CausalDrive: Real-time Causal World Models for Autonomous Driving

Tianyi Yan[1], Huan Zheng[1], Dubing Chen[1], Meizhi Qu[3], Yingying Shen[2], Lijun Zhou[2], Mingfei Tu[2], Bing Wang[2], Guang Chen[2], Hangjun Ye[2], Haiyang Sun[2], Cheng-zhong Xu[1], and Jianbing Shen[1⋆✉]

[1] SKL-IOTSC, CIS, University of Macau, Macau
[2] Xiaomi EV, Beijing, China
[3] CASIA, Beijing, China
tianyi.yan123@gmail.com

Abstract. World models have emerged as a promising paradigm for scaling autonomous driving (AD) data, yet existing video generative models fall short as interactive simulators. Layout-conditioned renderers rely on "oracle" future trajectories of all background agents, rendering them strictly non-reactive. Conversely, pure action-conditioned predictors lack semantic control over complex interactions and suffer from prohibitive diffusion latencies, hindering closed-loop policy learning. To bridge this gap, we present CausalDrive, a controllable, real-time foundation driving world renderer. CausalDrive operates solely on the initial front-view frame, the ego-vehicle's trajectory, and a macroscopic text prompt. By excluding future NPC layouts, we compel the model to intrinsically predict causal interactions, enabling text-driven control over Driving Sociology—allowing users to dynamically orchestrate diverse counterfactual reactions to identical ego-actions. To overcome the efficiency bottleneck and address the covariate shift in autoregressive generation, we propose a novel Context-Forced DMD architecture. This combines continuous flow-matching with a self-correcting distillation objective, achieving interactive speeds of 12 FPS. This breakthrough transforms the passive video generator into a playable neural simulator. We demonstrate its versatility across three downstream applications: (1) generative closed-loop evaluation with significantly mitigated collision artifacts, (2) large-scale Reinforcement Learning (RL) post-training driven by a Video2Reward module, and (3) real-time human-in-the-loop simulation. Extensive experiments validate that policies trained within CausalDrive's reactive scenarios exhibit superior interaction capabilities in the real world.

Keywords: World Model · Simulation · Video Diffusion

⋆ ✉ Corresponding author. This work was supported by the Science and Technology Development Fund of Macau SAR (FDCT) under grants 0134/2025/RIA2 and 0102/2023/RIA2, and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC).

1 Introduction

The paradigm of Autonomous Driving (AD) is shifting from modular pipelines to end-to-end learning [4, 15, 18, 22, 25]. Within this transition, learning a world model [9, 10, 35, 37, 40, 41, 46] capable of simulating plausible future outcomes is envisioned as a key milestone toward high-level machine intelligence. Recent advances [2,13] in generative AI have achieved remarkable success in high-fidelity visual synthesis, precise instruction controllability [9, 10, 24], and the simulation of safety-critical corner cases [33, 39].

However, a critical dichotomy remains. Existing models [9, 10, 23, 41] primarily function as "offline data engines" for augmentation, rather than interactive environments. SOTA driving world models [9,10,19,24,37] exhibit stunning photorealism but largely operate as "open-loop" video predictors. Relying heavily on log-replay, these models often ignore the ego-vehicle's novel actions or fail to model the causal reactions of surrounding agents (e.g., a neighbor yielding when the ego merges). Consequently, the generated environment remains a passive "background" rather than an active participant. Furthermore, the prohibitive inference latency of diffusion-based architectures (often seconds per frame) renders them impractical for computationally intensive tasks like online Reinforcement Learning (RL) [27, 28].

To serve as a viable alternative to real-world data collection and casual-aware simulators, we argue that the next generation of models must evolve into Foundation Driving World Renderers, holistic systems that are not just "dreamers" of video, but real-time, reactive, and physically grounded. Inspired by recent discussions on driving simulation [5, 7, 21, 40, 42, 44], we identify four essential pillars required to unlock the full potential of world models for downstream applications: (1)High-Fidelity Visual Synthesis (Photorealism): To minimize the domain gap, the model must produce sensor-realistic outputs where perception networks can generalize directly from simulation to reality. (2) Strict Action-Controllability: The generated dynamics must faithfully reflect control signals (steering, throttle) to prevent "action-ignoring" or posterior collapse, ensuring a "left turn" command results in a physically accurate visual transition. (3) Data-Driven Reactive Agents (Driving Sociology): Unlike rigid "log-replay" systems, the model must implicitly capture the "sociology" of driving, the causal, reactive interactions between the ego-vehicle and surrounding agents (e.g., yielding, merging, or reacting to hazards) (4) Real-Time Inference Efficiency: Simulation-based applications, such as large-scale reinforcement learning (RL) or human-in-the-loop testing, require interactive frame rates ($>$ 10 FPS) to be computationally feasible.

Despite individual advancements [11, 19, 28, 38, 42, 47], integrating these four pillars into a unified framework remains an open challenge. Current models often trade latency for quality or sacrifice interactivity for stability. In this work, we propose **CausalDrive**, a Foundation Driving World Renderer designed to satisfy these desiderata simultaneously.

To systematically address the challenges of driving sociology and real-time inference, we introduce **CausalDrive**, an Interactive Driving World Model de-

signed to unify photo-realistic generation with closed-loop reactive logic. Our framework encompasses three core innovations across data curation, model architecture, and acceleration. To cultivate predictive driving sociology without relying on "oracle" layout leaks, we establish SocioDrive-Bench. Unlike layout-conditioned models that passively render future bounding boxes, CausalDrive predicts futures strictly conditioned on the initial frame, ego-trajectory, and semantic prompts. By leveraging a vision-language pipeline to extract explicit causal interaction labels—such as active pressure, defensive driving, and complex negotiation—we transform raw geometric logs into sociologically-aware training curricula. Architecturally, directly converting a bidirectional video diffusion model into an autoregressive (AR) streaming simulator causes expectation collapse due to the violation of probability flow injectivity. We bridge this gap by constructing a Sociology-Aware Causal AR Teacher. Conditioned via adaptive layer normalization for geometric poses and cross-attention for semantic sociology prompts, this continuous flow-matching teacher ensures mathematically sound and dynamically accurate step-by-step rollouts.

