulman et al. 2017b).

This procedure effectively trains the diffusion policy network to generate high reward action-sequences (τ) that maximize cumulative rewards, improving autonomous driving performance.

EXPERIMENTS AND RESULTS

Experimental Setup

We consider the task of navigation along pre-defined routes in Town04 of CARLA (Dosovitskiy et al. 2017) version 0.9.11 running at 20 Hz. We utilize the predefined routes from the CARLA Leaderboard. We randomly spawn scenarios at several locations along each route. Each scenario presents unique environmental and lighting conditions, with varying number of lanes. Each scenario has a maximum length of upto 500 meters with a maximum of 20 traffic vehicles. Each dynamic traffic vehicle is assigned a velocity

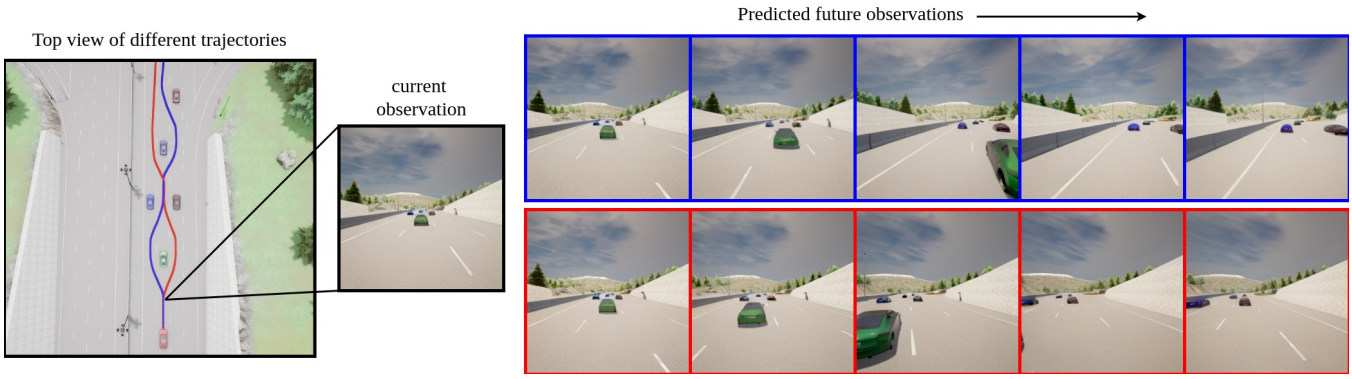

Figure 3: **Multi-Modal Nature of *DPA*:** The top-view visualization illustrates the multi-modal nature of the *Diffusion-based Policy Actor (DPA)*. Starting from the same initial state, the agent follows two distinct trajectories: BlueLeft (in blue) and RedRight (in red), generated using different random seeds. The corresponding predicted future sequences for both trajectories are shown in their respective colors, highlighting *DPA's* ability to model diverse behavioral modes and predict multiple plausible futures from a single state.

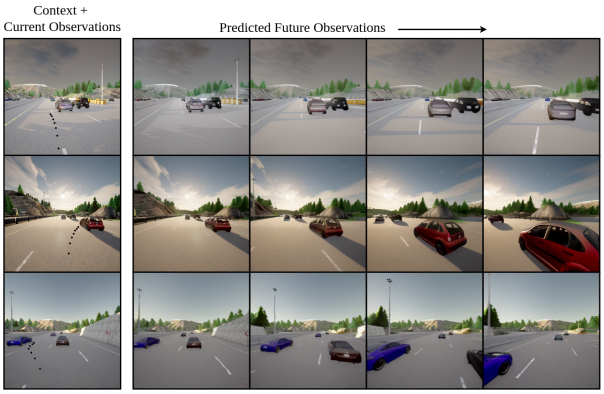

Figure 4: ***VISTAPlan* Future Prediction:** Future observation predictions from the *VISTAPlan World Model*, conditioned on the input trajectory (shown in black) and current observations. Demonstrates the *VISTAPlan's* ability to accurately predict future observations based on the provided context, highlighting its robust trajectory prediction capabilities.

between 1 and 5 km/hr. An episode is terminated upon collision with any traffic vehicle or going out-of-lane bounds. We assign our ego-agent a constant velocity of 7 km/hr and use the default PID controller to follow the output waypoint trajectory. We train and evaluate our model and the baselines on a subset of 10 scenarios.

