AD-R1: Closed-Loop Reinforcement Learning for End-to-End Autonomous Driving with Impartial World Models

Supplementary Material

AD-R1: Closed-Loop Reinforcement Learning for End-to-End Autonomous Driving with Impartial World Models

Supplementary Material

Overview

This supplementary material provides additional details to support the claims and ensure the reproducibility of the main paper. The content is organized as follows:

- **Sec. A**: Detailed network architecture, hyperparameters for the Impartial World Model (IWM), and the RL training configuration.
- **Sec. B**: Elaboration on the nuScenes-CF dataset and the mathematical formulation of Counterfactual Synthesis.
- **Sec. C**: Strict mathematical definitions for the Risk Foreseeing Benchmark (RFB).
- **Sec. D**: Additional quantitative results focusing on the sensitivity analysis of synthetic data ratios.
- **Sec. E**: Qualitative analysis and description of the attached demo.

A. Implementation Details

A.1. Impartial World Model (IWM) Architecture

Our IWM consists of a VQ-VAE-based tokenizer and a Transformer-based autoregressive forecaster based on I^2-World [28].

4D Scene Tokenizer. We voxelize the local scene into a grid of spatial resolution 200×200 and height resolution 16. The physical range covered is $[-40\text{m}, 40\text{m}]$ in X, Y and $[-1\text{m}, 5.4\text{m}]$ in Z, resulting in a voxel size of $0.4\text{m} \times 0.4\text{m} \times 0.4\text{m}$, following Occ-3D [39].

- **Encoder/Decoder:** We utilize a 3D-VAE architecture with 4 downsampling stages.
- **Vector Quantization:** We use a codebook of size $N = 512$ with an embedding dimension of $D = 128$.

4D Forecaster. The forecaster is a encoder-decoder Transformer tailored for 4D occupancy generation.

- **Structure:** 24 Attention Layers, 16 Attention Heads, Hidden Dimension $d_{model} = 1024$.
- **Context:** The model takes 2 seconds of historical context (4 frames at 2Hz) and predicts 3 seconds into the future (6 frames).

Training Details. We use AdamW optimizer with 1×10^{-3} as the base learning rate, 256 as the batch size. The Tokenizer is trained for 30 epochs while the Forecaster is trained for 50 epochs. All training is performed on $8\times$NVIDIA H20 GPUs.

A.2. RL Training Configuration (AD-R1)

We employ the Group Relative Policy Optimization (GRPO) algorithm adapted for continuous trajectory refinement.

Training Setup.

- **Policy Network:** We fine-tune the denoising head of the pre-trained DiffusionDrive [27] and ReCogDrive [24].
- **Group Sampling (G):** Due to the computational cost of 4D world model rollouts, we set the group size $G = 64$ and $G = 8$ for DiffusionDrive and ReCogDrive, respectively. This provides a sufficient baseline variance for advantage estimation while fitting within GPU memory constraints.
- **Optimization:** We use the AdamW optimizer with a learning rate of 1×10^{-5} and 4×10^{-5} for DiffusionDrive and ReCogDrive.
- **Hardware:** Training is performed on 8 $\times$ NVIDIA H20 GPUs. The refinement stage takes approximately 24 hours for 10 epochs on the navsim [6] training set.

A.3. Reward Function Details

The reward shaping is critical for guiding the agent towards safety without compromising progress. Table 4 details the specific weights used in Eq. (18) of the main paper.

B. The nuScenes-CF Dataset

B.1. Kinematic Trajectory Generation

To generate the unsafe ego-trajectory $\tilde{\mathcal{T}}_{ego}$ for Counterfactual Synthesis, we employ a Blending Kinematic Model. Let $\mathbf{p}_t, \mathbf{v}_t$ be the vehicle's current state and $\mathbf{p}_{target}$ be the designated collision point (e.g., the center of another vehicle or a wall).