Finally, to overcome exposure bias and achieve real-time simulation speeds, we propose Self-Corrective Forcing (SCF) for diffusion distillation. Autoregressive generation over long horizons inherently accumulates errors. Standard distillation fails in this regime due to a context mismatch between the teacher's perfect history and the student's flawed history. By explicitly forcing the AR teacher to evaluate gradients based on the student's imperfect self-rollout, SCF teaches the student to auto-correct temporal compounding errors. This reduces the required function evaluations to ≤ 4 steps, achieving a throughput of ≥ 12 FPS with sub-second latency.

By integrating these innovations, CausalDrive effectively transforms a passive video generator into a playable neural simulator. We demonstrate its versatility across three downstream applications: closed-loop evaluation with highly consistent kinematic dynamics, efficient RL training utilizing a Video-to-Reward (V2R) module, and human-in-the-loop simulation with real-time WASD control.

Our main contributions are summarized as follows:

- We propose **CausalDrive**, a real-time foundation world model that strictly avoids future-layout conditioning, unifying photo-realistic generation with closed-loop reactive logic for autonomous driving.
- We introduce **SocioDrive-Bench**, establishing a novel VLM/LLM-driven taxonomy to extract and formalize causal driving sociology (e.g., yielding, negotiating), enabling semantic control over multi-agent interactions.
- We identify the architectural and context-mismatch bottlenecks in streaming video generation, and propose a novel **Context-Forced DMD** framework over a **Causal AR Flow-Matching Teacher**. This enables robust, infinite-horizon generation at ≥ 12 FPS.
- Extensive experiments demonstrate that CausalDrive acts as a superior neural simulator. RL policies trained safely within its hallucinated, safety-critical scenarios exhibit significantly improved interaction capabilities.

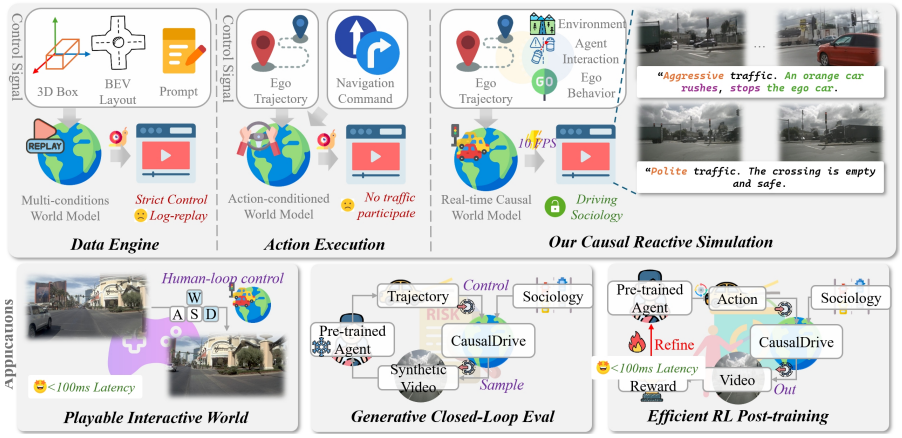

Fig. 1: Top: Semantic Control of Driving Sociology: Unlike layout-conditioned renderers that require "oracle" future boxes of all agents, CausalDrive generates interactive futures. Given the *exact same* initial state and ego-trajectory (nudging into an intersection), modifying the prompt allows us to dynamically simulate diverse counterfactual reactions from surrounding agents (e.g., politely yielding vs. aggressively rushing). **Down: The reactive API that powers three distinct downstream applications:** (1) **Playable Interactive World**, offering a playable, low-latency virtual world controlled via WASD/steering wheels. (2) **Closed-Loop Evaluation**, providing a "zero-ghosting" proving ground for planners; and (3) **RL Post-Training**, where an agent safely learns from simulated safety-critical scenarios via a Video2Reward (V2R) module.

2 Related Work

2.1 Generative World Models for Autonomous Driving

The advent of large-scale generative models has catalyzed the development of world models for autonomous driving (AD), aiming to synthesize infinite driving scenarios for data augmentation and policy learning [14, 35, 46]. Current video-based driving models broadly fall into two paradigms based on their conditioning strategies. The first paradigm, *Layout-Conditioned Renderers* (e.g., MagicDrive [9, 10], UniScene [19], Panacea [37]), synthesizes high-fidelity videos conditioned on the complete future trajectories or 3D bounding boxes of all traffic participants. While producing stable visuals, these models act as "oracles" rather than simulators; they are fundamentally non-reactive because the future states of surrounding agents are hard-coded inputs rather than dynamically inferred responses. The second paradigm, *Action-Conditioned Predictors* (e.g., Vista [11], Orbis [29], Drive-WM [36]), predicts future frames based solely on the ego-vehicle's action and past context. While conceptually closer to true world models, they often suffer from "posterior collapse" (ignoring ego-actions) and lack macroscopic semantic control over the diverse, long-tail behaviors of surrounding agents. In contrast, **CausalDrive** takes a hybrid approach: we strictly

omit future NPC layouts to compel genuine causal prediction, while introducing a macroscopic sociology prompt to semantically orchestrate the predicted interactions, achieving both strict action-controllability and semantic diversity.

2.2 Reactive Simulation and Driving Sociology

Closed-loop policy evaluation and Reinforcement Learning (RL) require environments where non-player characters (NPCs) react logically to the ego-vehicle. Traditional graphics-based simulators (e.g., CARLA [8], MetaDrive [20]) provide interactivity but suffer from a severe visual sim-to-real gap and utilize rigid, rule-based NPCs that fail to capture human-like stochasticity. To address the visual gap, data-driven approaches often rely on *Log-Replay* from real-world datasets (e.g., nuPlan [3,6]). However, log-replay is inherently open-loop; if the ego-vehicle deviates from the logged trajectory, surrounding vehicles fail to react, leading to the unrealistic "ghost effect" (false-positive collisions) [40, 44]. While eliminating these artifacts entirely remains an open challenge, CausalDrive significantly mitigates them by strictly enforcing causal dependencies. We introduce **SocioDrive-Bench**, shifting the focus from geometric trajectories to *Driving Sociology*. By leveraging Large Vision-Language Models (VLMs), we explicitly extract causal interaction labels (e.g., yielding, negotiating), bridging the gap between raw perceptual logs and cognitive reactive simulation.