Training Details and Architectural Details

For world model, we initialize the VISTAPlan model with pre-trained weights trained on approximately 1700 hours of driving videos. For faster convergence, we first give a warm-start to the World Model by training it for 5K steps with data collected using the randomly initialized diffusion policy actor. We use 4 A100s to train VISTAPlan for about 28 hours. Similar to VISTA, we accumulate gradients of 4 steps, giving an effective batch size of 16. We train all methods on each scenario for 1M timesteps.

For DPA, we use a UNet similar to (Janner et al. 2022). We use sampling strategy from DDIM (Song, Meng, and Ermon 2022) and use $T = 50$ denoising steps.

For training *Iso-Dreamer* (Pan et al. 2022) and *Dreamer-V3* (Hafner et al. 2024), we use the default training hyperparameters.

Implementation Details

Action Space Waypoint trajectory is represented in the X-Y cartesian space. To predict the trajectory effectively, we simplify the action space by predicting the incremental ΔX and ΔY.

$$a_t = (\Delta X, \Delta Y) \tag{10}$$

Simplifying ΔX: To reduce action space dimensionality and ensure forward progress, we fix $\Delta X = 1$, allowing the model to focus on lateral adjustments.

Discretizing ΔY: lateral movement ΔY is defined within $(0.5, 0.5)$ meters, divided into 11 equal bins representing specific lateral deviations.

Action Encoding: Each discrete action, representing ΔY, is encoded using a one-hot embedding.

Reward Function The reward function for lane-keeping and collision avoidance is defined as:

$$r_t = v_{ego}^T \hat{u}_h \cdot \Delta t - \xi_1 \cdot \left|Collision\_Cost\right| - \xi_2 \cdot \left|\Delta Y\right| + c \tag{11}$$

Here v_{ego} is the agent's velocity projected onto the highway direction $\hat{u}_h$, normalized and scaled by $\Delta t = 0.05$ to measure highway progression in meters. The $Collision\_cost = 10$ penalizes collisions, and ΔY penalizes large trajectory changes. The constant $c = 1$ encourages longer episodes, with ξ_1 and ξ_2 set to 1.

Horizon and Context Length We set prediction horizon length $H = 9$ and context length $P = 5$.

Baselines

We focus on world model-based approaches using front camera images as the sole input modality and compare with

the following methods:

1. *Dreamer-V3* (Hafner et al. 2024): The third generation in the Dreamer series (Hafner et al. 2020, 2022), incorporating robustness and normalization techniques for stable training.
2. *Iso-Dreamer* (Pan et al. 2022): which focuses on decomposing scenes into action controllable and non-controllable branches.
3. *DQN* (Mnih et al. 2013) and *PPO* (Schulman et al. 2017b): Standard Model-free algorithms.

Metrics

Success Rate (SR %) as the percentage of runs where the ego-agent is able to achieve at-least 90% *route completion* (RC %) with zero instances of *Infractions* which include the instances of going out-of-lane bounds and collisions with other traffic actors.

Episodic Return is defined as the cumulative sum of the *reward* function defined in eq. 11.

Results

Driving Scores and Episodic Returns In table 1, we compare the different RL experts on standard driving metrics across all the scenarios. Our model *Imagine-2-Drive* outperforms all the other experts across all the metrics significantly, with 20% and 14.6% gain in *SR* and *RC* metrics respectively. Infraction/km also reduced by about 50% This clearly highlights the superior combined performance of the DPA and VISTAPlan. All world model based methods outperform the model-free methods, being able to explicitly learn a model of the environment, allowing them to simulate future states and plan actions more efficiently. *Dreamer-V3*, which combines robustness and normalization techniques for stable training outperforms *Iso-Dreamer*. PPO outperforms DQN by directly optimizing the policy with stable updates, making it more suitable for both continuous and discrete action spaces while improving exploration through stochastic policies. Fig. 5 further supports these observations which shows the higher episodic returns for *Imagine-2-Drive* during training with 1M timesteps averaged over the scenarios.