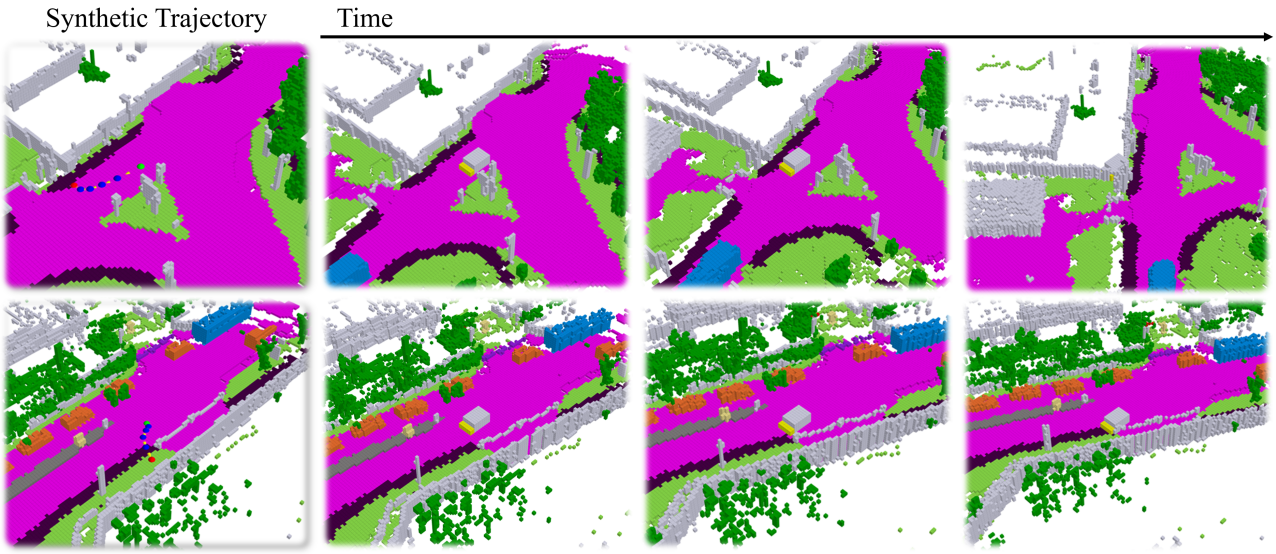

Figure 5. Examples of synthetic unsafe trajectories for the Counterfactual Data (Off-road). Red dot indicates the end of the trajectory while the green one indicates the start point.

The synthesized position $\mathbf{p}_{t+k}$ at future timestep k is computed as:

$$\mathbf{p}_{t+k} = (1 - \gamma_k) \cdot \mathbf{p}^{(k)}_{inertial} + \gamma_k \cdot \mathbf{p}^{(k)}_{intercept} \tag{18}$$

where:

- $\mathbf{p}^{(k)}_{inertial}$ represents the position if the vehicle continued with its historical momentum.
- $\mathbf{p}^{(k)}_{intercept}$ is the position required to linearly intercept the target $\mathbf{p}_{target}$.
- $\gamma_k \in [0, 1]$ is a time-varying blending coefficient, defined as $\gamma_k = (k/K)^2$. The quadratic growth ensures a smooth deviation that mimics a loss of control or a gradual drift, rather than an unrealistic sharp turn.

B.2. Dataset Statistics and Sampling Strategy

We constructed a dedicated pool of counterfactual scenarios consisting of **1,800 synthetic clips** derived from the nuScenes training set. For evaluation, we curated a separate set of **450 clips** to serve as the Risk Foreseeing Benchmark (RFB).
Training Sampling Strategy. It is important to note that the "80:20 ratio" refers to the data distribution seen by the model during training, not the absolute size of the datasets. We employ a **weighted sampling strategy**: in each training iteration, 80% of the samples are drawn from the original real-world nuScenes dataset (safe expert data), and 20% are sampled from our synthetic counterfactual pool (unsafe failure data). This ensures the model maintains strong priors on realistic scene dynamics while receiving sufficient exposure to rare failure modes.

The distribution of failure types within the synthetic pool is balanced as follows:

- **Dynamic Collisions (40%):** Inter-vehicle collisions and vehicle-pedestrian accidents.
- **Static Collisions (30%):** Collisions with barriers, traffic cones, and buildings.
- **Off-Road/Sidewalk (30%):** Excursions into non-drivable zones.

