Driver-WM: A Driver-Centric Traffic-Conditioned Latent World Model for In-Cabin Dynamics Rollout

Haozhuang Chi[1], Daosheng Qiu[2], Hao Su[3], Haochen Liu[1], Zirui Li[4], Haoruo Zhang[1], and Chen Lv[1]✉⋆

[1] Nanyang Technological University, Singapore
[2] Hubei University, Wuhan, China
[3] Osaka University, Osaka, Japan
Correspondence: lyuchen@ntu.edu.sg

Abstract. Safe L2/L3 driving automation requires anticipating human-in-the-loop reactions during shared-control transitions. While most driving world models forecast the external environment, in-cabin intelligence remains strictly recognition-oriented and lacks multi-step rollout capabilities for driver dynamics. We introduce **Driver-WM**, a driver-centric latent world model that rolls out in-cabin dynamics causally conditioned on out-cabin traffic context. This formulation unifies physical kinematics forecasting with auxiliary behavioral and emotional semantic recognition. Operating in a compact latent space constructed from frozen vision-language features, Driver-WM adopts a dual-stream architecture to separately encode external traffic and internal driver states. These streams are directionally coupled via a gated causal injection mechanism, which uses a learned vector gate to modulate external contextual perturbations while strictly enforcing temporal causality. Experiments on AIDE show robust long-horizon forecasting on reactive high-motion clips, improved driver/traffic semantic alignment, and controlled interventions that expose the external-to-internal mechanism.

Keywords: World model · Driver monitor system · Vision-language model · Human-in-the-loop driving automation · Autonomous driving

1 Introduction

Autonomous driving has evolved from isolated driver assistance functions toward human-centered intelligent systems [5, 6, 53]. Currently, the majority of real-world driving automation systems are deployed at the SAE L2/L3 level [36], where driving responsibility continuously shifts between the autonomous policy and human supervision in a shared-control (mixed-autonomy) setting [38, 45].

⋆ Corresponding author.

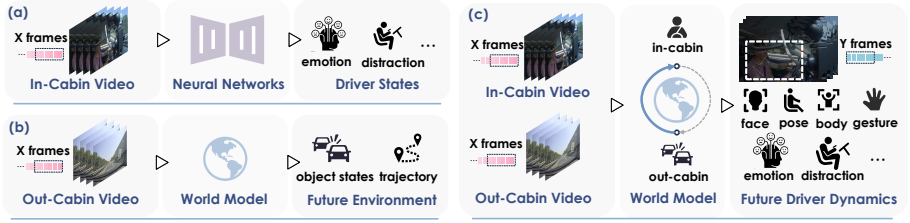

Fig. 1: The comparison of three paradigms: (a) Regular driver monitoring systems (DMS) for driver-state recognition. (b) Standard world models for future environment forecasting. (c) **Driver-WM (ours)** that performs multi-step rollout of internal driver dynamics explicitly conditioned on synchronized external traffic observations.

Although recent advances in end-to-end systems [16, 25] and vision-language-action (VLA) paradigms [12, 18, 64] have notably improved the understanding and reasoning capabilities of perception and planning [16, 20, 26, 29], the ultimate safety of current system domains still requires the human in the loop. In practice, many safety-critical failures are associated with inadequate takeover readiness rather than incorrect scene understanding [43], and the risk increases when the system operates beyond its functional domain [4] or under evolving traffic interactions [41, 56].

Motivated by the need to reason about long-horizon safety and interaction, world models have recently emerged as a principled framework for autonomous driving [22]. By predicting how the external environment evolves conditioned on current observations and actions, they have been widely adopted for forward simulation, maneuver reasoning, and policy training, ranging from unified full-stack driving systems to generative traffic simulators and driving foundation models [10, 13, 15, 17, 42, 47, 50]. To support efficient long-horizon prediction and scalable closed-loop rollout, latent world models have emerged that operate dynamic prediction in compressed latent space [25, 60]. However, existing world models are primarily environment-oriented (Fig. 1b): they model how roads, surrounding agents, and vehicles evolve, while the driver is usually treated as a post-hoc source of risk. They follow an environment-to-action paradigm and do not model how external driving events drive the driver's internal evolution, leaving posture, gaze, motor readiness, and reaction behavior outside the rollout loop.

In parallel, in-cabin modules, including DMS [8, 33, 51] (Fig. 1a) and emerging LLM-based cockpit assistants [37], remain recognition- and interaction-oriented rather than predictive. DMS primarily focuses on recognizing instantaneous states such as distraction or fatigue from short temporal windows [19, 32, 40], while LLM-based systems emphasize dialogue, intent understanding, and high-level reasoning [27, 30, 37]. Although effective for detection and interaction, both paradigms lack the capability to predict how the driver's physical and cognitive state will evolve over time in response to changing traffic conditions, treating the driver as an observed entity rather than a dynamical system. As

a result, in-cabin intelligence remains limited to recognition-based alerts and cannot support anticipatory safety reasoning or systematic analysis of how traffic context modulates driver responses over a prediction horizon. This is critical for L2/L3 driving automation, where safe design requires anticipating not only environment evolution but also driver responses seconds ahead and under hypothetical interventions. A world model that ignores the driver's internal dynamics is therefore incomplete for shared-control operation.

To address these issues, we propose **Driver-WM**, a driver-centric latent world model that rolls out future driver dynamics conditioned on synchronized in-cabin and out-cabin observations. In this context, *dynamics* denotes the temporal evolution of the driver's internal state in a compact latent space. This underlying evolution supports both physically grounded *kinematics* (decoded as future skeleton trajectories) and semantic factors (e.g., driver behavior and emotion, predicted as auxiliary regularizers). Instead of generating pixels, Driver-WM forecasts structured 2D skeleton keypoints of the driver's pose and motion [14, 63], providing an efficient target that is stable for long-horizon prediction and naturally aligned with in-cabin supervision. Driver-WM operates in a compact latent space, where frozen vision–language model (VLM) [3, 23] features serve as a perceptual encoder to compress high-dimensional visual inputs, and a lightweight world-model core learns latent dynamics with a past/future rollout protocol initialized by observed latents. It adopts a dual-stream latent architecture to separately represent traffic context and driver state, and couples them through a gated interaction module that controls how external driving events drive internal dynamics. This formulation explicitly models how the environment shapes human reactions, enabling both realistic driver response forecasting and controlled test-time interventions for mechanism analysis.

In summary, our contributions are threefold:

- **Driver-centric latent world model for dynamics rollout:** multi-step forecasting of in-cabin driver dynamics, unifying kinematic trajectories and auxiliary semantic factors, causally conditioned on observed traffic context.
- **Directionally coupled dual-stream dynamics with gated injection:** explicit external-to-internal coupling that supports controllable conditioning and controlled intervention analysis.
- **Foundation-derived state interface with unified decoding:** frozen VLM features serve as a compact perceptual interface, supporting geometric rollout with auxiliary semantic regularization.

2 Related Work

2.1 World Models in Autonomous Driving

Most driving world models formulate future prediction as conditional video generation in pixel space. Examples include GAIA-1 [15], DriveDreamer [42, 58], Vista [13] for controllable rollouts, DriveDreamer4D [57] for 4D scene generation,

UniDrive-WM, UniWM, and DriveVA [10, 31, 47] for unified planning and generation, and CausalDrive [50] for interactive causal simulation. While pixel-space rollouts offer high visual fidelity, they often entangle semantics, geometry, and physics with low-level rendering, and their iterative diffusion process suffers from inherent latency and extrapolation instability [59]. Recent efforts, such as MAD [35], attempt to explicitly decouple motion and appearance [7] to improve generation efficiency. However, these approaches remain strictly focused on the external environment; enforcing rigorous structural constraints [24] remains challenging in a pixel-generation formulation. Latent-state world models address this limitation by predicting and planning in a compact state space. In autonomous driving, LAW [25] regularizes end-to-end driving via temporal consistency, while World4Drive [60] constructs an intention-aware latent model for complex interactions. Yet, whether employing motion-decoupled video generators or latent-state planners, existing methods predominantly model the evolution of the external scene, leaving the human driver unmodeled in the predictive state. Driver-WM diverges from this scene-centric paradigm by introducing a driver-centric latent dynamics model causally conditioned on out-cabin traffic context. Instead of raw video synthesis, we utilize skeleton trajectory rollout as a physically grounded target, decoupling human kinematics from visual appearance [48] and enabling explicit analysis of how external driving events shape predicted driver behavior.