2.3 Efficient Video Generation and Distillation

Transforming video diffusion models into real-time interactive simulators remains a formidable challenge due to the high Number of Function Evaluations (NFE) required by iterative solvers. While generalized distillation techniques like Consistency Models [32] and Distribution Matching Distillation (DMD) [45] have accelerated text-to-image and short-video generation, applying them to open-ended, streaming driving simulation introduces unique bottlenecks. State-of-the-art video Diffusion Transformers (DiTs) [30] rely on bidirectional temporal attention. Directly distilling a bidirectional teacher into an Autoregressive (AR) streaming student violates Probability Flow ODE (PF-ODE) injectivity—where multiple valid futures collapse into blurred predictions [16, 48]. Furthermore, long-horizon AR generation suffers from *exposure bias*, where small student errors compound over time. To systematically resolve this, we propose a mathematically sound pipeline: we first train a **Causal AR Teacher** via Continuous Flow Matching to guarantee conditional alignment, and then introduce **Context-Forced DMD**. By forcing the teacher to evaluate gradients on the student's flawed autoregressive rollouts, we explicitly correct exposure bias, successfully compressing high-fidelity reactive generation to ≤ 4 steps (≥ 12 FPS).

3 Methodology

Our ultimate goal is to construct a real-time, interactive driving world model capable of modeling the causal interactions among various traffic participants

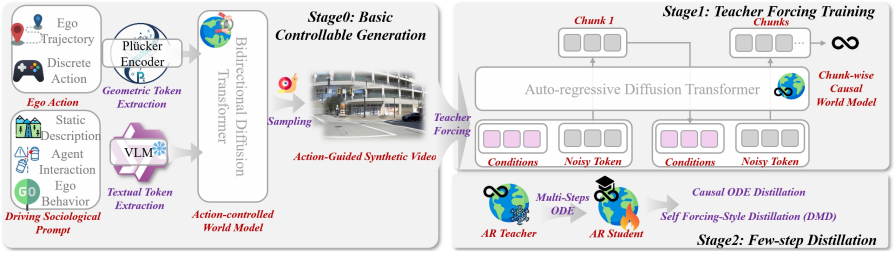

Fig. 2: Architecture and Training Pipeline of CausalDrive. (a) Stage 0: The base model employs Block Causal Attention for $O(1)$ streaming. Ego-trajectories ($\mathcal{P}$) and sociology prompts ($\mathcal{B}$) are injected via AdaLN and cross-attention, respectively. **(b) Stage 1:** To prevent probability flow injectivity violations, we train a continuous flow-matching Autoregressive (AR) teacher using strict Teacher Forcing on ground-truth history ($C_{<k}$). **(c) Stage 2: Self-Corrective DMD.** To mitigate exposure bias during long-horizon rollouts, the student performs an AR self-rollout to generate a flawed memory context ($\hat{C}_{\text{stu}}$). The distillation gradient $\nabla\mathcal{L}_{\text{DMD}}$ is evaluated by forcing the frozen teacher to condition on this flawed context, thereby teaching the student to auto-correct accumulated temporal errors.

(agents). Let $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ denote a visual observation. Instead of passively predicting video continuations, the model must act as a reactive simulator, approximating the transition dynamics $p_\theta(C_{1:K} \mid C_0, \mathcal{A}_{1:K})$, where C_k represents a temporal chunk of latent visual states, and $\mathcal{A}_{1:K}$ denotes a streaming sequence of external control signals.

Transforming a powerful but slow bidirectional video generative model into an ultra-fast autoregressive (AR) interactive simulator presents two fundamental challenges: (1) the *architectural gap*, where distilling a bidirectional teacher into a causal student violates PF-ODE injectivity, causing blurry generation; and (2) the *context mismatch*, where error accumulates over long-horizon autoregressive rollouts (exposure bias). To systematically address these issues, we formulate our method into three parts. We first provide the preliminaries in Sec. 3.1. Then, we introduce the construction of a driving sociology-aware causal autoregressive teacher in Sec. 3.2. Finally, we detail the few-step distillation framework via Causal ODE and Context-Forced DMD in Sec. 3.3.

3.1 Preliminaries

Bidirectional Video Diffusion Models. Recent large-scale video generation heavily relies on continuous-time diffusion models parameterized by Diffusion Transformers (DiTs). Let $\mathbf{z}_0 \sim p_{\text{data}}$ denote a real video chunk in the latent space, and $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote standard Gaussian noise. Under the Flow Matching (FM) framework [26], the PF-ODE governing the marginal distributions is defined as $\mathrm{d}\mathbf{z}_t = v_\theta(\mathbf{z}_t, t)\mathrm{d}t$, for $t \in [0, 1]$. The optimal transport path corresponds to linear interpolation: $\psi_t(\mathbf{z}_0) = t\mathbf{z}_1 + (1-t)\mathbf{z}_0$. The network v_θ is trained to

regress the velocity field:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1,t}\left[\left\|v_\theta(\psi_t(\mathbf{z}_0),t) - (\mathbf{z}_1 - \mathbf{z}_0)\right\|^2\right] \tag{1}$$

Standard video DiTs employ v_θ employs *Bidirectional Attention* across the temporal dimension. When denoising frame $x_t^{(i)}$, the model attends to all other frames, including the future. While excelling at fixed-length generation, this non-causal nature prevents real-time, open-ended interactivity.

Chunk-wise Autoregressive Generation. To enable streaming generation, we factorize the joint distribution of a video into non-overlapping temporal chunks $C_1, C_2, \ldots, C_K$:

$$p_\theta(C_{1:K}) = \prod_{k=1}^{K} p_\theta(C_k \mid C_{<k}, \mathcal{A}_{\leq k}) \tag{2}$$

We adopt a *Block Causal Attention* mask: tokens within the same chunk C_k attend to each other bidirectionally to preserve local spatio-temporal dynamics, while attention to past chunks $C_{<k}$ is strictly causal and cached via a Key-Value (KV) mechanism to ensure $O(1)$ streaming complexity.