B.3. Counterfactual Example

We visualize examples about synthetic counterfactual occupancy data to show **Penetration** (Fig. 6), **Off-road** (Fig. 5) and **Collision** (Fig. 7).

C. Experimental Setup and Benchmarks

C.1. Risk Foreseeing Benchmark (RFB)

To quantify the core contribution of our AD-R1 for overcoming the optimistic bias, we evaluate world models on our proposed **Risk Foreseeing Benchmark (RFB)**. The RFB consists of a curated set of **450 challenging scenarios** (clips) with ground-

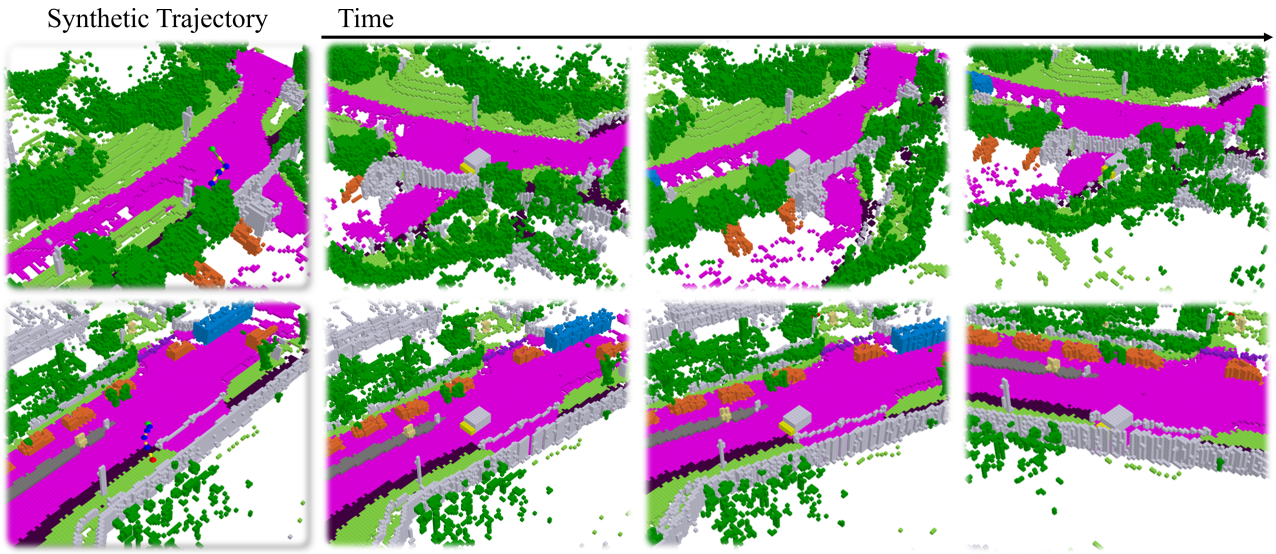

Figure 6. Examples of synthetic unsafe trajectories for the Counterfactual Data (Penetration). Red dot indicates the end of the trajectory while the green one indicates the start point.