2.2 Driver and Cabin State Modeling

In-cabin perception is commonly studied as discrete state recognition, aiming to infer discrete driver states or activities, such as distraction or drowsiness [46, 52], from visual observations [8, 40]. Drive&Act [33] provides a benchmark for fine-grained driver activity recognition with multi-modal streams, and AIDE [51] further links in-cabin observations with synchronized out-cabin traffic context. Recent unified multi-task frameworks have reported strong semantic recognition results on this benchmark; for example, MMTL-UniAD [32]. Driver-related prediction has also been explored in maneuver/intention anticipation, e.g., Brain4Cars [19]. These works primarily focus on predicting discrete categorical labels. As such, they do not explicitly model continuous driver motion responses to external driving events (e.g., ego maneuvers such as braking or turning), which are important for predictive safety assessment. Human motion forecasting predicts future skeletal trajectories from historical motion. SiMLPe [14] shows that compact architectures can capture strong kinematic priors, while MotionBERT [63] demonstrates the benefit of transferable skeleton representations via large-scale pretraining. Recent studies further incorporate context and interactions: TRiPOD [1] studies pose forecasting in the wild, and Waymo-3DSkelMo [62] models pedestrian skeletal reactions under traffic interactions. Additionally, explicitly predicting human kinematics, such as hand trajectories, has recently proven highly effective for conditioning future visual and action forecasting in egocentric scenarios [55]. In contrast, driver motion inside the cabin is still often treated as isolated estimation without explicit conditioning on the evolving traffic context. To bridge the gap between classification-based driver monitoring and motion forecasting

without explicit conditioning on out-cabin traffic context, Driver-WM explicitly parameterizes the directional coupling from out-of-cabin driving events to driver motion, formulating traffic-conditioned driver responses as latent dynamics and decoding them into continuous skeleton rollouts.

2.3 Foundation Models for Embodied Perception

Vision–language foundation models (VLMs) provide open-vocabulary and visual-knowledge representations [21] and are widely used as perception backbones; instruction-tuned families such as Qwen-VL [2, 3, 39] further improve semantic grounding. Many works adopt a frozen-backbone paradigm [23], keeping large encoders fixed while learning lightweight interfaces for downstream adaptation. In embodied AI and autonomous driving, VLM/LLM-based agents are frequently used for high-level reasoning and planning [28, 54], including general embodied systems such as PaLM-E [12], RT-2 [64], and FARE [27], language-conditioned navigation models [11], as well as driving-oriented VLMs like DriveLM [37]. Rather than utilizing VLMs to produce discrete symbolic outputs, Driver-WM treats frozen Qwen3-VL features as a continuous state interface for latent dynamics, ensuring that the rollout relies on grounded visual perception rather than language priors [44, 61]. This formulation enables efficient multi-step physical rollouts that remain semantically consistent with the visual scene, bypassing the need for pixel-level generation.

3 Method

3.1 Problem Formulation

We formulate driver-centric cabin world modeling as externally conditioned temporal forecasting. We use in-/out-cabin to denote temporally synchronized raw observations, and internal/external to denote their derived latent states. Given these streams, we learn a world model to roll out the driver's internal dynamics conditioned on the observed external context, and decode the resulting states into physically plausible skeleton trajectories. In contrast to conventional driver monitoring that focuses on per-frame recognition, our primary objective is multi-step rollout; semantic predictions are used only as auxiliary regularizers. Throughout this paper, *dynamics* refers to the latent-state evolution governing the driver's future responses, whereas *kinematics* refers specifically to the decoded geometric skeleton trajectories. Semantic predictions provide clip-level regularization and are not used as temporal conditioning inputs.

Let $\mathbf{o}^{\text{in}}_{1:T} = \{\mathbf{o}^{\text{in}}_t\}_{t=1}^{T}$ and $\mathbf{o}^{\text{out}}_{1:T} = \{\mathbf{o}^{\text{out}}_t\}_{t=1}^{T}$ denote synchronized in-cabin and out-cabin observations over horizon T. Driver-WM treats $\mathbf{o}^{\text{out}}$ as an external context that conditions the driver's internal evolution.

Driver-WM adopts a past/future protocol with $T = T_{\text{obs}} + T_{\text{pred}}$ with causal rollouts. Both streams are initialized with ground-truth latents for $t \leq T_{\text{obs}}$. For $t > T_{\text{obs}}$, Driver-WM performs closed-loop rollouts of external and internal

latents, while enforcing that the update at step $t{+}1$ only accesses histories up to step t. Crucially, the external rollout serves exclusively to sustain continuous environmental conditioning for the driver, rather than to reconstruct future traffic scenes explicitly.

Dual-stream latent states are formulated in Driver-WM. Specifically, we represent the driver state at time t as an internal latent $\mathbf{z}_t^{\text{int}} \in \mathbb{R}^D$ and a set of multi-view external latents $\mathbf{Z}_t^{\text{ext}} = \{\mathbf{z}_{t,v}^{\text{ext}} \in \mathbb{R}^D\}_{v=1}^V$, where V is the number of out-cabin views (e.g., front/left/right) and $D{=}2048$ is the hidden dimension of the frozen Qwen3-VL vision encoder. Accordingly, the external history is $\mathbf{Z}_{\le t}^{\text{ext}} = \{\mathbf{z}_{\tau,v}^{\text{ext}}\}_{\tau \le t,\, v}$. We further define a pooled external vector $\bar{\mathbf{z}}_t^{\text{ext}} = \frac{1}{V}\sum_{v=1}^V \mathbf{z}_{t,v}^{\text{ext}}$ and its history $\bar{\mathbf{Z}}_{\le t}^{\text{ext}} = \{\bar{\mathbf{z}}_\tau^{\text{ext}}\}_{\tau \le t}$, which is used by the dynamics core.

During the rollout stage, the external context of Driver-WM is advanced by $\hat{\mathbf{z}}_{t+1}^{\text{ext}} = f_\theta^{\text{ext}}(\hat{\mathbf{z}}_t^{\text{ext}})$, and the internal state is updated by an externally conditioned transition:

$$\hat{\mathbf{z}}_{t+1}^{\text{int}} = \mathcal{F}_\theta\Big(\hat{\mathbf{Z}}_{\le t}^{\text{int}},\ \hat{\bar{\mathbf{Z}}}_{\le t}^{\text{ext}}\Big). \tag{1}$$

We initialize $\hat{\mathbf{z}}_{1:T_{\text{obs}}}^{\text{int}} = \mathbf{z}_{1:T_{\text{obs}}}^{\text{int}}$ and $\hat{\bar{\mathbf{z}}}_{1:T_{\text{obs}}}^{\text{ext}} = \bar{\mathbf{z}}_{1:T_{\text{obs}}}^{\text{ext}}$, and compute rollout losses only on the future window. Driver-WM formulates multi-task decoding with auxiliary regularizers. A geometric head decodes rolled-out internal latents into skeleton keypoints:

$$\hat{\mathbf{s}}_t = D_{\text{skel}}\big(\hat{\mathbf{z}}_t^{\text{int}}\big), \quad \hat{\mathbf{s}}_t \in \mathbb{R}^{K\times 2} \text{ (or } \mathbb{R}^{K\times 3}), \tag{2}$$

where K is the number of keypoints (e.g., $K{=}136$ for HALPE-style skeleton). We also attach lightweight auxiliary heads on $\mathbf{z}^{\text{int}}$ and $\mathbf{z}^{\text{ext}}$ to regularize the latent space with semantic factors.