3.2 Driving Sociology-Aware Causal Autoregressive Teacher

Directly distilling a bidirectional model into an AR student causes conditional expectation collapse[casual forcing]. Therefore, we first construct a mathematically sound *Causal Autoregressive Teacher*, denoted as $v_{\theta^*}^{\mathrm{AR}}$.

Crucially, driving is not merely following geometric trajectories; it is a highly interactive multi-agent social process. Thus, we condition our AR teacher on a hybrid control signal $\mathcal{A} = (\mathcal{P}, \mathcal{B})$, establishing a causal hierarchy between low-level ego-mechanics and high-level sociology.

Given the ego-vehicle trajectory $\mathcal{P}$, we encode the continuous camera poses into Plücker coordinates and process them via a Multi-Layer Perceptron (MLP) to obtain spatial modulation parameters $\mathbf{h}_{\mathrm{geo}}$. Simultaneously, the discrete behavioral description $\mathcal{B}$ (e.g., *"ego vehicle yields to an aggressive pedestrian"*) is mapped into continuous semantic embeddings $\mathbf{E}_{\mathrm{socio}}$ via a pre-trained text encoder. It is important to note that $\mathcal{B}$ serves as a global "sociological prior" (e.g., a "Polite" scene implies a higher probability of yielding across all agents) rather than a rigid script for specific instance-level control.

Within each DiT block, we explicitly decouple these conditions to model causal interactions. Let $\mathbf{z}$ denote the intermediate latent feature. The geometric action imposes a hard spatial constraint via Adaptive Layer Normalization (AdaLN), while the sociology prompt modulates the multi-agent reaction via Cross-Attention:

$$\hat{\mathbf{z}} = \mathbf{z} + \mathrm{SelfAttn}\big(\mathrm{AdaLN}(\mathbf{z}, \mathbf{h}_{\mathrm{geo}})\big) \tag{3}$$

$$\mathbf{z}' = \hat{\mathbf{z}} + \mathrm{CrossAttn}\big(\mathrm{Q} = \hat{\mathbf{z}}, \mathrm{K,V} = \mathbf{E}_{\mathrm{socio}}\big) \tag{4}$$

This architectural design strictly enforces causality: Eq. 3 guarantees that the visual perspective precisely shifts according to the ego-vehicle's steering and throttle, preventing "action-ignoring." Concurrently, Eq. 4 allows the semantic prior $\mathbf{E}_{\text{socio}}$ to query the spatially-anchored features, hallucinating logical reactions from surrounding agents (e.g., synthesizing a braking neighbor) without interfering with the ego-trajectory's geometric fidelity.

We train this causal AR teacher v_θ^{AR} using **Teacher Forcing**. Specifically, when training the k-th chunk, the attention mechanism is strictly conditioned on the *clean, ground-truth* historical prefix $C_{<k}$, rather than a noisy prefix. This guarantees that the training objective perfectly aligns with the inference distribution, establishing a mathematically sound and dynamically accurate AR teacher.

3.3 Few-Step Distillation

While the AR teacher achieves high fidelity, it requires numerous ODE solver steps (e.g., 50 steps), violating the latency constraints of a real-time interactive world model. We compress the generation to 1-4 steps via a two-stage distillation process.

Causal ODE Distillation. A fundamental requirement for ODE distillation is *frame-level injectivity*: each noisy frame must map to a unique clean frame. Distilling an AR student from a bidirectional teacher violates this, as the absence of future frames causes one noisy state to map to multiple possible clean outcomes, resulting in blurred predictions. To bridge this architectural gap, we perform distillation strictly using trajectories generated by our Causal AR Teacher v_θ^{AR}. Given a clean history $C_{<k}$ and actions $\mathcal{A}_{\leq k}$, we sample exact PF-ODE trajectories from noise to data. The AR student G_ϕ is trained to regress the clean target $C_k^{(0)}$ from intermediate noisy states $C_k^{(t)}$:

$$\phi^* = \arg\min_\phi \mathbb{E}_{C_{<k},t}\left[\left\| G_\phi(C_k^{(t)}, C_{<k}, \mathcal{A}_{\leq k}, t) - C_k^{(0)} \right\|^2\right] \tag{5}$$

Since both the teacher and student are causal, frame-level injectivity perfectly holds, allowing the student to correctly learn the flow map without detail degradation.

Context-Forced DMD for Infinite Horizon. To further reduce sampling to 1-4 steps, we apply Distribution Matching Distillation (DMD). However, in autoregressive long-horizon generation, the student accumulates errors and deviates from the training distribution (*Exposure Bias*). Standard DMD fails here because it creates a *Context Mismatch*: the teacher evaluates gradients based on perfect ground-truth history, while the student generates based on its own flawed history. This phenomenon is akin to the "Covariate Shift" problem in imitation learning.

To resolve this, we propose **Context Forcing**. During training, the student performs an N-chunk *Self-Rollout* to construct a flawed memory context $\hat{C}_{stu}$. When computing the real score using the AR Teacher, we explicitly force the Teacher to condition on this student-generated history $\hat{C}_{stu}$, rather than the ground truth:

$$\nabla_\phi \mathcal{L}_{\text{DMD}} \approx \mathbb{E}\left[s_{\text{real}}\big(\hat{C}_k^{(t)}, t \mid \hat{C}_{stu}, \mathcal{A}\big) - s_{\text{fake}}\big(\hat{C}_k^{(t)}, t \mid \hat{C}_{stu}, \mathcal{A}\big) \right] \frac{\partial \hat{C}}{\partial \phi} \tag{6}$$

By forcing the teacher to "clean up the student's mess," the distillation gradient explicitly teaches the student how to recover from its own errors. Furthermore, to preserve high-frequency road textures (e.g., lane markings) over long simulations, we attach an adversarial discriminator to the fake score network, pushing the few-step generated driving chunks to be indistinguishable from real traffic videos.