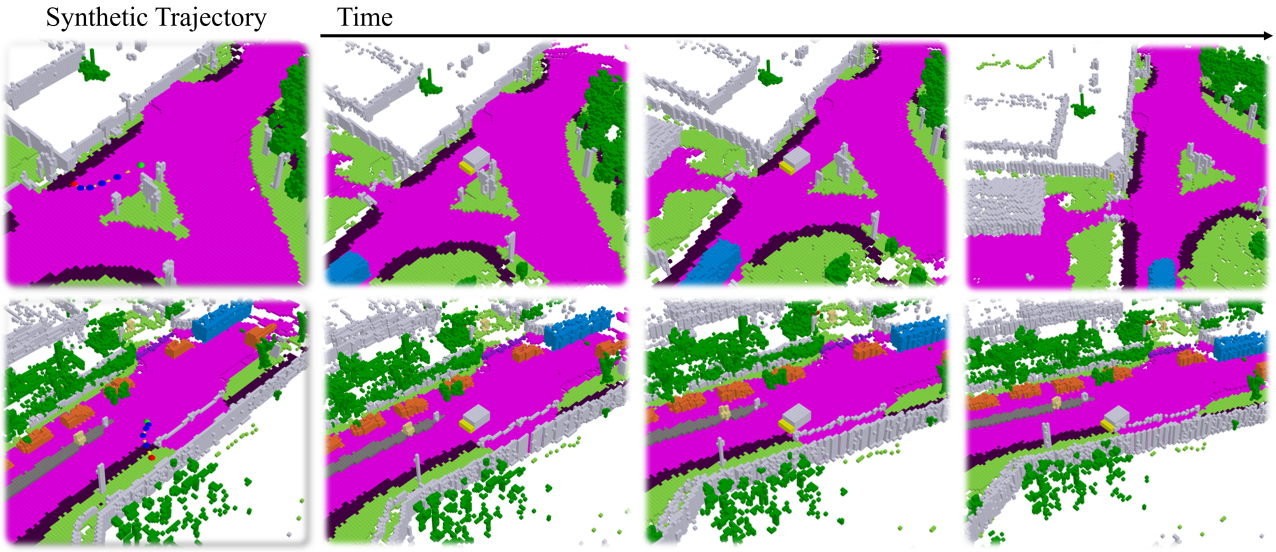

Figure 7. Examples of synthetic unsafe trajectories for the Counterfactual Data (Collision). Red dot indicates the end of the trajectory while the green one indicates the start point.

truth unsafe ego-trajectories spanning 3 seconds. Unlike standard benchmarks that evaluate on safe expert driving, RFB strictly tests the model's ability to predict negative outcomes.

We measure performance using the following three key metrics:

1. **Global Scene Fidelity (*G*-IoU):** This metric calculates the mean Intersection-over-Union (mIoU) across the entire scene (including static background and dynamic agents). It serves as a sanity check to penalize unrelated hallucinations and ensure the overall scene structure remains consistent.

$$G\text{-IoU} = \frac{1}{|C|} \sum_{c \in C} \frac{|\hat{O}_c \cap O_c|}{|\hat{O}_c \cup O_c|} \tag{19}$$

where C represents the semantic classes in the scene.

Table 4. **Reward Coefficients Configuration.** Negative weights indicate penalties.

Component	Symbol	Weight (w)	Description
Safety-Critical (Heavily Penalized)			
Collision	w_{coll}	-20.0	Penalty if VC-IoU > 0.
Off-Road	w_{off}	-10.0	Penalty for non-drivable areas.
Clearance	w_{clr}	-15.0	Vertical height violation.
Comfort & Task			
Stability	w_{stab}	-2.0	Variance of Z-axis under ego.
Progress	w_{prog}	$+1.0$	Reward for distance traveled.
Velocity	w_{vel}	$+0.5$	Alignment with speed limit.

2. **Failure IoU (f-IoU):** To strictly measure the prediction of danger, we define f-IoU. This metric focuses on the **critical failure regions** (e.g., collision zones surrounding the ego car). We define a local volume V_{crit} centered around the ego-vehicle's future trajectory. f-IoU calculates the IoU of occupied voxels specifically within this volume:

$$f\text{-IoU} = \frac{|\hat{O}_{occ} \cap O_{occ} \cap V_{crit}|}{|\hat{O}_{occ} \cup O_{occ} \cap V_{crit}|} \tag{20}$$

   A high f-IoU indicates the model correctly predicts that "something is there" (e.g., a wall or another car) directly in the ego-vehicle's path, rather than hallucinating empty space.
3. **Dynamic Agent Fidelity (DAF):** Previous optimistic models tend to make other agents vanish ("ghosting") to avoid collisions. DAF measures the average Instance-IoU of other dynamic agents (Vehicles, Pedestrians) to ensure that risk-inducing actors are preserved in the prediction.

$$\text{DAF} = \frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} \text{IoU}(\hat{M}_i, M_i) \tag{21}$$

   where M_i is the occupancy mask of the i-th dynamic agent.

C.2. Driving Policy Evaluation

On the NavSim benchmark, we evaluate the final agent's performance using safety and planning metrics. For planning quality, we use the **PDSM** [6], which provides a holistic score for driving quality, weighting progress, comfort, and safety.