3.2 VLM-based Perception and Latent Interface

We model temporal dynamics in a compact semantic space rather than raw pixels. We adopt **Qwen3-VL** [3] as a frozen perceptual backbone and use its representation as the latent interface, following fixed-encoder visual-interface adaptation paradigms (e.g., BLIP-2 [23] and training-free VLM adaptation [9]).

Frozen features. Let $E_{\text{vlm}}(\cdot)$ denote the visual perception stack of Qwen3-VL. Given synchronized observations at time t, we extract $\mathbf{f}_t^{\text{in}} = E_{\text{vlm}}(\mathbf{o}_t^{\text{in}})$ and $\mathbf{f}_{t,v}^{\text{out}} = E_{\text{vlm}}(\mathbf{o}_{t,v}^{\text{out}})$, where $v \in \{1,\dots,V\}$ indexes the out-cabin views and $D{=}2048$. We pre-extract and cache $\{\mathbf{f}_t^{\text{in}}, \mathbf{f}_t^{\text{out}}\}_{t=1}^T$ and exclude the VLM from the training graph to reduce computation and stabilize optimization, so that learning focuses on the world-model core for driver dynamics.

View-conditioned interface. To disambiguate camera sources in a shared semantic space, we inject a learnable view embedding

$$\bar{\mathbf{f}}_t^{\text{in}} = \mathbf{f}_t^{\text{in}} + \mathbf{e}_{\text{view}}(\text{in}), \qquad \bar{\mathbf{f}}_{t,v}^{\text{out}} = \mathbf{f}_{t,v}^{\text{out}} + \mathbf{e}_{\text{view}}(v), \qquad \mathbf{e}_{\text{view}}(\cdot) \in \mathbb{R}^D, \tag{3}$$

where v is the view identifier (e.g., $\{\text{front}, \text{left}, \text{right}, \text{in-cabin}\}$). We obtain $\bar{\mathbf{f}}_t^{\text{in}}$ and $\bar{\mathbf{f}}_{t,v}^{\text{out}}$ accordingly.

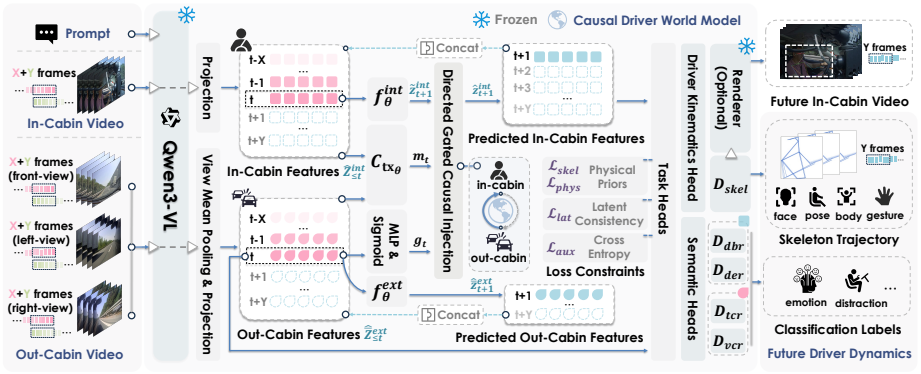

Fig. 2: Overall Architecture of Driver-WM. From synchronized in/out-cabin videos, a frozen Qwen3-VL extracts dual-stream latent features. Pooled external history $\hat{\mathbf{Z}}^{\text{ext}}_{\leq t}$ perturbs the internal transition via a directed Gated Causal Injection with a vector gate $\mathbf{g}_t$, yielding an updated internal latent $\hat{\mathbf{z}}^{\text{int}}_{t+1}$. Internal latents are autoregressively rolled out to forecast future states, decoded into skeleton trajectories and auxiliary semantic predictions (e.g., driver behavior) for regularization.

Latent State Interface. By default, we use an identity latent interface:

$$\mathbf{z}^{\text{int}}_t = \bar{\mathbf{f}}^{\text{in}}_t, \qquad \mathbf{z}^{\text{ext}}_{t,v} = \bar{\mathbf{f}}^{\text{out}}_{t,v}. \tag{4}$$

We apply mean pooling across views to obtain $\bar{\mathbf{z}}^{\text{ext}}_t = \frac{1}{V}\sum_{v=1}^{V} \mathbf{z}^{\text{ext}}_{t,v}$.

3.3 Causal Driver World Model

We formulate the world model as a lightweight gated module, in contrast to heavy block-based architectures. Given dual-stream latents $\{(\mathbf{z}^{\text{int}}_t, \bar{\mathbf{z}}^{\text{ext}}_t)\}_{t=1}^{T}$ (Sec. 3.1–3.2), we learn a *directed* world model that rolls out the driver's internal dynamics conditioned on external context. We enforce two inductive biases: **(i) temporal causality** (the update at $t{+}1$ only depends on context up to t) and **(ii) directed coupling** (external $\rightarrow$ internal via an explicit gate rather than symmetric fusion). In our main configuration, we apply a causal temporal self-attention pre-encoding to each stream before rollout; the bidirectional variant is used only as a non-causal upper bound.

Causal context summary. Let $\hat{\mathbf{Z}}^{\text{int}}_{\leq t}$ and $\hat{\mathbf{Z}}^{\text{ext}}_{\leq t}$ denote the internal/external histories available up to step t. We compute an external-to-internal context summary $\mathbf{m}_t \in \mathbb{R}^D$ via cross-attention:

$$\mathbf{m}_t = \text{Ctx}_\theta\left(\hat{\mathbf{Z}}^{\text{int}}_{\leq t}, \hat{\mathbf{Z}}^{\text{ext}}_{\leq t}\right). \tag{5}$$

Temporal causality is enforced by truncating the histories passed to Ctx_θ at each rollout step (i.e., only providing representations $\leq t$).

Internal transition. We predict a candidate next internal latent via a lightweight transition predictor f_θ^{int}:

$$\tilde{\mathbf{z}}_{t+1}^{\text{int}} = f_\theta^{\text{int}}\big(\hat{\mathbf{z}}_t^{\text{int}}\big)\,. \tag{6}$$

In practice, we parameterize a diagonal Gaussian transition during training and use the mean prediction for deterministic inference; KL regularization is enabled only in the probabilistic ablation (Supplementary).

Gated causal coupling. External driving events do not affect the driver equally; we introduce a vector-valued gate $\mathbf{g}_t \in (0,1)^D$ to regulate contextual perturbation:

$$\mathbf{g}_t = \sigma\big(\text{MLP}_g\big(\hat{\tilde{\mathbf{z}}}_t^{\text{ext}}\big)\big)\,, \qquad \hat{\mathbf{z}}_{t+1}^{\text{int}} = (\mathbf{1}-\mathbf{g}_t)\odot\tilde{\mathbf{z}}_{t+1}^{\text{int}} \;+\; \mathbf{g}_t\odot\mathbf{m}_t. \tag{7}$$

For controlled interventions analysis, we bypass the learned gate with a scalar override λ_{CA}, $\hat{\mathbf{z}}_{t+1}^{\text{int}} = (1-\lambda_{\text{CA}})\tilde{\mathbf{z}}_{t+1}^{\text{int}} + \lambda_{\text{CA}}\mathbf{m}_t$.

3.4 Unified Decoding and Optimization Objectives

Driver-WM rolls out internal latents $\hat{\mathbf{z}}_{1:T}^{\text{int}}$ and uses them as a shared substrate for (i) skeleton rollout (primary) and (ii) auxiliary semantic predictions (regularizers). We adopt a lightweight *one core, multiple heads* design to keep the temporal core compact.