3.4 Downstream Applications Formulation

By satisfying real-time ($\geq$ 12 FPS) and reactive requirements, CausalDrive naturally serves as a universal interactive API. We formulate three specific downstream integrations:

Generative Closed-loop Simulation. We formulate CausalDrive as an OpenAI Gym environment. At step k, a planner π_{plan} observes the generated visual state $\mathbf{x}_k$ and outputs geometric action $\mathcal{P}_k$. ForceDrive acts as the transition function: $\mathbf{x}_{k+1} = \text{ForceDrive}(\mathbf{x}_k, \mathcal{P}_k, \mathcal{B}_{\text{prompt}})$. This eliminates the "ghost effect" found in standard log-replay, as neighbors dynamically react based on sociology priors $\mathcal{B}$.

Efficient Policy Post-training via RL. CausalDrive functions as a safe, infinite RL hallucination. We define a continuous MDP where state $s_k = \mathbf{x}_k$ and action $a_k = \mathcal{P}_k$. To provide dense feedback, we append a differentiable **Video-to-Reward (V2R)** module R_ψ that maps generated chunks to scalar rewards: $r_k = R_\psi(\mathbf{x}_{k+1}) = w_1 r_{\text{col}} + w_2 r_{\text{lane}}$. The policy π_ω is optimized to maximize expected returns entirely within the simulated domain.

Human-in-the-Loop Simulation. The sub-100ms latency of our Context-Forced distillation enables human drivers to inject streaming control signals $\mathcal{A}_k$ via hardware (e.g., steering wheels), facilitating real-time counterfactual safety testing and high-quality DAgger (Dataset Aggregation) collection.

4 SocioDrive-Bench: A Benchmark for Driving Sociology

Current autonomous driving datasets (e.g., nuScenes, Waymo) primarily function as open-loop geometric logs. While they provide precise trajectories, they lack explicit annotations of *causal interactions*—the underlying social contracts and reactive negotiations among traffic participants. To train and evaluate the sociology-aware world model proposed in Section 3, we introduce **SocioDrive-Bench**, a large-scale, multimodal benchmark explicitly annotated with driving sociology and causal reactions.

SocioDrive-Bench comprises 20K video clips, which is sourced heterogeneously: 80% of the data is mined from real-world logs (nuPlan [3]) to capture nominal, photorealistic human interactions, while the remaining 20% is generated via the CARLA simulator [8] to inject safety-critical and collision events, thus mitigating the "survival bias" inherent in expert datasets.

4.1 Taxonomy of Causal Interactions

We formalize driving sociology into three distinct behavioral categories. Let $\mathcal{T}_{\text{ego}} = (\mathbf{p}, \mathbf{v}, \mathbf{a}, \theta)$ denote the ego-vehicle state, $\mathcal{T}_{\text{obj}}$ denote the state of surrounding agents, and $\mathcal{M}$ represent the map topology. Rather than relying on simple proximity, we systematically mine scenarios based on strict kinematic triggers to form explicit causal pairs.

Ego-Initiated Interactions (Active Pressure). This category encapsulates scenarios where the environment dynamically reacts to the ego-vehicle's proactive maneuvers, providing crucial supervision for predicting yielding or evasion behaviors. For instance, an *aggressive cut-in* event is triggered when the ego crosses lane boundaries $\mathcal{M}_{\text{lane}}$ while a rear vehicle is within a 15m proximity. A valid interaction is recorded only if the rear vehicle exhibits significant deceleration ($a_{\text{obj}} < -1.5$ m/s^2) or a sudden drop in Time-to-Collision (TTC) within Δt seconds, forming the causal label: `[Ego Cut-in]` $\rightarrow$ `[Forced Rear Car Brake]`. Similarly, *tailgating* is identified when $v_{\text{ego}} > 30$ km/h and the Time Headway (THW) to the lead vehicle remains under 1.0s for at least 3s, leading to a forced yield or acceleration from the leading agent.

Ego-Reactive Interactions (Defensive Driving). Conversely, this category captures environmental anomalies that force the ego-vehicle to alter its state. These scenarios are essential for providing the "negative rewards" required in downstream RL safety training. We identify *hazard responses* when a neighboring vehicle laterally invades the ego's lane, or a Vulnerable Road User (VRU) intersects the future path, necessitating an emergency maneuver ($a_{\text{ego}} < -3$ m/s^2 or excessive lateral jerk). Another common trigger is the *lead car hard brake*, recorded when $a_{\text{lead}} < -3$ m/s^2 is followed by a corresponding deceleration from the ego with a reaction delay $\tau < 1.0$s. These extract explicit defensive causalities, such as `[Neighbor Cut-in]` $\rightarrow$ `[Forced Ego Emergency Brake]`.

Complex Negotiation and Game Theory. The highest density of driving sociology occurs during low-speed, mutual probing where right-of-way is ambiguous. *Unprotected intersections* are prime examples, triggered when the ego's intended path conflicts with oncoming traffic. We detect "creeping" behaviors ($0 < v_{\text{ego}} < 5$ km/h for > 2s), which branch into mutually exclusive outcomes: either the oncoming traffic yields, or the ego halts. Similarly, *narrow passages* restrict dual passage, forcing implicit negotiation until one party yields. These scenarios train the world model to synthesize the nuanced hesitation and game-theoretic resolutions inherent in human driving.

4.2 Automated Annotation Pipeline via Vision-Language Models

Extracting high-level semantic sociology from raw pixels and numeric trajectories is non-trivial. Rule-based scripts are notoriously brittle against the long-tail diversity of the real world. To automate the annotation of millions of frames, we propose a two-stage, Large Vision-Language Model (VLM) powered pipeline.

Phase 1: Clip-Level Micro-Analysis (VLM). We slice the multi-view (Left, Front, Right) video stream into 4-second clips (2Hz sampling). A VLM (e.g., GPT-4o) processes these frames alongside system prompts to output a strictly formatted JSON structure. Crucially, we enforce **Feature Disentanglement** at this stage: the VLM must separately parse `road_structure`, and `vulnerable_road_users` (VRUs). For world modeling, decoupling VRUs from vehicles is essential, as pedestrian kinematics require entirely different attention priors than rigid bodies.