C.3. Baselines and Comparison Schemes

We conduct a comprehensive comparison to validate both the world model and the downstream policy refinement.
1. For World Model Evaluation: We compare our **Impartial World Model (IWM)** against strong baseline world models trained only on safe data to demonstrate the impact of optimistic bias:
- **I^2-World** [29]: A state-of-the-art 4D occupancy world model using VQ-VAE and autoregressive transformers.
- **DOME** [14]: A diffusion-based occupancy world model.

2. For Policy Refinement: We apply our plug-and-play framework, powered by IWM, to two state-of-the-art, publicly available, pre-trained end-to-end driving agents. We compare the original performance against our refined versions:
- **DiffusionDrive** [27]: A diffusion-based planning policy.
- **ReCogDrive** [24]: A VLA-based agent integrating vision-language features.

D. Additional Experimental Results

D.1. Ablation: Synthetic Data Ratio

We investigate the impact of the ratio of synthetic counterfactual data mixed into the training set. Table 5 shows that a balance is crucial.

While increasing synthetic data to 50% slightly improves f-IoU (collision prediction), it degrades the general scene generation quality (G-IoU) because the synthetic physics are simplified approximations. The 80:20 ratio provides the optimal balance.

Table 5. **Effect of Synthetic Data Ratio.** We find that an 80:20 mix offers the best trade-off between scene fidelity (G-IoU) and safety awareness (f-IoU).

Real : Syn Ratio	G-IoU ↑	f-IoU ↑	PDMS ↑
100 : 0 (Baseline)	21.01	14.21	85.3
90 : 10	35.40	38.12	88.2
80 : 20 (Ours)	**40.21**	45.91	**89.8**
50 : 50	38.50	**46.10**	87.4

Table 6. **Ablation on Reward Components.** Each component plays a distinct role: Safety penalties are crucial for collision avoidance, Progress rewards prevent the "frozen robot" problem, and Comfort rewards ensure smooth 3D trajectory planning.

Configuration	Coll. Rate ↓	PDMS ↑
DiffusionDrive	1.8%	88.1
Full AD-R1 (Ours)	**1.6%**	**89.8**
w/o Safety ($w_{coll}, w_{off}, w_{clr} = 0$)	1.8%	88.5
w/o Comfort ($w_{stab} = 0$)	1.8%	87.6
w/o Progress ($w_{prog}, w_{vel} = 0$)	2.0%	76.8

D.2. Ablation: Reward Function Components

To validate the contribution of each component in our multi-faceted reward function R_{total}, we conduct an ablation study based on DiffusionDrive by selectively removing specific reward terms during the RL training phase. Table 6 summarizes the impact on the final policy performance.

Impact of Safety Penalties: When safety penalties are removed (Row 3), the Collision Rate reverts to the baseline level of 1.8%. While the absolute difference (0.2%) appears small, it represents a meaningful improvement given that the baseline policy is already highly optimized. More importantly, our Impartial World Model provides a mechanism to penalize subtle unsafe behaviors that heuristics miss. **Impact of Comfort:** Removing the 3D stability reward (Row 4) results in a PDMS of 87.6, which is even lower than the original baseline (88.1). This suggests that without comfort constraints, the RL process may optimize for safety or speed at the expense of ride smoothness, generating jerky trajectories that degrade the overall driving quality. **Impact of Progress:** Removing task-oriented rewards (Row 5) causes a catastrophic drop in PDMS to 76.8. Interestingly, the collision rate slightly increases to 2.0%, implying that an agent lacking goal-directed behavior may behave unpredictably or fail to exit dangerous zones efficiently, thereby increasing risk.

E. Qualitative Visualization

In Figure 8, we show the behavior of a baseline agent versus our refined agent in a critical scenario. The original agent, failing to anticipate a sudden cut-in, generates a plan that leads to a collision. Our refined agent, having learned from countless imagined failures, exhibits more defensive behavior, correctly predicting the risk and executing a safe braking maneuver.

Figure 8. Behavior of an agent with and without *AD-R1* refinement. **Left:** The original agent's plan results in a collision or off-road. **Right:** Our refined agent safely avoids the hazard.