Geometric decoding. We decode the rolled-out internal latents into skeleton keypoints via a Spatial-Temporal Graph Convolutional Network (ST-GCN):

$$\hat{\mathbf{s}}_t = D_{\text{skel}}\big(\hat{\mathbf{z}}_t^{\text{int}}\big)\,. \tag{8}$$

For 2D rollout, $\hat{\mathbf{s}}_t \in [0,1]^{K\times 2}$ stores normalized joint coordinates (sigmoid output), where K is the number of joints (e.g., K=136 for HALPE). D_{skel} uses human skeletal adjacency as the graph topology. We supervise $\{\hat{\mathbf{s}}_t\}$ with regression loss $\mathcal{L}_{\text{skel}}$ over the rollout horizon; model selection is driven by multi-step skeleton accuracy (e.g., MPJPE). These are image-plane quantities; ROI-/vehicle-frame use requires calibrated cabin references and camera calibration, while metric 3D can use priors, monocular lifting, or depth/multi-view sensors.

Physical priors for plausible rollouts. Beyond pointwise regression, we regularize decoded trajectories with

$$\mathcal{L}_{\text{phys}} = \lambda_{\text{bone}}\mathcal{L}_{\text{bone}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{seat}}\mathcal{L}_{\text{seat}}, \tag{9}$$

where $\mathcal{L}_{\text{bone}}$ preserves kinematic edge lengths, $\mathcal{L}_{\text{smooth}}$ suppresses temporal jitter, and $\mathcal{L}_{\text{seat}}$ is an optional cabin/seat feasibility term. For the kinematic tree $\mathcal{E}$, we use

$$\mathcal{L}_{\text{bone}} = \sum_t \sum_{(i,j)\in\mathcal{E}} \Big|\|\hat{\mathbf{s}}_{t,i}-\hat{\mathbf{s}}_{t,j}\|_2 - \|\mathbf{s}_{t,i}-\mathbf{s}_{t,j}\|_2\Big|\,. \tag{10}$$

Temporal smoothness is enforced by matching first-order motion and penalizing second-order jitter:

$$\mathcal{L}_{\text{smooth}} = \sum_t \|\Delta\hat{\mathbf{s}}_t - \Delta\mathbf{s}_t\|_2^2 + \sum_t \big\|\Delta^2\hat{\mathbf{s}}_t\big\|_2^2\,, \tag{11}$$

where $\Delta\hat{\mathbf{s}}_t = \hat{\mathbf{s}}_{t+1} - \hat{\mathbf{s}}_t$ and $\Delta^2\hat{\mathbf{s}}_t = \hat{\mathbf{s}}_{t+2} - 2\hat{\mathbf{s}}_{t+1} + \hat{\mathbf{s}}_t$. The 3D counterpart and implementation details of $\mathcal{L}_{\text{seat}}$ are provided in the Supplementary Material.

Auxiliary semantic heads. We apply last-step pooling for clip-level recognition:

$$\hat{\mathbf{y}}^{\text{dbr}} = D_{\text{dbr}}(\hat{\mathbf{z}}_T^{\text{int}}), \qquad \hat{\mathbf{y}}^{\text{der}} = D_{\text{der}}(\hat{\mathbf{z}}_T^{\text{int}}), \tag{12}$$

$$\hat{\mathbf{y}}^{\text{tcr}} = D_{\text{tcr}}(\bar{\mathbf{z}}_T^{\text{ext}}), \qquad \hat{\mathbf{y}}^{\text{vcr}} = D_{\text{vcr}}(\bar{\mathbf{z}}_T^{\text{ext}}), \tag{13}$$

trained with cross-entropy losses and used exclusively as auxiliary regularizers. External semantics are decoded from the final external context $\bar{\mathbf{z}}_T^{\text{ext}}$, while internal semantics use the rolled-out state $\hat{\mathbf{z}}_T^{\text{int}}$.

Latent rollout consistency. Besides decoded skeleton supervision, we regularize future internal rollouts by either direct latent matching or velocity matching:

$$\mathcal{L}_{\text{lat}}^{\text{direct}} = \sum_t \|\hat{\mathbf{z}}_{t+1}^{\text{int}} - \mathbf{z}_{t+1}^{\text{int}}\|_2^2, \quad \mathcal{L}_{\text{lat}}^{\text{vel}} = \sum_t \|(\hat{\mathbf{z}}_{t+1}^{\text{int}} - \hat{\mathbf{z}}_t^{\text{int}}) - (\mathbf{z}_{t+1}^{\text{int}} - \mathbf{z}_t^{\text{int}})\|_2^2. \tag{14}$$

The final objective aggregates latent consistency, skeleton supervision, physical priors, auxiliary semantics:

$$\mathcal{L} = \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{skel}}\mathcal{L}_{\text{skel}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}, \tag{15}$$

4 Experiments

4.1 Experimental Setup

Dataset, protocol, and targets. We evaluate Driver-WM on the AIDE assistive driving benchmark [51] to validate its capacity for both kinematic rollout and auxiliary semantic recognition. Each video is segmented into 3-second clips with 10 uniformly sampled frames ($\sim$3.3 Hz). We follow a 5$\rightarrow$5 time-causal protocol: given T_{obs}=5 observed steps, we roll out T_{pred}=5 future steps in latent space, decoding the internal states into HALPE-136 2D skeletons (K=136). Rollout losses (latent and skeleton) are computed exclusively on the future window ($t > T_{\text{obs}}$). We attach four auxiliary classification heads as semantic regularizers: 7-class Driver Behavior (DBR) and 5-class Emotion (DER) for the in-cabin stream, alongside 3-class Traffic Context (TCR) and 5-class Vehicle Condition (VCR) for the out-cabin stream. Ground-truth semantic labels are used exclusively to supervise these auxiliary heads. They are strictly excluded from serving as conditioning inputs to the temporal transition or gating pathways (i.e., label embeddings are disabled) to ensure a purely vision-driven rollout.

Implementation details. We pre-extract per-frame features (D=2048) using a frozen Qwen3-VL-2B [3] with a fixed prompt to avoid label leakage. Models are trained for 80 epochs using AdamW (learning rate 5×10^{-5}, batch size 32) on A5000 GPUs. In inference, temporal causality is enforced by truncating the internal/external histories passed to the context interaction and gating modules at each step (strictly $\leq t$). Unless stated otherwise, models are trained from scratch;

Table 1: Main results on AIDE under the 5→5 rollout. We report horizon-averaged geometric errors (MPJPE in px, d-nMPJPE in %) and semantic Macro-F1.

Model	All		HM		PCK	DBR	DER	TCR	VCR
	MPJPE↓	d-nMPJPE↓	MPJPE↓	d-nMPJPE↓	@0.05↑	F1↑	F1↑	F1↑	F1↑
Motion-only & offline baselines									
Zero-Velocity† [34]	52.89	2.40	139.19	6.32	85.95	8.27	14.94	–	–
ST-GCN [49]	110.98	5.04	158.10	7.18	60.97	18.96	22.66	–	–
SiMLPe [14]	106.38	4.83	156.29	7.09	63.45	37.26	41.69	–	–
MotionBERT [63]	73.51	3.34	141.53	6.42	78.01	56.70	52.71	–	–
Offline Enc-Dec [34]	75.87	3.45	144.05	6.54	77.17	53.82	50.68	–	–
Controlled contextual baselines (identical VLM interface)									
Static pooling (no rollout)‡	**68.50**	**3.11**	**134.56**	**6.11**	**72.55**	54.04	60.07	26.39	14.00
Single-stream	90.87	4.13	147.44	6.69	59.63	24.56	24.35	26.39	14.00
Late fusion	82.59	3.75	143.31	6.51	65.07	45.87	49.10	87.85	65.97
Cross-Attn only	80.14	3.64	142.41	6.47	66.43	62.52	68.75	87.94	**69.09**
Ours									
Driver-WM (main)	**71.47**	**3.24**	138.03	6.27	**71.66**	68.07	72.61	90.15	68.34
Driver-WM (full pretrained)	72.68	3.30	138.03	6.27	70.99	73.22	73.97	**90.54**	62.09
Non-causal (bidir)	72.15	3.27	**136.50**	**6.20**	71.39	**73.35**	**74.46**	88.65	**68.82**

† Copy-last-frame baseline favored by ℓ_2 metrics under inertial continuity.
‡ Static pooling baseline (no rollout) can yield low horizon-averaged ℓ_2 errors via mean-regression.
– Externally grounded semantics (TCR/VCR) are inapplicable for motion-only baselines.
HM The High-Motion subset: the Top 10% clips ranked by a ground-truth motion score computed on the future window as the mean per-joint frame-to-frame displacement in pixel coordinates (contains 60/609 test clips; see Supplementary).

warm-start variants are detailed in the Supplementary. We additionally verify strict zero-lookahead numerically: replacing the unobservable future external suffix with noise or zeros changes the predicted skeletal tensors by at most 5×10^{-7} (max abs diff; see Supplementary).