Table 1: **Driving Metrics Comparison:** We compare the *Imagine-2-Drive* and the baselines on standard driving metrics across all the scenarios. We run each model on all scenarios with 3 random seeds.

Model	**SR(%)** ↑	**Infraction/Km** ↓	**RC(%)** ↑
DQN	0.0	6.27	27.63
PPO	16.66	4.61	50.02
Iso-Dreamer	56.66	1.65	60.33
Dreamer-V3	63.33	1.52	67.53
Imagine-2-Drive	**83.33**	**0.70**	**82.13**

Prediction Fidelity Table 3 shows a quantitative comparison of world models prediction fidelity. *Imagine-2-Drive* outperforms in temporal consistency and fidelity metrics by

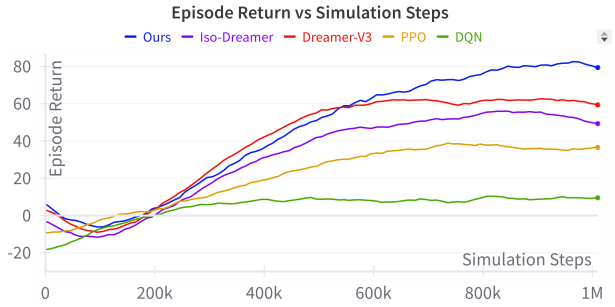

Figure 5: Episodic return of *Imagine-2-Drive* (Ours) and other different RL agents averaged over the scenarios. A moving average with a window size of 50 is applied

Table 2: **Driving Metrics Comparison for Ablations:** We compare the *Imagine-2-Drive* and the baselines, along with combining (DPA) with other world-model based methods on standard driving metrics across all the scenarios. We run each model on all scenarios with 3 random seeds.

Model	**SR(%)** ↑	**Infraction/Km** ↓	**RC(%)** ↑
Iso-Dreamer	56.66	1.65	60.33
Dreamer-V3	63.33	1.52	67.53
DPA + Iso-Dreamer	73.33	1.17	70.87
DPA + Dreamer-V3	83.33	0.97	75.09
Imagine-2-Drive	**83.33**	**0.70**	**82.13**

72 and 290 in FID and FVD scores respectively. Unlike autoregressive models, which degrade over time, SVD models enhance temporal consistency by incorporating temporal attention. Fig. 4 qualitatively highlights the high fidelity prediction of our world model conditioned on an input trajectory.

Table 3: Comparison of prediction fidelity scores of different world models over CARLA driving sequences

World Model	**FID** ↓	**FVD** ↓
DriveGAN	67.1	281.9
Iso-Dreamer	102.24	421.56
Dreamer-V3	89.29	324.07
VISTAPlan	**17.09**	**130.39**

ABLATION STUDIES

DPA analysis

Table 2 evaluates the DPA within the Iso-Dreamer and Dreamer-V3 frameworks by comparing performance with and without the actor (rows 1 vs. 3) and (rows 2 vs. 4). DPA coupled models outperform by atleast 10% and 8% on SR and RC metrics. This analysis underscores the influence of the DPA on trajectory prediction and demonstrates its effectiveness across different world model architectures. The results demonstrate that the diffusion policy actor enhances

the ability to model complex, multi-modal action distributions, improving trajectory prediction accuracy across different world model architectures. This underscores the actor's adaptability and potential for broader applications in reinforcement learning and decision-making tasks.