Phase 2: Sequence-Level Macro-Synthesis (LLM). Independent 4-second clip analysis often introduces temporal truncation (e.g., unnatural transitions between states). To resolve this, a Large Language Model (LLM) aggregates the sequential JSON annotations into a modular, macroscopic scene summary. This stage guarantees three functional properties for our world model: *Temporal Smoothing:* The LLM synthesizes an `ego_behavior_narrative`, linguistically smoothing the 2Hz sampling gaps to provide a continuous, chronological condition for the trajectory generation. *Interaction Aggregation:* By extracting a `key_interaction_summary`, we explicitly map the spatial awareness (Left/Front/Right views) to the dynamic agents, forming the precise sociology prior $\mathcal{B}_k$ injected into our cross-attention blocks (Sec. 3). *Queryability and Hard-Case Mining:* The pipeline extracts `critical_events` into structured arrays. This allows us to programmatically query our database for "safety-critical scenarios involving pedestrians," facilitating targeted curriculum generation for RL training without requiring repeated semantic processing.

Through this pipeline, SocioDrive-Bench transforms raw driving logs into a highly structured, sociologically-aware dataset, directly bridging the gap between perception data and cognitive driving simulation.

5 Experiments

Our empirical evaluation is designed to systematically validate CausalDrive across three progressive dimensions: (1) **Simulation Reliability**, ensuring the generated world is visually faithful, controllable, and real-time; (2) **Driving Sociology & Reactivity**, quantifying the model's ability to synthesize causal multi-agent interactions via our proposed benchmark; and (3) **Downstream Applications**, demonstrating its efficacy as a closed-loop evaluation engine and a generative RL post-training environment.

5.1 Experimental Setup

Training Datasets and Curriculum. We employ a progressive, multi-stage data curriculum to train CausalDrive. For foundational pretraining, we utilize

1,700 hours of front-view driving videos from the large-scale OpenDV dataset [43] to learn robust real-world visual priors. Subsequently, we fine-tune the model using the nuPlan dataset [3] processed through our Phase-1 annotation pipeline, alongside the proposed SocioDrive-Bench to explicitly instill driving sociology priors. Crucially, following the insights from ReSim [42], we augment our training corpus with unsafe, safety-critical scenarios synthesized via Bench2Drive [17]. This out-of-distribution (OOD) augmentation is vital for enhancing the model's action-controllability when conditioned on uncommon, non-expert trajectories (e.g., aggressive steering or imminent collisions) typically absent in expert logs.

Implementation Details. Our core flow-matching DiT backbone is initialized from the state-of-the-art Wan2.1-1.3B [34] video generation model. The Stage-0 training is distributed across 8 NVIDIA H20 GPUs with a total batch size of 4. We optimize the network using AdamW with a constant learning rate of 1×10^{-5}. Due to space constraints, the comprehensive hyperparameters for the Stage-1 Causal AR Teacher-Forcing and the Stage-2 Context-Forced DMD algorithms are detailed in the Supplementary Material.

Baselines and Evaluation Metrics. Our evaluation spans three dimensions: visual fidelity (FID, FVD), trajectory controllability (Average/Final Displacement Error: ADE/FDE), and our novel sociology metric—Yielding Compliance Rate (YCR). More details are included in supplementary materials

5.2 Simulation Reliability: Fidelity, Efficiency, and Controllability

Following the evaluation protocols of ReSim [42] and Orbis [29], a foundation world model must strictly adhere to the injected ego-trajectory while maintaining high-fidelity visuals at interactive speeds.

Visual Fidelity and Inference Speed. As shown in Table 1, we achieve an FVD of 121.6, outperforming or matching existing video diffusion models (Vista, Orbis). This proves that our Context-Forced DMD effectively preserves spatial-temporal consistency without blurring. Crucially, CausalDrive achieves an inference throughput of **12.4 FPS** on a single NVIDIA A100 GPU—an order of magnitude faster than baseline diffusion simulators, successfully crossing the threshold required for human-in-the-loop and online RL operations.

Action Controllability. A reactive simulator must not suffer from "posterior collapse" (ignoring control signals). To quantify this, we deploy a pre-trained inverse dynamics model (tracker) to estimate the ego-vehicle's trajectory directly from the generated video pixels, and compute the Average/Final Displacement Errors (ADE/Frechet) against the injected condition trajectory $\mathcal{P}_{1:K}$. Table 1 demonstrates that CausalDrive achieves best metrics, confirming that our continuous flow-matching and AdaLN conditioning precisely translate geometric actions into visual dynamics.

5.3 Evaluating Driving Sociology and Reactivity

Beyond trajectory following, a reactive simulator must synthesize logical multi-agent interactions. We utilize **SocioDrive-Bench** to quantify this reactivity. We

Table 1: Simulation Reliability Comparison. We evaluate visual fidelity (FID, FVD), inference speed (FPS), and Action Controllability (ADE/FDE between the injected trajectory and the trajectory estimated from generated pixels) on nuplan dataset. CausalDrive achieves comparable visual quality and SOTA controllability while running at real-time speeds.

Method	Fidelity ↓	ADE		Frechet		Speed ↑	Real-time
	FVD	Prec. ↑	Rec. ↑	Prec. ↑	Rec. ↑	FPS	
Cosmos [1]	291.80	-	-	-	-	≈ 0.7	✗
Vista [11]	323.37	0.25	0.48	0.39	0.45	≈ 0.3	✗
GEM [12]	291.84	0.27	0.47	0.33	0.54	≈ 0.5	✗
Orbis* [29]	196.21	0.39	0.48	0.46	0.55	≈ 1.2	✗
CausalDrive^{+}	**113.6**	**0.47**	**0.55**	**0.56**	**0.61**	≈ 0.5	✗
CausalDrive	**121.6**	**0.45**	**0.52**	**0.53**	**0.59**	**12.4**	✓

input aggressive ego-trajectories (e.g., forced cut-ins) and measure the *Yielding Compliance Rate* (YCR)—the frequency at which generated neighbors correctly decelerate.

As shown in Table 2, pure action-conditioned predictors like Vista often result in unrealistic collisions ("ghosting"), yielding only 12.5% of the time. In contrast, by conditioning on the semantic prompt $\mathcal{B}$ (*"Polite"*), CausalDrive explicitly models the causal reaction, achieving an **82.0% YCR** and reducing the False-Collision Rate (FCR) to **18.0%**. This confirms our model's ability to synthesize diverse sociological behaviors controlled by language prompts.

Table 2: Reactivity Evaluation on SocioDrive-Bench. We measure the Yielding Compliance Rate (YCR) and False-Collision Rate (FCR) during aggressive ego-maneuvers. CausalDriveeliminates the ghost effect and synthesizes highly reactive NPC behaviors.

Method	Sociology Control	Yielding Rate (YCR) ↑	False-Collision (FCR) ↓
Log-Replay(GT)	✗	0.0%	100.0% (Ghosting)
Vista [11]	✗	12.5%	87.5%
CausalDrive	✓(*"Polite"*)	**82.0%**	**18.0%**

5.4 Ablations

Table 3 isolates each design choice. Here are some findings: (1) Removing the sociology prompt drops YCR by ∼37pp, confirming its critical role in steering reactive behavior. (2) The Causal AR Teacher is the most impactful single component (YCR drops to 38.6% without it). (3) Context-Forced DMD (SCF) contributes ∼10pp YCR gain and substantially improves long-horizon stability. (4) N=4 is the sweet spot: N=8 gives marginal gains at higher training cost.

Table 3: Component ablation on SocioDrive-Bench (Polite). SCF = Self-Corrective Forcing (Context-Forced DMD). N = rollout chunks.

Variant	**FVD↓**	**YCR↑**
Full CausalDrive	121.6	82.0%
w/o Sociology Prompt	128.4	45.2%
w/o Causal AR Teacher	157.3	38.6%
w/o SCF (standard DMD)	142.1	72.5%
1 Taxonomy only	133.7	58.3%
2 Taxonomies	126.2	73.1%
3 Taxonomies (full)	121.6	82.0%
N=1 (no self-rollout)	139.8	68.4%
N=2	128.5	77.1%
N=4 (default)	121.6	82.0%
N=8	120.9	82.3%

5.5 Downstream Applications

Generative Closed-Loop Simulation. To prove CausalDriveis a reliable proving ground, we evaluate planners (e.g., UniAD [15]) inside our simulator. As shown in Table 4, evaluating under the PDM-Closed protocol reveals that planners tested in CausalDrive experience significantly fewer "ghost" collisions and achieve higher success rates compared to Log-Replay or other generative baselines, demonstrating our fidelity and reactive realism.

Table 4: Generative Closed-Loop Evaluation. Planner performance (e.g., UniAD) evaluated under the PDM-Closed protocol. CausalDrive provides a highly realistic closed-loop testbed, significantly reducing unrealistic "ghost" collisions compared to Log-Replay.

Simulation Env.	**PDMS ↑**	**RC ↑**	**ADS ↓**
DriveArena [44]	0.6901	0.0641	0.0508
DrivingSphere [40]	0.7281	0.1170	0.0851
CausalDrive(Ours)	**0.7792**	**0.1752**	**0.1365**

Efficient RL Post-Training. Finally, we validate CausalDrive as a generative data engine for policy improvement. We benchmark our RL-post-trained agent against SOTA learning-based planners (e.g., UniAD, DiffusionDrive) on the Navsim evaluation protocol. As shown in Table 5, our agent achieves state-of-the-art performance with a **PDMS score of 90.7**. More importantly, the agent trained in CausalDrive achieves a perfect **Comfort score of 100** and the highest **Time-to-Collision (TTC) of 94.7**. The ability to safely experience

"counterfactual crashes" inside CausalDrive allows the policy to discover emergent defensive driving strategies, leading to superior safety metrics compared to models trained purely on static expert datasets.

Table 5: RL Post-Training Performance on Navsim. Learning from counterfactual interactions in our impartial environment using DiffusionDrive [25] leads to superior safety (lower collision rates) and success rates.

Policy Method	**NC ↑**	**DAC ↑**	**TTC ↑**	**Comf. ↑**	**EP ↑**	**PDMS ↑**
UniAD [15]	97.8	91.9	92.9	100	78.8	83.4
DriveDPO [31]	98.5	98.1	94.8	99.9	84.3	90.0
DiffusionDrive [25] [25]	98.2	96.2	94.7	100	82.2	88.1
AD-R1 [39]	98.5	97.5	94.6	100	84.8	89.8
DiffusionDrive-v2 [25]	98.4	98.3	94.6	100	85.0	90.3
RL in CausalDrive	98.7	98.2	94.7	100	85.1	90.7

6 Conclusion

In this paper, we presented **CausalDrive**, a real-time foundation driving world renderer that transforms passive video generation into a reactive neural simulator. By omitting "oracle" future layouts and leveraging our VLM-curated **SocioDrive-Bench**, we compel the model to predict causal multi-agent interactions strictly from the initial frame, ego-trajectory, and semantic prompts. To overcome the latency and exposure bias of autoregressive generation, we proposed a novel **Context-Forced DMD** framework. This distillation strategy explicitly corrects temporal compounding errors, compressing inference to ≤ 4 steps and achieving ≥ 12 FPS. Extensive experiments demonstrate that CausalDrive effectively eliminates the "ghost effect" in closed-loop evaluations and serves as a highly scalable generative environment for RL post-training, yielding policies with superior defensive driving capabilities. Extending this monocular paradigm to multi-camera surround-view systems remains our future work.

Acknowledgements

We use publicly available datasets, models ,and code resources, including nuScenes, Bench2Drive, nuplan and other public research resources. We confirm that all such resources are used solely for academic research purposes and not for any commercial activity, in accordance with their respective licenses and terms of use.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810 (2021)
4. Chen, S., Jiang, B., Gao, H., Liao, B., Xu, Q., Zhang, Q., Huang, C., Liu, W., Wang, X.: Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. arXiv preprint arXiv:2402.13243 (2024)
5. Chi, H., Qiu, D., Su, H., Liu, H., Li, Z., Zhang, H., Lv, C.: Driver-wm: A driver-centric traffic-conditioned latent world model for in-cabin dynamics rollout. arXiv preprint arXiv:2605.05092 (2026)
6. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., et al.: Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. Advances in Neural Information Processing Systems **37**, 28706–28719 (2024)
7. Deng, T., Chen, X., Chen, Y., Chen, Q., Xu, Y., Yang, L., Xu, L., Zhang, Y., Zhang, B., Huang, W., Wang, H.: Gaussiandwm: 3d gaussian driving world model for unified scene understanding and multi-modal generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10656–10667 (June 2026)
8. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
9. Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrivedit: High-resolution long video generation for autonomous driving with adaptive control. arXiv preprint arXiv:2411.13807 (2024)
10. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601 (2023)
11. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems **37**, 91560–91596 (2024)
12. Hassan, M., Stapf, S., Rahimi, A., Rezende, P., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., et al.: Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22404–22415 (2025)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
15. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)
16. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
17. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. Advances in Neural Information Processing Systems **37**, 819–844 (2024)
18. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
19. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. arXiv preprint arXiv:2412.05435 (2024)
20. Li, Q., Peng, Z., Feng, L., Zhang, Q., Xue, Z., Zhou, B.: Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. IEEE transactions on pattern analysis and machine intelligence **45**(3), 3461–3475 (2022)
21. Li, R., Li, X., Xie, M., Shan, H., Qiu, S., Chang, X., Fan, Y., Xiong, F., Jiang, H., Ren, Y., Yu, H., Xu, M., Long, Y., Ojha, V., Cui, Z.: Amap: Distilling future priors for ahead-aware online hd map construction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24906–24917 (June 2026)
22. Li, Y., Xiong, K., Guo, X., Li, F., Yan, S., Xu, G., Zhou, L., Chen, L., Sun, H., Wang, B., et al.: Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. arXiv preprint arXiv:2506.08052 (2025)
23. Liang, A., Kong, L., Yan, T., Liu, H., Yang, W., Huang, Z., Yin, W., Zuo, J., Hu, Y., Zhu, D., et al.: Worldlens: Full-spectrum evaluations of driving world models in real world. arXiv preprint arXiv:2512.10958 (2025)
24. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. Advances in Neural Information Processing Systems **35**, 10421–10434 (2022)
25. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., et al.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12037–12047 (2025)
26. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
27. Lu, G., Guo, W., Zhang, C., Zhou, Y., Jiang, H., Gao, Z., Tang, Y., Wang, Z.: Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. arXiv preprint arXiv:2505.18719 (2025)
28. Mai, J., Liao, B., Zhao, Z., Zeng, Y., Li, H., Civera, J., Wu, T., Zhou, Y., Liu, P.: Neural predictor-corrector: Solving homotopy problems with reinforcement learning. arXiv preprint arXiv:2602.03086 (2026)
29. Mousakhan, A., Mittal, S., Galesso, S., Farid, K., Brox, T.: Orbis: Overcoming challenges of long-horizon prediction in driving world models. arXiv preprint arXiv:2507.13162 (2025)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)

31. Shang, S., Chen, Y., Wang, Y., Li, Y., Zhang, Z.: Drivedpo: Policy learning via safety dpo for end-to-end autonomous driving. arXiv preprint arXiv:2509.17940 (2025)
32. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
33. Tian, H., Li, T., Liu, H., Yang, J., Qiu, Y., Li, G., Wang, J., Gao, Y., Zhang, Z., Wang, L., et al.: Simscale: Learning to drive via real-world simulation at scale. arXiv preprint arXiv:2511.23369 (2025)
34. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European Conference on Computer Vision. pp. 55–72. Springer (2024)
36. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: 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. pp. 14749–14759 (2024)
37. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6902–6912 (2024)
38. Yan, S., Shi, P., Zhao, Z., Wang, K., Cao, K., Wu, J., Li, J.: Turboreg: Turboclique for robust and efficient point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26371–26381 (2025)
39. Yan, T., Tang, T., Gui, X., Li, Y., Zhesng, J., Huang, W., Kong, L., Han, W., Zhou, X., Zhang, X., et al.: Ad-r1: Closed-loop reinforcement learning for end-to-end autonomous driving with impartial world models. arXiv preprint arXiv:2511.20325 (2025)
40. Yan, T., Wu, D., Han, W., Jiang, J., Zhou, X., Zhan, K., Xu, C.z., Shen, J.: Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. arXiv preprint arXiv:2411.11252 (2024)
41. Yan, T., Yin, J., Lang, X., Yang, R., Xu, C.Z., Shen, J.: Olidm: Object-aware lidar diffusion models for autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9121–9129 (2025)
42. Yang, J., Chitta, K., Gao, S., Chen, L., Shao, Y., Jia, X., Li, H., Geiger, A., Yue, X., Chen, L.: Resim: Reliable world simulation for autonomous driving. arXiv preprint arXiv:2506.09981 (2025)
43. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., et al.: Generalized predictive model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14662–14672 (2024)
44. Yang, X., Wen, L., Ma, Y., Mei, J., Li, X., Wei, T., Lei, W., Fu, D., Cai, P., Dou, M., et al.: Drivearena: A closed-loop generative simulation platform for autonomous driving. arXiv preprint arXiv:2408.00415 (2024)
45. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
46. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation.

In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10412–10420 (2025)
47. Zhao, Z., Yang, H., Liao, B., Zeng, Y., Yan, S., Gu, Y., Liu, P., Zhou, Y., Li, H., Civera, J.: Advances in global solvers for 3d vision. arXiv preprint arXiv:2602.14662 (2026)
48. Zhu, H., Zhao, M., He, G., Su, H., Li, C., Zhu, J.: Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. arXiv preprint arXiv:2602.02214 (2026)