Metrics and baselines. We report horizon-averaged MPJPE (px), diagonal-normalized MPJPE (d-nMPJPE, %), and PCK@0.05; PCK@0.10 and per-head Accuracy are deferred to the Supplementary. We compute d-nMPJPE as $100 \times \text{MPJPE}/d$ with the image diagonal $d = \sqrt{1920^2 + 1080^2}$, and use a threshold of $0.05d$ for PCK@0.05. Semantic performance is measured by Macro-F1. We compare against motion-only forecasters (Zero-Velocity [34], ST-GCN [49], SiMLPe [14], MotionBERT [63]) and an offline encoder-decoder reference [34] representing the non-causal motion bound. To rigorously isolate architectural gains from visual backbone disparities, we additionally construct controlled contextual baselines (e.g., static pooling, single-stream, late fusion) under the exact same frozen Qwen3-VL interface. All learnable methods follow the identical 5→5 rollout protocol.

4.2 Main Results

Table 1 reports the comparison under the fixed 5→5 causal rollout. To focus on paradigm-level differences, we defer dedicated ablations and controlled variants to Sec. 4.3.

Motion-only baselines and an inertia effect under ℓ_2-style metrics. An apparent outlier in Table 1 is the Zero-Velocity baseline, which attains the lowest overall d-nMPJPE. This is consistent with naturalistic driving clips dominated by mild motion and inertial continuity: copying the last observed pose exploits

this inertia to reduce an ℓ_2-style geometric metric globally. However, autonomous safety hinges on reactive, high-dynamic moments. As illustrated in Fig. 3(b), when focusing on the safety-critical High-Motion tail, the Zero-Velocity baseline's error grows sharply as the horizon extends. While simpler baselines benefit from strong spatial priors at initial steps, Driver-WM successfully captures reactive dynamics, significantly mitigating the rapid error accumulation observed in motion-only models and achieving competitive long-horizon fidelity (e.g., h=5 MPJPE 155.82 vs. S0 178.62) on the High-Motion tail. While stronger motion-only forecasters (e.g., MotionBERT) improve kinematic fidelity, they do not explicitly condition on out-cabin context and degrade on the High-Motion tail at longer horizons. Finally, the offline variant serves strictly as a non-causal reference.

Contextual paradigms and a static mean-regression trap. Introducing external visual context changes the trade-off substantially under the same frozen Qwen3-VL interface. Notably, the static pooling baseline (no rollout) can appear strong on horizon-averaged ℓ_2 metrics, including on the HM subset, since a time-invariant, time-averaged hypothesis may reduce the average penalty when reactive maneuvers exhibit temporal ambiguity under sparse sampling. However, this baseline does not produce an autoregressive trajectory and therefore cannot support horizon-wise degradation analysis or test-time mechanism probes. In addition, its externally grounded Macro-F1 remains low (Table 1), indicating limited balanced recognition of traffic context. In contrast, Driver-WM explicitly models multi-step dynamics conditioned on traffic observations and yields robust long-horizon rollouts on the safety-critical High-Motion tail (Fig. 3(b)).

Driver-WM. Driver-WM maintains competitive horizon-averaged errors while improving long-horizon behavior on the HM tail (Fig. 3b). The static pooling baseline can reduce the averaged ℓ_2 error by mean regression, but produces no autoregressive trajectory and cannot support mechanism probes. Driver-WM instead performs time-causal rollouts conditioned on traffic observations, improving semantic consistency under the same frozen interface. The non-causal variant is an offline reference and does not yield consistent gains, supporting causal modeling as a deployment-consistent bias.

Robustness. Across three random initialization seeds, the main Driver-WM configuration demonstrates high stability, yielding a consistent All d-nMPJPE of $3.25\pm0.01\%$ alongside a robust TCR F1 score of 89.26 ± 1.33. In contrast, suboptimal fusion designs exhibit notably larger variance; for instance, the M2 variant fluctuates with an All MPJPE of 79.93 ± 2.35 px. The complete multi-seed table and a reference comparison to recent classification-only methods are provided in the Supplementary Material. Note that these approaches do not perform kinematic rollout and are therefore not directly comparable to our setting.

Risk-ranking signal. Using only Driver-WM outputs, a composite of predicted motion, DBR/DER probabilities, and uncertainty ranks GT-HM events better than motion-only: AUPRC 0.1785 vs. 0.1364, with random $\sim$0.10 and paired-bootstrap CI $[0.013, 0.094]$ (p=0.0009). This suggests that structured driver rollouts provide safety-relevant risk cues.

Table 2: Ablations on Driver-WM. Protocol and metrics follow Table 1. Geometric metrics are inapplicable for the *no pose head* variant.

Variant	All		HM		PCK	DBR	DER	TCR	VCR
	MPJPE$_\downarrow$	d-nMPJPE$_\downarrow$	MPJPE$_\downarrow$	d-nMPJPE$_\downarrow$	@0.05$_\uparrow$	F1$_\uparrow$	F1$_\uparrow$	F1$_\uparrow$	F1$_\uparrow$
Driver-WM (main)	71.47	3.24	138.03	6.27	71.66	68.07	72.61	90.15	68.34
External Context									
w/o ext context	71.74	3.26	136.94	6.21	71.13	71.33	63.56	26.39	14.00
Dynamics Core & Training Choices									
RSSM/GRU dyn.	78.42	3.56	136.46	6.19	67.48	67.15	72.29	88.26	63.74
+Skeleton-Pre	70.92	3.22	136.04	6.18	72.09	69.55	75.65	88.33	70.24
Regularization & Interface Diagnostics									
KL bottleneck	70.56	3.20	136.50	6.20	71.71	69.89	71.54	88.84	64.68
no-prompt feat.	76.54	3.47	142.08	6.45	68.28	68.88	69.92	76.87	58.59
no pose head	–	–	–	–	–	67.22	70.92	88.83	69.44

4.3 Ablation Studies on Driver-WM Architecture

To isolate the contributions of key design choices in Driver-WM, we evaluate controlled variants around the main model under the identical 5→5 causal rollout. Table 2 summarizes representative ablations.

External context is necessary for externally grounded semantics. Severing the out-cabin stream leads to a substantial degradation on the externally grounded head (TCR), while also reducing DER. Although re-training without external context yields competitive averaged MPJPE due to dominant low-motion segments, it lacks access to environmental triggers. This gap is revealed by test-time interventions on the jointly trained model: removing external context increases rollout error substantially (e.g., +15.2 px at h=5; Table 3), indicating that Driver-WM strictly relies on traffic cues for reactive forecasting.

Dynamics core and training choices. Replacing the transition kernel with an RSSM/GRU-style recurrent dynamics degrades geometric fidelity and lowers semantic consistency, suggesting that self-attention is better suited for long-horizon conditional rollouts under high-dimensional context features. Skeleton pre-training yields only a limited effect relative to our main model and does not consistently improve semantics, indicating that the main gains of Driver-WM are not driven by warm-starting the skeleton predictor.

Regularization and interface diagnostics. Introducing a KL bottleneck results in marginal changes across metrics, implying that the dominant performance driver in this setting is the causal injection structure rather than probabilistic regularization. Removing prompted external features consistently degrades both geometric and semantic metrics, highlighting the importance of the external feature interface for aligning driver dynamics with traffic context. Finally, removing the pose forecasting head retains competitive semantic performance on TCR but removes kinematic rollouts altogether; this diagnostic underscores the distinction between discriminative semantics and physically grounded forecasting, motivating kinematic prediction as the core foundation in Driver-WM.

Table 3: Controlled interventions on Driver-WM (main, test set). We report absolute deviations (Δ) in pixel space between the intervened rollouts and the factual (unintervened) rollouts. Larger deviations indicate higher sensitivity to the corresponding intervention. h=5 denotes the last predicted step.

Intervention $do(\cdot)$	**Target**	ΔAll	ΔHM	Δh=5	ΔHead	ΔHands
Factual (Driver-WM, main)	None	0.000	0.000	0.000	0.000	0.000
do(Ext = Swap_clip)	Cross-video content swap	5.363	4.794	8.042	5.462	3.502
do(Ext = $\emptyset$)	Remove out-cabin features	12.953	8.691	15.208	7.385	12.937
do(Ext = Shift_large)	Large temporal offset	0.793	0.892	1.009	0.661	0.018
do(Ext = DropView)	Drop one external view	0.419	0.100	0.478	0.392	0.265
$do(\lambda_{\text{CA}}=0)$ / $do(g=0)$	Disable pathway	89.641	62.232	116.751	36.783	95.666
$do(\lambda_{\text{CA}}=1)$ / $do(g=1)$	Force injection	26.733	11.480	42.170	15.861	26.727
$do(\lambda_{\text{CA}}=2)$	Over-injection	43.015	23.959	52.142	33.909	43.695

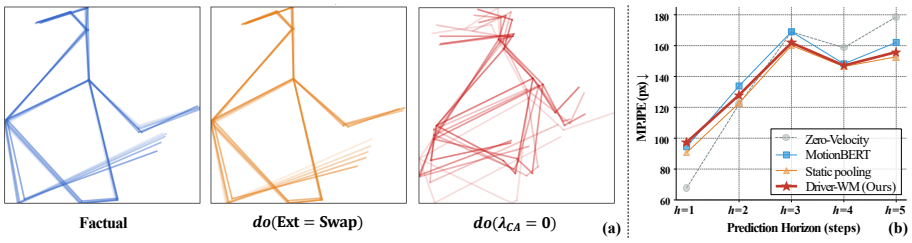

Fig. 3: Mechanism and dynamics. (a) **Controlled interventions:** On the same clip, swapping the out-cabin context or disabling injection (λ_{CA}=0) alters reactive hand motion; frames are aligned to the maximal injection step. (b) **High-Motion tail:** Horizon-wise MPJPE shows the zero-velocity baseline degrades with horizon, while Driver-WM substantially reduces long-horizon error compared to motion-only baselines.

4.4 Interventions and Qualitative Analysis

We perform post-hoc test-time interventions on the trained Driver-WM (main) to assess (i) whether external context is effectively used and (ii) whether the injection pathway is necessary and controllable. Specifically, we intervene on the external stream (`Ext`) via cross-video content swapping, large temporal shifting, view dropping, or complete removal; or we intervene on the injection pathway by either clamping the learned gate g to a constant or bypassing it with a scalar override λ_{CA}; $do(g{=}0/1)$ is equivalent to $do(\lambda_{\text{CA}}{=}0/1)$ under clamping or bypass, while $\lambda_{\text{CA}} > 1$ serves as an over-injection stress test. We use interventions as mechanism probes, without making causal effect estimation claims.

Pathway necessity and injection strength. Table 3 shows that disabling the injection pathway (λ_{CA}=0 or g=0) yields the largest factual-rollout deviation, especially at h=5 and on hand joints. Forcing unmodulated injection ($\lambda_{\text{CA}} \in \{1, 2\}$) also causes substantial deviations, indicating that external signals should be injected with a learned, time-varying intensity rather than applied unconditionally. Direct-GT errors in the Supplementary Material confirm the same ordering.

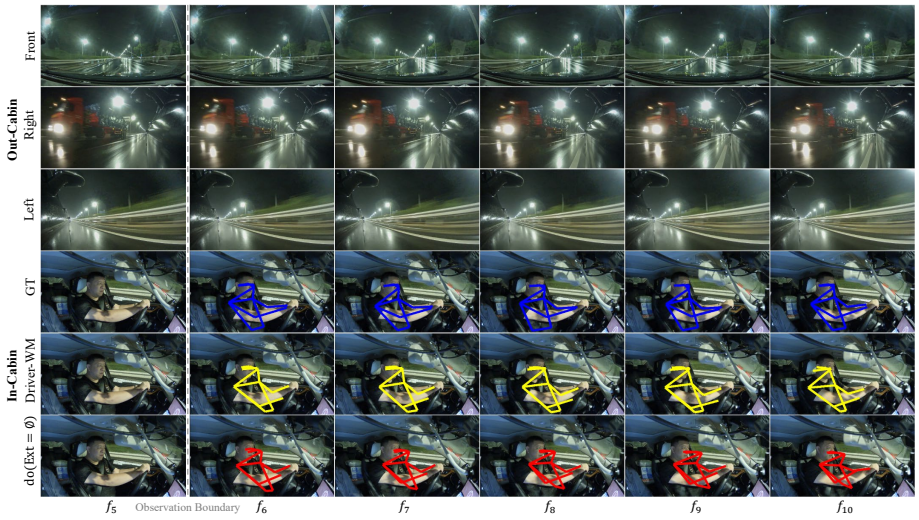

Fig. 4: Qualitative results of driver dynamics rollout and causal interventions. Setting: 5→5 rollout (T_{obs}=5 frames observed, T_{pred}=5 frames predicted). We compare the factual rollout of Driver-WM against the intervened rollout without external traffic context ($do(\text{Ext}=\emptyset)$). To avoid any video-generation artifacts, we directly overlay the 5-step predicted skeletons ($f_6 \sim f_{10}$) onto the corresponding ground-truth in-cabin frames in pixel space. The intervention yields a clearly different hand/upper-body kinematic trajectory in the future window.

Effects of external context. Both swapping external context (Swap_clip) and removing it entirely (Ext=$\emptyset$) induce consistent deviations, with a larger degradation when the external stream is removed. The effect is more pronounced at longer horizons and on reactive joints (hands). In contrast, large temporal shifts and dropping a single view produce comparatively minor deviations in this setting, suggesting empirical robustness to coarse temporal misalignment and partial viewpoint loss. This is consistent with early mean-pooling of multi-view external latents before entering the dynamics core.

Visualizing causal responses. Figures 3 and 4 provide qualitative counterparts of Table 3. In particular, Fig. 3 visualizes the degradation under $do(\lambda_{\text{CA}}=0)$. Fig. 4 overlays the predicted factual and intervened skeleton trajectories onto the corresponding ground-truth in-cabin frames (future window $f_6 \sim f_{10}$) to highlight the kinematic divergence induced by intervention.

5 Conclusion

We present Driver-WM, a dual-stream autoregressive world model that forecasts in-cabin driver dynamics explicitly conditioned on out-cabin traffic context. The architecture unifies the prediction of geometric skeleton trajectories and auxiliary semantic factors via a time-causal, gated cross-attention pathway,

which dynamically injects external environmental cues into the internal latent rollout. Evaluations on the AIDE benchmark expose a critical inertia effect in standard ℓ_2 metrics, where motion-only baselines exploit static extrapolation to attain deceptively low errors but degrade substantially on reactive maneuvers. Driver-WM circumvents this mean-regression trap, achieving robust kinematic fidelity alongside strong semantic alignment with the driving scene. Controlled test-time interventions corroborate the structural design: disabling the causal injection causes severe rollout degradation, while unmodulated injection leads to destabilization. This confirms the necessity of learned, time-varying contextual gating. Overall, Driver-WM shows that driver forecasting is best treated as a traffic-grounded rollout problem, bridging recognition-oriented cabin monitoring and human-aware risk reasoning.

Acknowledgements

This research/project is supported by A*STAR under the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative (Award: I2501E0045), as well as cash and in-kind contribution from the industry partner(s).

References

1. Adeli, V., Ehsanpour, M., Reid, I., Niebles, J.C., Savarese, S., Adeli, E., Rezatofighi, H.: TRiPOD: Human Trajectory and Pose Dynamics Forecasting in the Wild . In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13370–13380. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01314, https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01314
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond (2024), https://openreview.net/forum?id=qrGjFJVl3m
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025)
4. Cao, Q., Grubmüller, S., Dikic, B., Solmaz, S., Innerwinkler, P., Šikić, H., Veas, E.: A simulation-based benchmark for lidar slam in featureless tunnels with a novel robustness evaluation. In: The IEEE Intelligent Vehicles Symposium. IEEE, United States (2026), https://ieee-iv.org/2026/, 2026 IEEE Intelligent Vehicles Symposium, IV 2026 ; Conference date: 22-06-2026 Through 22-06-2026
5. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: Challenges and frontiers. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 10164–10183 (Dec 2024). https://doi.org/10.1109/TPAMI.2024.3435937, https://doi.org/10.1109/TPAMI.2024.3435937

6. Chen, L., Li, Y., Huang, C., Li, B., Xing, Y., Tian, D., Li, L., Hu, Z., Na, X., Li, Z., Teng, S., Lv, C., Wang, J., Cao, D., Zheng, N., Wang, F.Y.: Milestones in autonomous driving and intelligent vehicles: Survey of surveys. IEEE Transactions on Intelligent Vehicles **8**(2), 1046–1056 (2023). https://doi.org/10.1109/TIV.2022.3223131
7. Cheng, Z., Zhang, X., Di, D., Wei, C., Li, H., Yang, X.: Moca: Modeling object consistency for 3d camera control in video generation. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=DZcpnudp7f
8. Chi, H., Yang, H., Yang, L., Lv, C.: Vlm-dm: Visual language models for multitask domain adaptation in driver monitoring. In: 2025 IEEE Intelligent Vehicles Symposium (IV). pp. 1280–1285 (2025). https://doi.org/10.1109/IV64158.2025.11097620
9. Debnath, S., Manh, B.D., Liu, Z., Wang, L.: Llmind: Bio-inspired training-free adaptive visual representations for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3133–3142 (June 2026)
10. Dong, Y., Wu, F., Chen, G., Kong, L., Zhu, X., Hu, Q., Zhou, Y., Sun, J., He, J.Y., Dai, Q., Hauptmann, A.G., Cheng, Z.Q.: Towards unified world models for visual navigation via memory-augmented planning and foresight (2026), https://arxiv.org/abs/2510.08713
11. Dong, Y., Wu, F., Dai, Y., Kong, L., Chen, G., Zhu, X., Hu, Q., Wang, T., Garnica, J., Liu, F., Huang, S., Dai, Q., Cheng, Z.Q.: Language-conditioned world modeling for visual navigation (2026), https://arxiv.org/abs/2603.26741
12. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: an embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
13. 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. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 91560–91596. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-2906, https://proceedings.neurips.cc/paper_files/paper/2024/file/a6a066fb44f2fe0d36cf740c873b8890-Paper-Conference.pdf
14. Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., Moreno-Noguer, F.: Back to mlp: A simple baseline for human motion prediction. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4798–4808 (2023). https://doi.org/10.1109/WACV56688.2023.00479
15. 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 e-prints arXiv:2309.17080 (Sep 2023). https://doi.org/10.48550/arXiv.2309.17080
16. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., Li, H.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17853–17862 (June 2023)

17. Huang, W., Zhang, S., Chua, C., Liang, Y., Mao, Z., Yang, H., Lv, C.: Towards safe mobility: A unified transportation foundation model enabled by open-ended vision-language dataset. arXiv preprint arXiv:2604.22260 (2026)
18. Huang, W., Zhang, S., Huang, Q., Wang, Z., Mao, Z., Chua, C., Chen, Z., Chen, L., Lv, C.: Automot: A unified vision-language-action model with asynchronous mixture-of-transformers for end-to-end autonomous driving. arXiv preprint arXiv:2603.14851 (2026)
19. Jain, A., Koppula, H.S., Soh, S., Raghavan, B., Singh, A., Saxena, A.: Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture (2016), https://arxiv.org/abs/1601.00740
20. 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: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8306–8316 (2023). https://doi.org/10.1109/ICCV51070.2023.00766
21. Jiang, T., Xia, S., Xu, Y., Wu, L., Zeng, X., Wang, L., Qiao, Y., Wang, Y.: Vknowu: Evaluating visual knowledge understanding in multimodal llms (2025), https://arxiv.org/abs/2511.20272
22. Kong, L., Yang, W., Mei, J., Liu, Y., Liang, A., Zhu, D., Lu, D., Yin, W., Hu, X., Jia, M., Deng, J., Zhang, K., Wu, Y., Yan, T., Gao, S., Wang, S., Li, L., Pan, L., Liu, Y., Zhu, J., Tsang Ooi, W., Hoi, S.C.H., Liu, Z.: 3D and 4D World Modeling: A Survey. arXiv e-prints arXiv:2509.07996 (Sep 2025). https://doi.org/10.48550/arXiv.2509.07996
23. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
24. Li, S., Ye, N., Li, T., Chitta, K., An, T., Su, P., Wang, B., Liu, H., Lv, C., Li, H.: Optimization-guided diffusion for interactive scene generation (2026), https://arxiv.org/abs/2512.07661
25. Li, Y., Fan, L., He, J., Wang, Y., Chen, Y., Zhang, Z., Tan, T.: Enhancing End-to-End Autonomous Driving with Latent World Model. arXiv e-prints arXiv:2406.08481 (Jun 2024). https://doi.org/10.48550/arXiv.2406.08481
26. Liang, J., Cao, Y., Ma, Y., Zhao, H., Sartoretti, G.: Hdplanner: Advancing autonomous deployments in unknown environments through hierarchical decision networks. IEEE Robotics and Automation Letters **10**(1), 256–263 (2025). https://doi.org/10.1109/LRA.2024.3506281
27. Liao, S., Lv, X., Lew, J., Zhang, S., Liang, J., Li, P., Cao, Y., Wu, W., Sartoretti, G.: Fare: Fast-slow agentic robotic exploration (2026), https://arxiv.org/abs/2601.14681
28. Lin, J., Zhu, C., Kneuertz, P.J., Bai, Y., Xue, Y.: When models learn to ask why: Adaptive causal reasoning for trustworthy medical vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5556–5568 (2026)
29. Liu, H., Huang, Z., Huang, W., Yang, H., Mo, X., Lv, C.: Hybrid-prediction integrated planning for autonomous driving. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2597–2614 (2025). https://doi.org/10.1109/TPAMI.2025.3526936
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)

31. Liu, M., Zhang, D., Liu, J., Cui, J., Xie, H., Chen, G., Ye, H., Yang, M.Y., Nex, F., Cheng, H.: Driveva: Video action models are zero-shot drivers (2026), https://arxiv.org/abs/2604.04198
32. Liu, W., Wang, W., Qiao, Y., Guo, Q., Zhu, J., Li, P., Chen, Z., Yang, H., Li, Z., Wang, L., Tan, T., Liu, H.: MMTL-UniAD: A Unified Framework for Multimodal and Multi-Task Learning in Assistive Driving Perception . In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6864–6874. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2025). https://doi.org/10.1109/CVPR52734.2025.00644, https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00644
33. Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, S., Voit, M., Stiefelhagen, R.: Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2801–2810 (2019). https://doi.org/10.1109/ICCV.2019.00289
34. Martinez, J., Black, M.J., Romero, J.: On Human Motion Prediction Using Recurrent Neural Networks . In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4674–4683. IEEE Computer Society, Los Alamitos, CA, USA (Jul 2017). https://doi.org/10.1109/CVPR.2017.497, https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.497
35. Rahimi, A., Gerard, V., Zablocki, E., Cord, M., Alahi, A.: Mad: Motion appearance decoupling for efficient driving world models (2026), https://arxiv.org/abs/2601.09452
36. SAE On-Road Automated Vehicle Standards Committee: Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. SAE Standard J3016 (2014)
37. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LII. p. 256–274. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72943-0_15, https://doi.org/10.1007/978-3-031-72943-0_15
38. Su, C., Yang, H., Li, J., Wu, X.: Steering authority allocation strategy for human-machine shared control based on driver take-over feasibility. Computers and Electrical Engineering **120**, 109753 (2024). https://doi.org/https://doi.org/10.1016/j.compeleceng.2024.109753, https://www.sciencedirect.com/science/article/pii/S0045790624006803
39. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution (2024), https://arxiv.org/abs/2409.12191
40. Wang, T., Lin, X., Xu, Y., Ye, Q., Guo, D., Escalera, S., Khoriba, G., Yu, Z.: Micro-gesture recognition: A comprehensive survey of datasets, methods, and challenges. Machine Intelligence Research **23**(2), 308–330 (2026)
41. Wang, W., Wang, L., Zhang, C., Liu, C., Sun, L.: Social interactions for autonomous driving: A review and perspectives. Found. Trends Robot **10**(3–4), 198–376 (Nov 2022). https://doi.org/10.1561/2300000078, https://doi.org/10.1561/2300000078
42. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: Computer Vision – ECCV

2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVIII. p. 55–72. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73195-2_4, https://doi.org/10.1007/978-3-031-73195-2_4

43. Weaver, B.W., DeLucia, P.R.: A systematic review and meta-analysis of takeover performance during conditionally automated driving. Human Factors **64**(7), 1227–1260 (2022)
44. Wu, L., Jiang, T., Dong, Y., Yang, H., Zhang, F., Meng, S., Xuan, A., Song, L., Keung, J.: Lavit: Aligning latent visual thoughts for multi-modal reasoning (2026), https://arxiv.org/abs/2601.10129
45. Xing, Y., Lv, C., Cao, D., Hang, P.: Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving. Transportation Research Part C: Emerging Technologies **128**, 103199 (2021). https://doi.org/https://doi.org/10.1016/j.trc.2021.103199
46. Xing, Y., Lv, C., Wang, H., Cao, D., Velenis, E., Wang, F.Y.: Driver activity recognition for intelligent vehicles: A deep learning approach. IEEE Transactions on Vehicular Technology **68**(6), 5379–5390 (2019). https://doi.org/10.1109/TVT.2019.2908425
47. Xiong, Z., Ye, X., Yaman, B., Cheng, S., Lu, Y., Luo, J., Jacobs, N., Ren, L.: Unidrive-wm: Unified understanding, planning and generation world model for autonomous driving (2026), https://arxiv.org/abs/2601.04453
48. Yan, R., Yang, M., Zhang, X., Tang, H., Yin, Q., Deng, Z., Zhang, K., Zhang, L., Ma, S.: Gaumvc: Generative decoupled gaussian representation for human-centric multi-view video compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4963–4972 (June 2026)
49. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'18/IAAI'18/EAAI'18, AAAI Press (2018)
50. Yan, T., Zheng, H., Chen, D., Qu, M., Shen, Y., Zhou, L., Tu, M., Wang, B., Chen, G., Ye, H., Sun, H., zhong Xu, C., Shen, J.: Causaldrive: Real-time causal world models for autonomous driving (2026), https://arxiv.org/abs/2606.15341
51. Yang, D., Huang, S., Xu, Z., Li, Z., Wang, S., Li, M., Wang, Y., Liu, Y., Yang, K., Chen, Z., Wang, Y., Liu, J., Zhang, P., Zhai, P., Zhang, L.: Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20402–20413 (2023). https://doi.org/10.1109/ICCV51070.2023.01871
52. Yang, H., Liu, H., Hu, Z., Nguyen, A.T., Guerra, T.M., Lv, C.: Quantitative identification of driver distraction: A weakly supervised contrastive learning approach. IEEE Transactions on Intelligent Transportation Systems **25**(2), 2034–2045 (2024). https://doi.org/10.1109/TITS.2023.3316203
53. Yang, H., Zhou, Y., Wu, J., Liu, H., Yang, L., Lv, C.: Human-guided continual learning for personalized decision-making of autonomous driving. IEEE Transactions on Intelligent Transportation Systems **26**(4), 5435–5447 (2025). https://doi.org/10.1109/TITS.2024.3524609
54. Yue, Z., Jin, B., Zeng, H., Zhuang, H., Qin, Z., Yoon, J., Shang, L., Han, J., Wang, D.: Hybrid latent reasoning via reinforcement learning. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 5501–5530. Curran Associates,

Inc. (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/087f7678af2abbe0c37ecc30bd7d1adf-Paper-Conference.pdf

55. Zhang, B., Shou, M.Z.: Ego-centric predictive model conditioned on hand trajectories (2025), https://arxiv.org/abs/2508.19852
56. Zhang, S., Liang, J., Zhou, Z., Ye, S., Wang, Y., Tan, D.M.S., Chiun, J., Cao, Y., Sartoretti, G.: Orion: Option-regularized deep reinforcement learning for cooperative multi-agent online navigation. IEEE Robotics and Automation Letters **11**(7), 9016–9023 (2026). https://doi.org/10.1109/LRA.2026.3699158
57. Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., Huang, G., Chen, X., Wang, B., Zhang, Y., Mei, W., Wang, X.: Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12015–12026 (June 2025)
58. 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 (2024), https://arxiv.org/abs/2403.06845
59. Zhao, Q., Li, Y., Sun, Q., Yan, Z.: Resilphase: Plug-and-play phase mapping and noise-resilient macro-trajectory extrapolation for diffusion acceleration (2026), https://arxiv.org/abs/2606.26769
60. Zheng, Y., Yang, P., Xing, Z., Zhang, Q., Zheng, Y., Gao, Y., Li, P., Zhang, T., Xia, Z., Jia, P., Lang, X., Zhao, D.: World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28632–28642 (October 2025)
61. Zhu, C., Zeng, J., Jiang, J., Lin, J., Wang, Y.: Medsynapse-v: Bridging visual perception and clinical intuition via latent memory evolution (2026), https://arxiv.org/abs/2604.26283
62. Zhu, G., Fan, S., Dai, H., Ho, E.S.L.: Waymo-3dskelmo: A multi-agent 3d skeletal motion dataset for pedestrian interaction modeling in autonomous driving. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 13184–13190. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3758273, https://doi.org/10.1145/3746027.3758273
63. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: MotionBERT: A Unified Perspective on Learning Human Motion Representations . In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15039–15053. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.01385, https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01385
64. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P.R., Salazar, G., Ryoo, M.S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.W.E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N.J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K.A., Driess, D., Ding, T., Choromanski, K.M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M.G., Han, K.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Tan, J., Toussaint, M., Darvish, K. (eds.) Proceedings of The 7th Conference on Robot Learning.

Proceedings of Machine Learning Research, vol. 229, pp. 2165–2183. PMLR (06–09 Nov 2023), https://proceedings.mlr.press/v229/zitkovich23a.html