VISTAPlan analysis

Table 2 assesses prediction fidelity and environmental dynamics across various world models, using our diffusion policy actor consistently across all models. This isolates the impact of each world model and allows for a focused assessment of each world model's ability to capture underlying environmental dynamics and its influence on trajectory prediction performance. Results in (rows 2, 3 and 4) highlight the VISTAPlan's superior performance with a 20% and 14.6% gain in *SR* and *RC* metrics respectively, underscoring the importance of accurate future state predictions and validating VISTA's role as the foundation for our world model framework.

Context Length analysis

To analyze the effect of POMDP approximation with MDP using the *State Encoder*, Fig. 6 compares the episodic returns of *Imagine-2-Drive* using different *context lengths* (CL) for the State Encoder, with $P = 5$ fixed for *VISTA*. The performance drops with $CL = 1$, compared to $CL = 3$ and 5. This can be attributed to shorter CLs which provide insufficient historical data, limiting the model's ability to infer the current state in partially observable environments. In contrast, longer context lengths $(CL = 3, 5)$ capture more temporal information, improving state estimation and decision-making, and resulting in better performance.

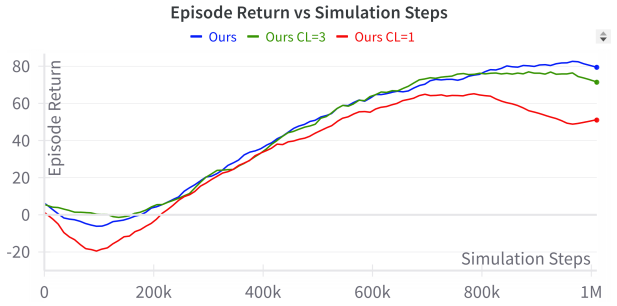

Figure 6: Episodic return of *Imagine-2-Drive* (Ours with CL=5) with different *Context Length* (CL) averaged over the scenarios. A moving average with a window size of 50 is applied.

CONCLUSION & FUTURE WORK

Conclusion: In this work, we introduced *Imagine-2-Drive* , a high-fidelity world model leveraging the VISTA framework for long-horizon trajectory planning in autonomous driving. By incorporating a diffusion-based prediction model, we overcame the limitations of single-step transition models, enabling simultaneous trajectory generation and enhanced temporal consistency. DPA, a diffusion based policy improves exploration and robustness by effectively capturing multimodal action distributions. Our empirical results in CARLA demonstrate significant gains in trajectory prediction and planning accuracy, showcasing the superiority of our approach over SOTA world model and model-free methods.

Future Work: We intend to extend *Imagine-2-Drive* to other domains such as robotic manipulation, where accurate long-horizon planning and multimodal prediction are crucial. Expanding the versatility of our approach across domains will further validate its generalizability and adaptability.

References

Ankile, L.; Simeonov, A.; Shenfeld, I.; and Agrawal, P. 2024. JUICER: Data-Efficient Imitation Learning for Robotic Assembly.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2024. Training Diffusion Models with Reinforcement Learning.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; Jampani, V.; and Rombach, R. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion.

Clark, K.; Vicol, P.; Swersky, K.; and Fleet, D. J. 2024. Directly Fine-Tuning Diffusion Models on Differentiable Rewards.

Deng, F.; Jang, I.; and Ahn, S. 2021. DreamerPro: Reconstruction-Free Model-Based Reinforcement Learning with Prototypical Representations.

Ding, Z.; Zhang, A.; Tian, Y.; and Zheng, Q. 2024. Diffusion World Model: Future Modeling Beyond Step-by-Step Rollout for Offline Reinforcement Learning.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An Open Urban Driving Simulator.

Fan, Y.; and Lee, K. 2023. Optimizing DDPM Sampling with Shortcut Fine-Tuning.

Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; and Lee, K. 2023. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models.

Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; and Levine, S. 2021. D4RL: Datasets for Deep Data-Driven Reinforcement Learning.

Gao, R.; Chen, K.; Xie, E.; Hong, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2024a. MagicDrive: Street View Generation with Diverse 3D Geometry Control.

Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024b. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2020. Dream to Control: Learning Behaviors by Latent Imagination.

Hafner, D.; Lillicrap, T.; Norouzi, M.; and Ba, J. 2022. Mastering Atari with Discrete World Models.

Hafner, D.; Pasukonis, J.; Ba, J.; and Lillicrap, T. 2024. Mastering Diverse Domains through World Models.

Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; and Corrado, G. 2023. GAIA-1: A Generative World Model for Autonomous Driving.

Janner, M.; Du, Y.; Tenenbaum, J. B.; and Levine, S. 2022. Planning with Diffusion for Flexible Behavior Synthesis.

Jiang, C. M.; Cornman, A.; Park, C.; Sapp, B.; Zhou, Y.; and Anguelov, D. 2023. MotionDiffuser: Controllable Multi-Agent Motion Prediction using Diffusion.

Kakade, S.; and Langford, J. 2002. Approximately Optimal Approximate Reinforcement Learning.

Kim, S. W.; Philion, J.; Torralba, A.; and Fidler, S. 2021. DriveGAN: Towards a Controllable High-Quality Neural Simulation.

Li, Q.; Jia, X.; Wang, S.; and Yan, J. 2024. Think2Drive: Efficient Reinforcement Learning by Thinking in Latent World Model for Quasi-Realistic Autonomous Driving (in CARLA-v2).

Liu, J.; Hang, P.; Zhao, X.; Wang, J.; and Sun, J. 2024. DDM-Lag : A Diffusion-based Decision-making Model for Autonomous Vehicles with Lagrangian Safety Enhancement.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing Atari with Deep Reinforcement Learning.

Pan, M.; Zhu, X.; Wang, Y.; and Yang, X. 2022. Iso-Dream: Isolating and Leveraging Noncontrollable Visual Dynamics in World Models.

Pearce, T.; Rashid, T.; Kanervisto, A.; Bignell, D.; Sun, M.; Georgescu, R.; Macua, S. V.; Tan, S. Z.; Momennejad, I.; Hofmann, K.; and Devlin, S. 2023. Imitating Human Behaviour with Diffusion Models.

Reuss, M.; Li, M.; Jia, X.; and Lioutikov, R. 2023. Goal-Conditioned Imitation Learning using Score-based Diffusion Policies.

Saha, K.; Mandadi, V.; Reddy, J.; Srikanth, A.; Agarwal, A.; Sen, B.; Singh, A.; and Krishna, M. 2023. EDMP: Ensemble-of-costs-guided Diffusion for Motion Planning.

Schulman, J.; Levine, S.; Moritz, P.; Jordan, M. I.; and Abbeel, P. 2017a. Trust Region Policy Optimization.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017b. Proximal Policy Optimization Algorithms.

Shah, D.; Sridhar, A.; Dashora, N.; Stachowicz, K.; Black, K.; Hirose, N.; and Levine, S. 2023. ViNT: A Foundation Model for Visual Navigation.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models.

Sridhar, A.; Shah, D.; Glossop, C.; and Levine, S. 2023. NoMaD: Goal Masked Diffusion Policies for Navigation and Exploration.

Tan, M.; and Le, Q. V. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2023. Diffusion Model Alignment Using Direct Preference Optimization.

Wang, L.; Zhao, J.; Du, Y.; Adelson, E. H.; and Tedrake, R. 2024. PoCo: Policy Composition from and for Heterogeneous Robot Learning.

Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; Zhu, J.; and Lu, J. 2023. DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving.

Yang, B.; Su, H.; Gkanatsios, N.; Ke, T.-W.; Jain, A.; Schneider, J.; and Fragkiadaki, K. 2024. Diffusion-ES: Gradient-free Planning with Diffusion for Autonomous Driving and Zero-Shot Instruction Following.

Ze, Y.; Zhang, G.; Zhang, K.; Hu, C.; Wang, M.; and Xu, H. 2024. 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations.